CMOS ELECTRONICS

IEEE Press
445 Hoes Lane
Piscataway, NJ 08855

Kenneth Moore, *Director of Book and Information Services (BIS)*
Catherine Faduska, *Senior Acquisitions Editor*
Christina Kuhnen, *Associate Acquisitions Editor*

IEEE Computer Society, *Sponsor*
C-S Liaison to IEEE Press, Michael Williams

Technical Reviewers
Cecilia Metra, University of Bologna, Italy
Antonio Rubio, Polytechnical University of Catalonia, Barcelona, Spain
Robert Aitken, Artisan Corp., San Jose, California
Robert Madge, LSI Logic Corp., Gresham, Oregon
Joe Clement, William Filler, Duane Bowman, and David Monroe, Sandia National Laboratories, Albuquerque, New Mexico
Jose Luis Rosselló and Sebastia Bota, University of the Balearic Islands, Spain
Harry Weaver and Don Neamen, Univerisity of New Mexico, Albuquerque, New Mexico
Manoj Sachdev, University of Waterloo, Ontario, Canada

CMOS ELECTRONICS

HOW IT WORKS, HOW IT FAILS

JAUME SEGURA
Universitat de les Illes Balears

CHARLES F. HAWKINS
University of New Mexico

IEEE Computer Society, *Sponsor*

IEEE

IEEE PRESS

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 0-471-47669-2

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

A Mumara, a la seva memòria, i al meu Pare.
To Patricia, Pau, and Andreu for love, for sharing life, and for happiness

—Jaume Segura

To Jan, who has shared the mostly ups, the sometimes downs, and the eternal middle road of life. To our children, Andrea, David, and Shannon, their spouses, and the grandchildren who are following, I hope this work will be a small record of an intense period in which all of you were close to my thoughts.

—Charles F. Hawkins

CONTENTS

PART II FAILURE MODES, DEFECTS, AND TESTING OF CMOS ICs

FOREWORD

Advances in electronics have followed Moore's Law, by scaling feature sizes every generation, doubling transistor integration capacity every two years, resulting in complex chips of today—a treadmill that we take for granted. No doubt that design complexity has grown tremendously and therefore gets tremendous attention, but often overlooked is the technology underlying reliability, and cost-effective test and product engineering of these complex chips. This is becoming especially crucial as we transition from yesterday's micro-electronics to today's nano-electronics.

This book, *CMOS Electronics: How It Works, How It Fails,* written by Professor Segura and Professor Hawkins, addresses just that—the technology underlying failure analysis, testing, and product engineering. The book starts with fundamental device physics, describes how MOS transistors work, how logic circuits are built, and then eases into failure mechanisms of these circuits. Thus the reader gets a very clear picture of failure mechanisms, how to detect them, and how to avoid them. The book covers the latest advances in failure analysis and test and product engineering, such as defects due to bridges, opens, and parametrics, and formulates test strategies to observe these defects in defect-based testing.

As technology progresses with even smaller geometries, you will have to comprehend test and product engineering upfront in the design. That is why this book is a refreshing change as it introduces design and test together.

I would like to thank Professor Segura and Professor Hawkins for giving me an opportunity to read a draft of this book; I surely enjoyed it, learned a lot from it, and I am sure that you the readers will find it rewarding, too.

SHEKHAR BORKAR
Intel Fellow
Director, Circuit Research

PREFACE

If you find the mysteries and varieties of integrated circuit failures challenging, then this book is for you. It is also for you if you work in the CMOS integrated circuits (IC) industry, but admit to knowing little about the electronics itself or the important electronics of failure. The goal of this book is knowledge of the electronic behavior that failing CMOS integrated circuits exhibit to customers and suppliers. The emphasis is on electronics at the transistor circuit level, to gain a deeper understanding of why failing circuits act as they do.

There are two audiences for this book. The first are those in the industry who have had little or no instruction in CMOS electronics, but are surrounded by the electronic symptoms that permeate a CMOS customer or supplier. You may have a physics, chemical engineering, chemistry, or biology education and need to understand the circuitry that you help manufacture. Part I of the book is designed to bring you up to speed in preparation for Part II, which is an analysis of the nature of CMOS failure mechanisms. The second audience is the electrical engineering professional or student who will benefit from a systematic description of the electronic behavior mechanisms of abnormal circuits. Part II describes the material reliability failures, bridge and open circuit defects, and the subtle parametric failures that are plaguing advanced technology ICs. The last chapter assembles this information to implement a defect-based detection strategy used in IC testing. Five key groups should benefit from this approach: test engineers, failure analysts, reliability engineers, yield improvement engineers, and designers. Managers and others who deal with IC abnormalities will also benefit.

CMOS manufacturing environments are surrounded with symptoms that can indicate serious test, design, yield, or reliability problems. Knowledge of how CMOS circuits work and how they fail can rapidly guide you to the nature of a problem. Is the problem related to test, design, reliability, failure analysis, yield–quality issues, or problems that may occur during characterization of a product? Is the symptom an outcome of random defects, or is it systematic, with a common failure signature? Is the defect a bridging problem, an open circuit problem, or a subtle speed-related problem?

We know how bridge and open circuit defects cause performance degradation or failure, and we are closing in on a complete picture of the IC failures that depend upon parametric properties of the circuit environment. These properties include temperature and power supply values. There are cost savings when you can shorten the time to diagnosis and intelligently plan test strategies that minimize test escapes. Rapid insights into abnormal electronic behavior shorten the time to diagnosis and also allow more deterministic planning of test strategies. Determination of root cause of failure is much more efficient if you can isolate one of the failure modes from their electronic properties early in the process.

Circuit abnormalities are a constant concern, and the ability to rapidly estimate the defect type is the first step: namely, what are we looking for? Or when planning test strategies, what test methods should we use to match against particular defects? The effect of the many varieties of defects on circuit operation is analyzed in the book with numerical examples. The early portion of the book assumes little knowledge of circuits and transistors, but builds to a mature description of the electronic properties of defects in Part II. The book adopts a self-learning style with many problems, examples, and self-exercises

Part I (Chapters 1–5) describes how defect-free CMOS circuits work and Part II (Chapters 6–10) describes CMOS circuit failure properties and mechanisms. Chapter 1 begins with Ohm's and Kirchhoff's laws, emphasizing circuit analysis by inspection, and concludes with an analysis of diodes and capacitors. Chapter 2 introduces semiconductor physics and how that knowledge leads to transistor operation in Chapter 3. Chapter 3 has abundant problem-solving examples to gain confidence with transistor operation and their interplay when connected to resistors. Chapter 4 builds to transistor circuit operation with two and four transistors such as CMOS inverters, NAND and NOR logic gates, and transmission gates. Chapter 5 shows how CMOS transistor circuits are synthesized from Boolean algebra equations, and also compares different design styles.

Part II builds on the foundation of how good, or defect-free, CMOS circuits operate and extends that to circuit failures. Chapter 6 examines why IC metal and oxide materials fail in time, producing the dreaded reliability failures. Chapters 7–9 analyze the electronic failure properties of bridge, open, and parametric variation to formulate a test strategy that matches suspected defects to a detection method sensitive to that defect type. Chapter 10 brings all of this together in a test engineering approach called defect-based testing (DBT). The links to test, failure analysis, and reliability are emphasized.

The math used in CMOS electronics varies from simple algebraic expressions to complex physical and timing models whose solutions are suitable only for computers. Fortunately, algebraic equations serve most needs when we manually analyze a circuit's current and voltage response in the presence of a defect. A few simple calculus equations appear, particularly for introducing timing and power relations. A goal is to understand the subtle pass/fail operation or loss of noise margin for defective circuits.

The rapid pace of CMOS technologies influenced these electronic descriptions. Virtually all modern IC transistors are now short-channel devices with different model equations than their long-channel transistor predecessors. Although first-order models for short-channel transistors may at first seem simpler than long-channel equivalents, we found that this simplicity bred inaccuracy and clumsiness when used for manual transistor circuit analysis. We describe these problems and one approach for analyzing short-channel devices in Chapter 3, but chose long-channel transistor models to illustrate how to calculate voltages and currents. It was the only expedient way to give readers the intuitive insights into the nature of transistors when these devices are subjected to various voltage biases.

CMOS field effect transistors do not exist in isolation in the IC. Parasitic bipolar transistors have a small but critical role in the destructive CMOS latchup condition, in electrostatic discharge (ESD) protection circuits, and in some parasitic defect structures. However, we chose to contain the size of the book by only including a brief description of bipolar transistors in sections where they are mentioned.

The knowledge poured into this book came from several sources. The authors teach electronics at their universities, and much data and knowledge were taken from collaborative work done at Sandia National Labs and at Intel Corporation. We are indebted to many persons in our defect electronics education and particularly acknowledge those with whom we worked closely. These include Jerry Soden and Ed Cole of Sandia National Labs, Antonio Rubio of the Polytechnical University of Catalonia, Jose Luis Rosselló of the University of the Balearic Islands, Alan Righter of Analog Devices, and Ali Keshavarzi, who hosted each of us on university sabbaticals at the Intel facilities in Portland, Oregon and Rio Rancho, New Mexico.

Each chapter had from one to three reviewers. All reviewers conveyed a personal feeling of wanting to make this a better book. These persons are: Sebastià Bota of the University of Barcelona, Harry Weaver and Don Neamen of the University of New Mexico, Jose Luis Rosselló of the University of the Balearic Islands, Antonio Rubio of the Polytechnic University of Catalonia, Manoj Sachdev of the University of Waterloo, Joe Clement, Dave Monroe, and Duane Bowman of Sandia National Labs, Cecilia Metra of the University of Bologna, Bob Madge of LSI Logic Corp., and Rob Aitken of Artisan Corp. We also thank editing and computer support from Gloria Ayuso of the University of the Balearic Islands and Francesc Segura from DMS, Inc.

The seminal ideas and original drafts for the book began in the Spanish city of Palma de Mallorca on the Balearic Islands. Over the next four years, about two-thirds of the book was written in Mallorca with the remainder at the University of New Mexico in Albuquerque in the United States. Countless editorial revisions occurred, with memorable ones on long flights across the Atlantic, in the cafes of the ancient quarter of Palma, and in the coffee shops around the University of New Mexico, where more than four centuries earlier, Spanish farmers grew crops.

The book is intended to flow easily for those without an EE degree and can be a one-semester course in a university. We found from class teaching with this material that the full book is suitable for senior or graduate students with non-EE backgrounds. The EE students can skip or review Chapters 1–2, and go directly to Chapters 3–10. Chapters 3–5 are often taught at the undergraduate EE level, but we found that the focus on CMOS electronics here is more than typically taught. The style blends the descriptive portions of the text with many examples and exercises to encourage self-study. The learning tools are a pad of paper, pen, pocket calculator, isolated time, and motivation to learn. The rewards are insights into the deep mysteries of CMOS IC behavior. For additional material related to this book, visit http://omaha.uib.es/cmosbook/index.html

JAUME SEGURA
CHARLES F. HAWKINS

January 2004

PART I

CMOS FUNDAMENTALS

CHAPTER 1

ELECTRICAL CIRCUIT ANALYSIS

1.1 INTRODUCTION

We understand complex integrated circuits (ICs) through simple building blocks. CMOS transistors have inherent parasitic structures, such as diodes, resistors, and capacitors, whereas the whole circuit may have inductor properties in the signal lines. We must know these elements and their many applications since they provide a basis for understanding transistors and whole-circuit operation.

Resistors are found in circuit speed and bridge-defect circuit analysis. Capacitors are needed to analyze circuit speed properties and in power stabilization, whereas inductors introduce an unwanted parasitic effect on power supply voltages when logic gates change state. Transistors have inherent diodes, and diodes are also used as electrical protective elements for the IC signal input/output pins. This chapter examines circuits with resistors, capacitors, diodes, and power sources. Inductance circuit laws and applications are described in later chapters. We illustrate the basic laws of circuit analysis with many examples, exercises, and problems. The intention is to learn and solve sufficient problems to enhance one's knowledge of circuits and prepare for future chapters. This material was selected from an abundance of circuit topics as being more relevant to the later chapters that discuss how CMOS transistor circuits work and how they fail.

1.2 VOLTAGE AND CURRENT LAWS

Voltage, current, and resistance are the three major physical magnitudes upon which we will base the theory of circuits. Voltage is the potential energy of a charged particle in an electric field, as measured in units of volts (V), that has the physical units of Newton ·

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

m/coulomb. Current is the movement of charged particles and is measured in coulombs per second or amperes (A). Electrons are the charges that move in transistors and interconnections of integrated circuits, whereas positive charge carriers are found in some specialty applications outside of integrated circuits.

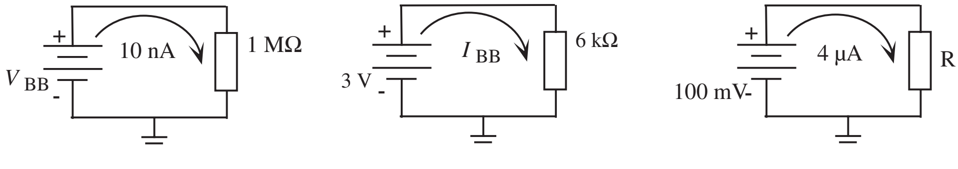
Three laws define the distribution of currents and voltages in a circuit with resistors: Kirchhoff's voltage and current laws, and the volt–ampere relation for resistors defined by Ohm's law. Ohm's law relates the current and voltage in a resistor as

$$V = R \times I \tag{1.1}$$

This law relates the voltage drop (V) across a resistor R when a current I passes through it. An electron loses potential energy when it passes through a resistor. Ohm's law is important because we can now predict the current obtained when a voltage is applied to a resistor or, equivalently, the voltage that will appear at the resistor terminals when forcing a current.

An equivalent statement of Ohm's law is that the ratio of voltage applied to a resistor to subsequent current in that resistor is a constant $R = V/I$, with a unit of volts per ampere called an ohm (Ω). Three examples of Ohm's law in Figure 1.1 show that any of the three variables can be found if the other two are known. We chose a rectangle as the symbol for a resistor as it often appears in CAD (computer-aided design) printouts of schematics and it is easier to control in these word processing tools.

The ground symbol at the bottom of each circuit is necessary to give a common reference point for all other nodes. The other circuit node voltages are measured (or calculated) with respect to the ground node. Typically, the ground node is electrically tied through a building wire called the common to the voltage generating plant wiring. Battery circuits use another ground point such as the portable metal chassis that contains the circuit. Notice that the current direction is defined by the positive charge with respect to the positive terminal of a voltage supply, or by the voltage drop convention with respect to a positive charge. This seems to contradict our statements that all current in resistors and transistors is due to negative-charge carriers. This conceptual conflict has historic origins. Ben Franklin is believed to have started this convention with his famous kite-in-a-thunderstorm experiment. He introduced the terms positive and negative to describe what he called electrical fluid. This terminology was accepted, and not overturned when we found out later that current is actually carried by negative-charge carriers (i.e., electrons). Fortunately, when we calculate voltage, current, and power in a circuit, a positive-charge hypothesis gives the same results as a negative-charge hypothesis. Engineers accept the positive convention, and typically think little about it.



V_{BB} = (10 nA)(1 MΩ) = 10 mV I_{BB} = 3 V/6 kΩ = 500 µA R = 100 mV/4 µA = 25 kΩ

Figure 1.1. Ohm's law examples. The battery positive terminal indicates where the positive charge exits the source. The resistor positive voltage terminal is where positive charge enters.

An electron loses energy as it passes through a resistance, and that energy is lost as heat. Energy per unit of time is power. The power loss in an element is the product of voltage and current, whose unit is the watt (W):

$$P = VI \tag{1.2}$$

1.2.1 Kirchhoff's Voltage Law (KVL)

This law states that "the sum of the voltage drops across elements in a circuit loop is zero." If we apply a voltage to a circuit of many serial elements, then the sum of the voltage drops across the circuit elements (resistors) must equal the applied voltage. The KVL is an energy conservation statement allowing calculation of voltage drops across individual elements: energy input must equal energy dissipated.

Voltage Sources. An ideal voltage source supplies a constant voltage, no matter the amount of current drawn, although real voltage sources have an upper current limit. Figure 1.2 illustrates the KVL law where V_{BB} represents a battery or bias voltage source. The polarities of the driving voltage V_{BB} and resistor voltages are indicated for the clockwise direction of the current.

Naming V_1 the voltage drop across resistor R_1, V_2 that across resistor R_2, and, subsequently, V_5 for R_5, the KVL states that

$$V_{BB} = V_1 + V_2 + V_3 + V_4 + V_5 \tag{1.3}$$

Note that the resistor connections in Figure 1.2 force the same current I_{BB} through all resistors. When this happens, i.e., when the same current is forced through two or more resistors, they are said to be connected in series. Applying Ohm's law to each resistor of Figure 1.2, we obtain $V_i = R_i \times I_{BB}$ (where i takes any value from 1 to 5). Applying Ohm's law to each voltage drop at the right-hand side of Equation (1.3) we obtain

$$\begin{aligned} V_{BB} &= R_1I_{BB} + R_2I_{BB} + R_3I_{BB} + R_4I_{BB} + R_5I_{BB} \\ &= (R_1 + R_2 + R_3 + R_4 + R_5)I_{BB} \\ &= R_{eq}I_{BB} \end{aligned} \tag{1.4}$$

where $R_{eq} = R_1 + R_2 + R_3 + R_4 + R_5$. The main conclusion is that when a number of resistors are connected in series, they can be reduced to an equivalent single resistor whose value is the sum of the resistor values connected in series.

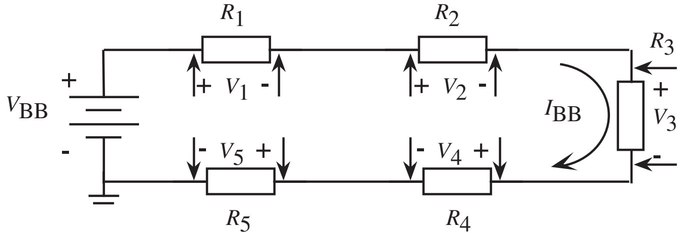


Figure 1.2. KVL seen in series elements.

■ EXAMPLE 1.1

Figure 1.3(a) shows a 5 V source driving two resistors in series. The parameters are referenced to the ground node. Show that the KVL holds for the circuit.

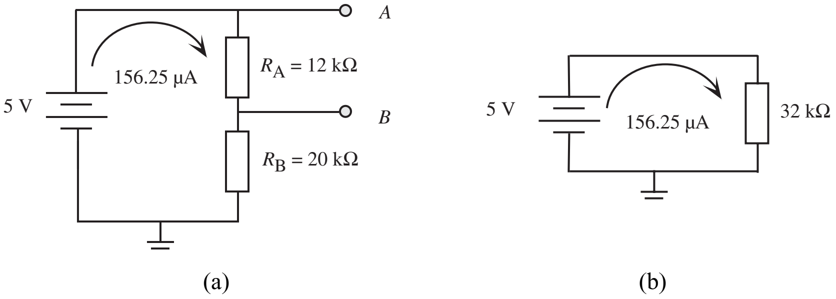


Figure 1.3. (a) Circuit illustrating KVL. (b) Equivalent circuit. The power supply cannot tell if the two series resistors or their equivalent resistance are connected to the power source terminals.

The voltage drop across R_A is

$$V_A = R_A \times I = 12 \text{ k}\Omega \times 156.25 \ \mu\text{A} = 1.875 \text{ V}$$

Similarly, the voltage across R_B is 3.125 V. The applied 5 V must equal the series drops across the two resistors or 1.875 V + 3.125 V = 5 V. The last sentence is a verification of the KVL. ■

The current in Figure 1.3(a) through the 12 kΩ in series with a 20 kΩ resistance is equal to that of a single 32 kΩ resistor (Figure 1.3(b)). The voltage across the two resistors in Figure 1.3(a) is 5 V and, when divided by the current (156.25 μA), gives an equivalent series resistance of (5 V/156.25 μA) = 32 kΩ. Figure 1.3(b) is an equivalent reduced circuit of that in Figure 1.3(a).

Current Sources. We introduced voltage power sources first since they are more familiar in our daily lives. We buy voltage batteries in a store or plug computers, appliances, or lamps into a voltage socket in the wall. However, another power source exists, called a current source, that has the property of forcing a current out of one terminal that is independent of the resistor load. Although not as common, you can buy current power sources, and they have important niche applications.

Current sources are an integral property of transistors. CMOS transistors act as current sources during the crucial change of logic state. If you have a digital watch with a microcontroller of about 200k transistors, then about 5% of the transistors may switch during a clock transition, so 10k current sources are momentarily active on your wrist.

Figure 1.4 shows a resistive circuit driven by a current source. The voltage across the current source can be calculated by applying Ohm's law to the resistor connected between the current source terminals. The current source as an ideal element provides a fixed current value, so that the voltage drop across the current source will be determined by the element or elements connected at its output. The ideal current source can supply an infinite voltage, but real current sources have a maximum voltage limit.

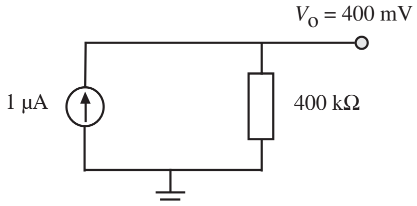


Figure 1.4. A current source driving a resistor load.

1.2.2 Kirchhoff's Current Law (KCL)

The KCL states that "the sum of the currents at a circuit node is zero." Current is a mass flow of charge. Therefore the mass entering the node must equal the mass exiting it. Figure 1.5 shows current entering a node and distributed to three branches. Equation (1.5) is a statement of the KCL that is as essential as the KVL in Equation (1.3) for computing circuit variables. Electrical current is the amount of charge (electrons) Q moving in time, or dQ/dt. Since current itself is a flow (dQ/dt), it is grammatically incorrect to say that "current flows." Grammatically, charge flows, but current does not.

$$I_0 = I_1 + I_2 + I_3 \tag{1.5}$$

The voltage across the terminals of parallel resistors is equal for each resistor, and the currents are different if the resistors have different values. Figure 1.6 shows two resistors connected in parallel with a voltage source of 2.5 V. Ohm's law shows a different current in each path, since the resistors are different, whereas all have the same voltage drop.

$$\begin{aligned} I_A &= \frac{2.5\ \text{V}}{100\ \text{k}\Omega} = 25\ \mu\text{A} \\ I_B &= \frac{2.5\ \text{V}}{150\ \text{k}\Omega} = 16.675\ \mu\text{A} \end{aligned} \tag{1.6}$$

Applying Equation (1.5), the total current delivered by the battery is

$$I_{BB} = I_A + I_B = 25\ \mu\text{A} + 16.67\ \mu\text{A} = 41.67\ \mu\text{A} \tag{1.7}$$

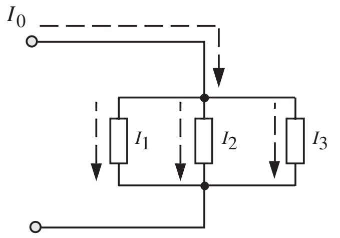


Figure 1.5. KCL and its current summation at a node.

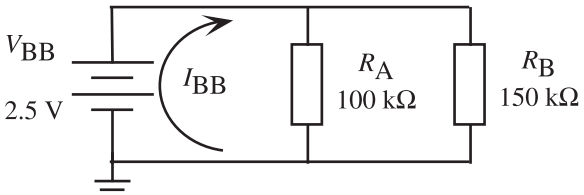


Figure 1.6. Parallel resistance example.

and

$$\frac{V_{\mathrm{BB}}}{I_{\mathrm{BB}}} = \frac{2.5\ \mathrm{V}}{41.67\ \mu\mathrm{A}} = 60\ \mathrm{k}\Omega \tag{1.8}$$

Notice that the sum of the currents in each resistor branch is equal to the total current from the power supply, and that the resistor currents will differ when the resistors are unequal. The equivalent parallel resistance R_{eq} in the resistor network in Figure 1.6 is $V_{\mathrm{BB}} = R_{\mathrm{eq}}(I_{\mathrm{A}} + I_{\mathrm{B}})$. From this expression and using Ohm's law we get

$$\begin{aligned} \frac{V_{\mathrm{BB}}}{R_{\mathrm{eq}}} &= I_{\mathrm{A}} + I_{\mathrm{B}} \\ I_{\mathrm{A}} &= \frac{V_{\mathrm{BB}}}{R_{\mathrm{A}}} \\ I_{\mathrm{B}} &= \frac{V_{\mathrm{BB}}}{R_{\mathrm{B}}} \end{aligned} \tag{1.9}$$

and

$$\begin{aligned} \frac{V_{\mathrm{BB}}}{R_{\mathrm{eq}}} &= \frac{V_{\mathrm{BB}}}{R_{\mathrm{A}}} + \frac{V_{\mathrm{BB}}}{R_{\mathrm{B}}} \\ \frac{1}{R_{\mathrm{eq}}} &= \frac{1}{R_{\mathrm{A}}} + \frac{1}{R_{\mathrm{B}}} \\ R_{\mathrm{eq}} = \frac{1}{\frac{1}{R_{\mathrm{A}}} + \frac{1}{R_{\mathrm{B}}}} &= \frac{1}{\frac{1}{100\ \mathrm{k}\Omega} + \frac{1}{150\ \mathrm{k}\Omega}} = 60\ \mathrm{k}\Omega \end{aligned} \tag{1.10}$$

This is the expected result from Equation (1.8). R_{eq} is the equivalent resistance of R_{A} and R_{B} in parallel, which is notationly expressed as $R_{\mathrm{eq}} = R_{\mathrm{A}} \| R_{\mathrm{B}}$. In general, for n resistances in parallel,

$$R_{\mathrm{eq}} = (R_1 \| R_2 \| \cdots \| R_n) = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}} \tag{1.11}$$

The following examples and self-exercises will help you to gain confidence on the concepts discussed.

Self-Exercise 1.1

Calculate V_0 and the voltage drop V_p across the parallel resistors in Figure 1.7. *Hint:* Replace the 250 kΩ and 180 kΩ resistors by their equivalent resistance and apply KVL to the equivalent circuit.

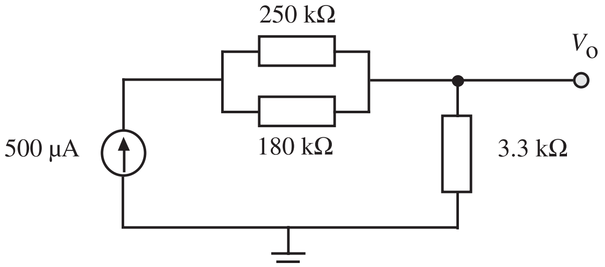


Figure 1.7.

Self-Exercise 1.2

Calculate R_3 in the circuit of Figure 1.8.

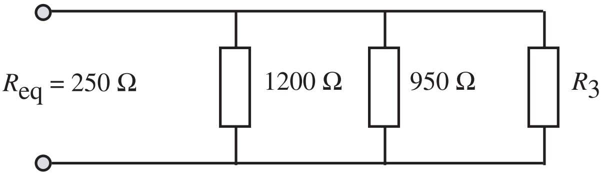


Figure 1.8. Equivalent parallel resistance.

When the number of resistors in parallel is two, Equation (1.11) reduces to

$$R_p = \frac{R_A \times R_B}{R_A + R_B} \tag{1.12}$$

■ EXAMPLE 1.2

Calculate the terminal resistance of the resistors in Figures 1.9(a) and (b).

$$\begin{aligned} R_{eq} &= 1\ \text{M}\Omega \| 2.3\ \text{M}\Omega \\ &= \frac{(10^6)(2.3 \times 10^6)}{10^6 + 2.3 \times 10^6} \\ &= 697\ \text{k}\Omega \end{aligned}$$

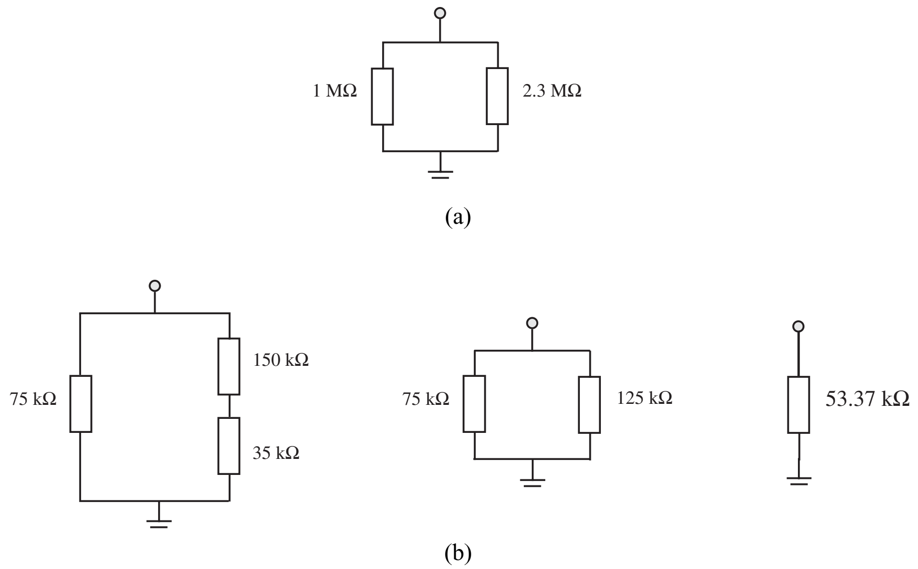


Figure 1.9. Parallel resistance calculations and equivalent circuits.

The equivalent resistance at the network in Figure 1.9(b) is found by combining the series resistors to 185 kΩ and then calculating the parallel equivalent:

$$R_{eq} = 75\ \text{k}\Omega \| 185\ \text{k}\Omega$$

$$= \frac{(75 \times 10^3)(185 \times 10^3)}{75 \times 10^3 + 185 \times 10^3}$$

■

$$= 53.37\ \text{k}\Omega$$

Self-Exercise 1.3

Calculate the resistance at the voltage source terminals R_{in}, I_{BB}, and V_0 at the terminals in Figure 1.10. If you are good, you can do this in your head.

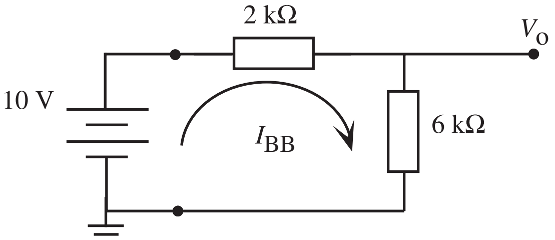


Figure 1.10.

Self-Exercise 1.4

Use Equations (1.11) or (1.12) and calculate the parallel resistance for circuits in Figures 1.11(a)–(d). Estimates of the terminal resistances for circuits in (a) and (b) should be done in your head. Circuits in (c) and (d) show that the effect of a large parallel resistance becomes negligible.

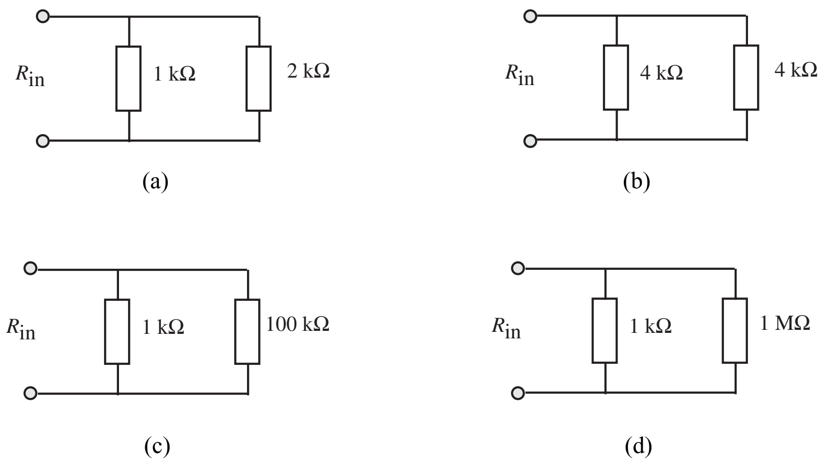


Figure 1.11.

Self-Exercise 1.5

Calculate R_{in}, I_{BB}, and V_0 in Figure 1.12. Estimate the correctness of your answer in your head.

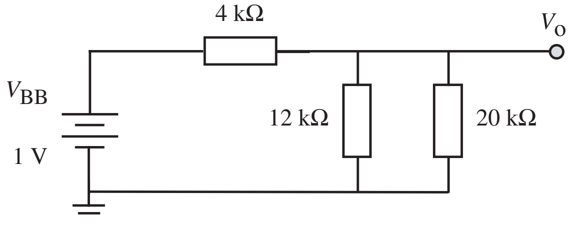


Figure 1.12.

Self-Exercise 1.6

(a) In Figure 1.13, find I_3 if $I_0 = 100$ μA, $I_1 = 50$ μA, and $I_2 = 10$ μA. (b) If $R_3 = 50$ kΩ, what are R_1 and R_2?

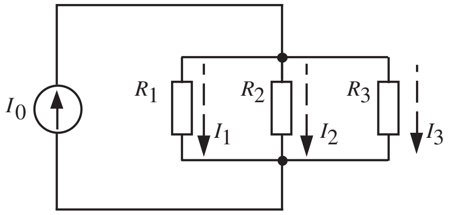


Figure 1.13.

Self-Exercise 1.7

Calculate R_1 and R_2 in Figure 1.14.

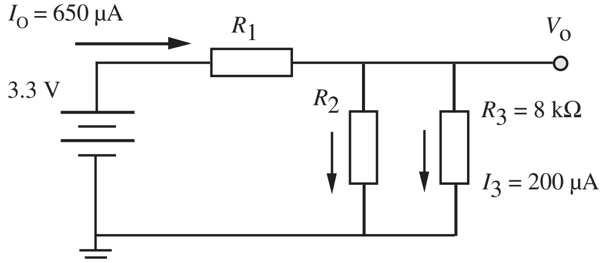


Figure 1.14.

Self-Exercise 1.8

If the voltage across the current source is 10 V in Figure 1.15, what is R_1?

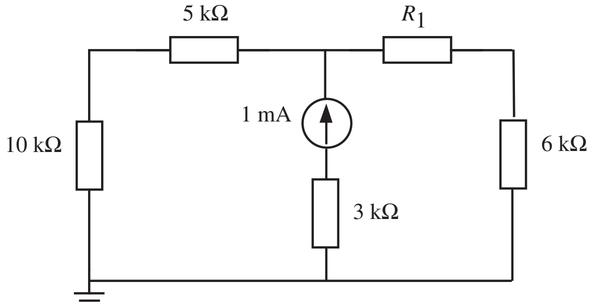


Figure 1.15.

Resistance Calculations by Inspection. A shorthand notation for the terminal resistance of networks allows for quick estimations and checking of results. The calculations are defined before computing occurs. Some exercises below will illustrate this. Solutions are given in the Appendix.

Self-Exercise 1.9

Write the shorthand notation for the terminal resistance of the circuits in Figures 1.16(a) and (b).

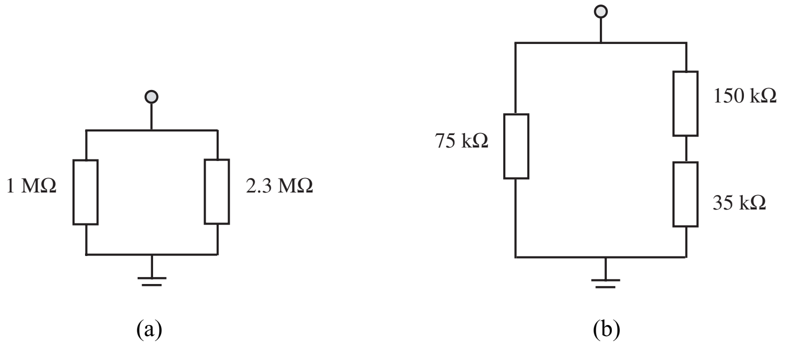


Figure 1.16.

Self-Exercise 1.10

Write the shorthand notation for the terminal resistance of the three circuits in Figure 1.17.

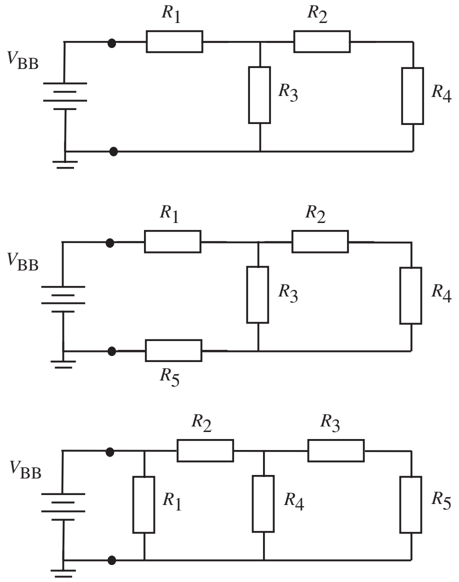


Figure 1.17. Terminal resistance using shorthand notation.

Self-Exercise 1.11

In the lower circuit of Figure 1.17, $R_1 = 20$ kΩ, $R_2 = 15$ kΩ, $R_3 = 25$ kΩ, $R_4 = 8$ kΩ, $R_5 = 5$ kΩ. Calculate R_{eq} for these three circuits.

Dividers. Some circuit topologies are repetitive and lend themselves to analysis by inspection. Two major inspection techniques use *voltage divider* and *current divider* concepts that take their analysis from the KVL and KCL. These are illustrated below with derivations of simple circuits followed by several examples and exercises. The examples have slightly more elements, but they reinforce previous examples and emphasize analysis by inspection.

Figure 1.18 shows a circuit with good visual voltage divider properties that we will illustrate in calculating V_3.

The KVL equation is

$$V_{BB} = I_{BB}(R_1) + I_{BB}(R_2) + I_{BB}(R_3) = I_{BB}(R_1 + R_2 + R_3)$$
$$I_{BB} = \frac{V_{BB}}{R_1 + R_2 + R_3} = \frac{V_3}{R_3} \tag{1.13}$$

and

$$V_3 = \frac{R_3}{R_1 + R_2 + R_3} V_{BB} \tag{1.14}$$

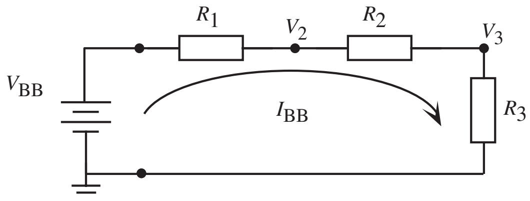


Figure 1.18. Voltage divider circuit.

Equation (1.14) is a shorthand statement of the voltage divider. It is written by inspection, and calculations follow. The voltage dropped by each resistor is proportional to their fraction of the whole series resistance. Figure 1.18 is very visual, and you should be able to write the voltage divider expression by inspection for any voltage drop. For example, the voltage from node V_2 to ground is

$$V_2 = \frac{R_2 + R_3}{R_1 + R_2 + R_3} V_{BB} \tag{1.15}$$

Self-Exercise 1.12

Use inspection and calculate the voltage at V_0 (Figure 1.19). Verify that the sum of the voltage drops is equal to V_{BB}. Write the input resistance R_{in} by inspection and calculate the current I_{BB}.

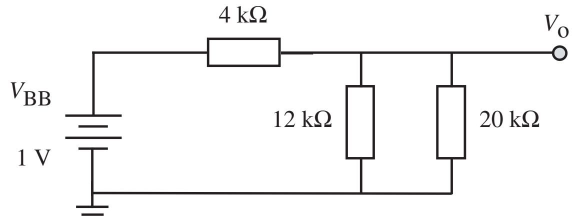


Figure 1.19. Voltage divider analysis circuit.

Self-Exercise 1.13

Write the expression for R_{in} at the input terminals, V_0, and the power supply current (Figure 1.20).

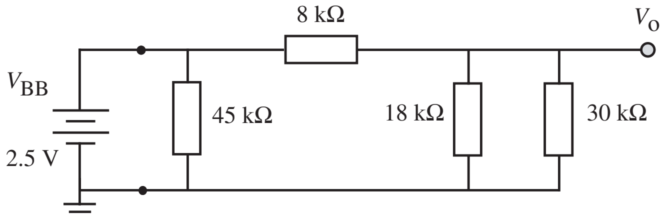


Figure 1.20.

Current divider expressions are visual, allowing you to see the splitting of current as it enters branches. Figure 1.21 shows two resistors that share total current I_{BB}.

KVL gives

$$V_{BB} = (R_1 \| R_2) I_{BB} = \frac{R_1 \times R_2}{R_1 + R_2} I_{BB} = (I_1)(R_1) = (I_2)(R_2) \tag{1.16}$$

then

$$I_1 = \frac{R_2}{R_1 + R_2} I_{BB} \quad \text{and} \quad I_2 = \frac{R_1}{R_1 + R_2} I_{BB} \tag{1.17}$$

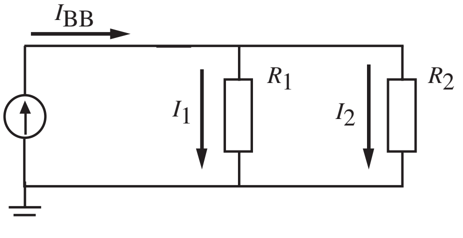


Figure 1.21. Current divider.

Currents divide in two parallel branches by an amount proportional to the opposite leg resistance divided by the sum of the two resistors. This relation should be memorized, as was done for the voltage divider.

Self-Exercise 1.14

Write the current expression by inspection and solve for currents in the 12 kΩ and 20 kΩ paths in Figure 1.22.

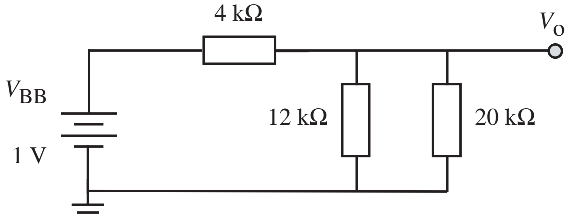


Figure 1.22.

Self-Exercise 1.15

(a) Write the current expression by inspection and solve for currents in all resistors in Figure 1.23, where I_{BB} = 185.4 μA. (b) Calculate V_{BB}.

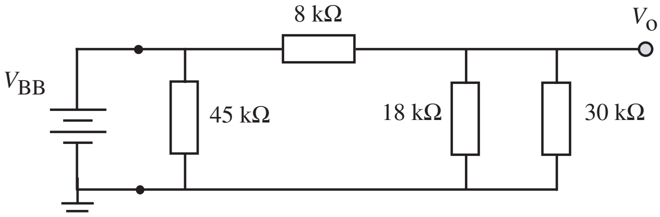


Figure 1.23.

Self-Exercise 1.16

(a) Solve for current in all resistive paths in Figure 1.24 using the technique of inspection. (b) Calculate a new value for the 20 kΩ resistor so that its current is 5 μA.

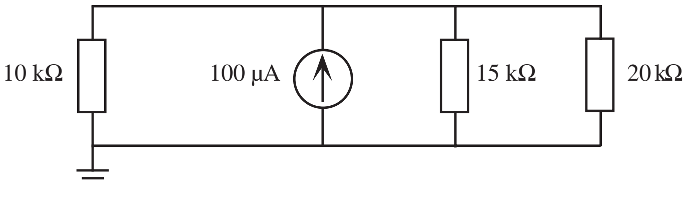


Figure 1.24.

Self-Exercise 1.17

In Figure 1.25, calculate V_0, I_2, and I_9.

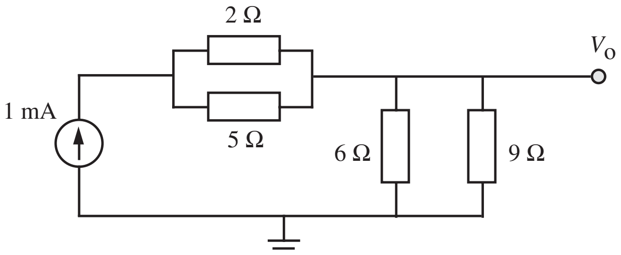


Figure 1.25.

Self-Exercise 1.18

(a) Write R_{in} between the battery terminals by inspection and solve (Figure 1.26). (b) Write the $I_{1.5k}$ expression by inspection and solve. This is a larger circuit, but it presents no problem if we adhere to the shorthand style. We write R_{in} between battery terminals by inspection, and calculate $I_{1.5k}$ by current divider inspection.

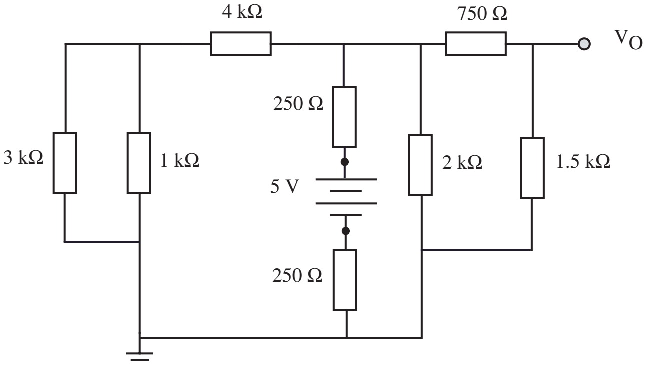


Figure 1.26.

1.3 CAPACITORS

Capacitors appear in CMOS digital circuits as parasitic elements intrinsic to transistors or with the metals used for interconnections. They have an important effect on the time for a transistor to switch between on and off states, and also contribute to propagation delay between gates due to interconnection capacitance. Capacitors also cause a type of noise called cross-talk. This appears especially in high-speed circuits, in which the voltage at one interconnection line is affected by another interconnection line that is isolated but located close to it. Cross-talk is discussed in later chapters.

The behavior and structure of capacitors inherent to interconnection lines are significantly different from the parasitic capacitors found in diodes and transistors. We introduce ideal parallel plate capacitors that are often used to model wiring capacitance. Capacitors inherent to transistors and diodes act differently and are discussed later.

A capacitor has two conducting plates separated by an insulator, as represented in Figure 1.27(a). When a DC voltage is applied across the conducting plates (terminals) of the capacitor, the steady-state current is zero since the plates are isolated by the insulator. The effect of the applied voltage is to store charges of opposite sign at the plates of the capacitor.

The capacitor circuit symbol is shown in Figure 1.27(b). Capacitors are characterized by a parameter called capacitance (C) that is measured in Farads. Strictly, capacitance is defined as the charge variation ∂Q induced in the capacitor when voltage is changed by a quantity ∂V, i.e.,

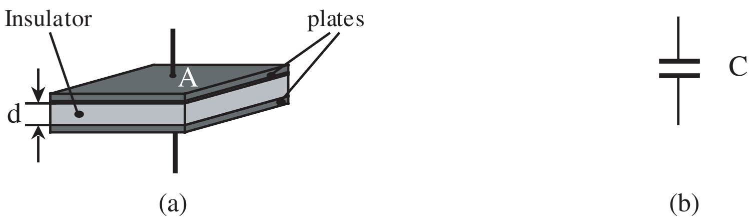
$$C = \frac{\partial Q}{\partial V} \tag{1.18}$$

This ratio is constant in parallel plate capacitors, independent of the voltage applied to the capacitor. Capacitance is simply the ratio between the charge stored and the voltage applied, i.e., $C = Q/V$, with units of Coulombs per Volt called a Farad. This quantity can also be computed from the geometry of the parallel plate and the properties of the insulator used to construct it. This expression is

$$C = \frac{\varepsilon_{\text{ins}} A}{d} \tag{1.19}$$

where ε_{ins} is an inherent parameter of the insulator, called permittivity, that measures the resistance of the material to an electric field; A is the area of the plates used to construct the capacitor; and d the distance separating the plates.

Although a voltage applied to the terminals of a capacitor does not move net charge through the dielectric, it can displace charge within it. If the voltage changes with time, then the displacement of charge also changes, causing what is known as displacement current, that cannot be distinguished from a conduction current at the capacitor terminals. Since this current is proportional to the rate at which the voltage across the capacitor changes with time, the relation between the applied voltage and the capacitor current is



(a) (b)

Figure 1.27. (a) Parallel plate capacitor. (b) Circuit symbol.

$$i = C\frac{dV}{dt} \tag{1.20}$$

If the voltage is DC, then $dV/dt = 0$ and the current is zero. An important consequence of Equation (1.20) is that the voltage at the terminals of a capacitor cannot change instantaneously, since this would lead to an infinite current. That is physically impossible. In later chapters, we will see that any logic gate constructed within an IC has a parasitic capacitor at its output. Therefore, the transition from one voltage level to another will always have a delay time since the voltage output cannot change instantaneously. Trying to make these output capacitors as small as possible is a major goal of the IC industry in order to obtain faster circuits.

1.3.1 Capacitor Connections

Capacitors, like resistors, can be connected in series and in parallel. We will show the equivalent capacitance calculations when they are in these configurations.

Capacitors in parallel have the same terminal voltage, and charge distributes according to the relative capacitance value differences (Figure 1.28(a)). The equivalent capacitor is equal to the sum of the capacitors:

$$\begin{gathered} C_1 = \frac{Q_1}{V}, \qquad C_2 = \frac{Q_2}{V} \\ C_1 + C_2 = \frac{Q_1}{V} + \frac{Q_2}{V} = \frac{Q_1 + Q_2}{V} \\ C_{\text{eq}} = \frac{Q_{\text{eq}}}{V} \end{gathered} \tag{1.21}$$

where $C_{\text{eq}} = C_1 + C_2$, and $Q_{\text{eq}} = Q_1 + Q_2$. Capacitors connected in parallel simply add their values to get the equivalent capacitance.

Capacitors connected in series have the same charge stored, whereas the voltage depends on the relative value of the capacitor (Figure 1.28(b)). In this case, the expression for the equivalent capacitor is analogous to the expression obtained when connecting resistors in parallel:

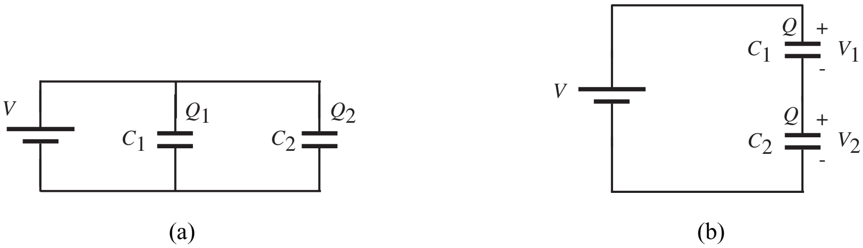


Figure 1.28. Capacitance interconnection. (a) Parallel. (b) Series.

$$C_1 = \frac{Q}{V_1}, \qquad C_2 = \frac{Q}{V_2}$$

$$\frac{1}{C_1} + \frac{1}{C_2} = \frac{V_1}{Q} + \frac{V_2}{Q} = \frac{V_1 + V_2}{Q} \tag{1.22}$$

$$C_{\text{eq}} = \frac{Q}{V_{\text{eq}}}$$

where $V_{\text{eq}} = V_1 + V_2$, and

$$C_{\text{eq}} = \frac{C_1 C_2}{C_1 + C_2}$$

■ **EXAMPLE 1.3**

In Figures 1.28(a) and (b), C_1 = 20 pF and C_2 = 60 pF. Calculate the equivalent capacitance seen by the voltage source.

(a) $$C_{\text{eq}} = C_1 + C_2 = 20\text{ pF} + 60\text{ pF} = 80\text{ pF}$$

(b) $$C_{\text{eq}} = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2}} = \frac{1}{1/20\text{ pF} + 1/60\text{ pF}} = 15\text{ pF}$$

■

Self-Exercise 1.19

Calculate the terminal equivalent capacitance for the circuits in Figure 1.29.

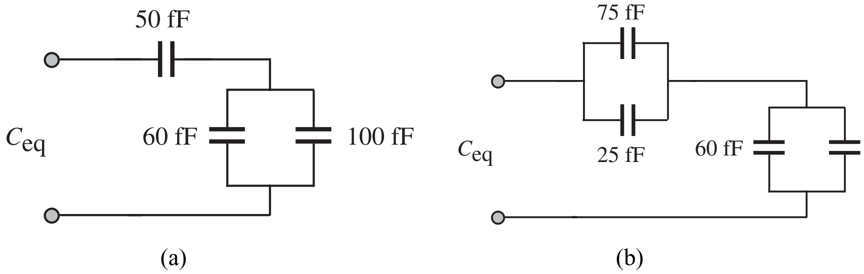


Figure 1.29.

1.3.2 Capacitor Voltage Dividers

There are open circuit defect situations in CMOS circuits in which capacitors couple voltages to otherwise unconnected nodes. This simple connection is a capacitance voltage di-

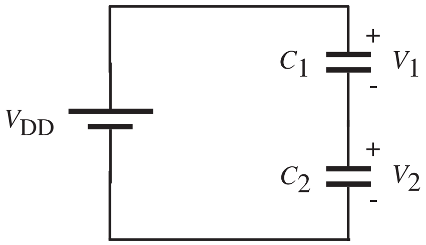


Figure 1.30. Capacitance voltage divider.

vider circuit (Figure 1.30). The voltage across each capacitor is a fraction of the total voltage V_{DD} across both terminals.

■ EXAMPLE 1.4

Derive the relation between the voltage across each capacitor C_1 and C_2 in Figure 1.30 to the terminal voltage V_{DD}.

The charge across the plates of the series capacitors is equal so that $Q_1 = Q_2$. The capacitance relation $C = Q/V$ allows us to write

$$Q_1 = Q_2 \qquad C_1 V_1 = C_2 V_2$$

or

$$V_2 = \frac{C_1}{C_2} V_1$$

Since

$$V_{DD} = V_1 + V_2$$

then

$$V_2 = V_{DD} - V_1 = \frac{C_1}{C_2} V_1$$

Solve for

$$V_1 = \frac{C_2}{C_1 + C_2} V_{DD}$$

and get

$$V_2 = \frac{C_1}{C_1 + C_2} V_{DD}$$

The form of the capacitor divider is similar to the resistor voltage divider except the numerator term differs. ■

Self-Exercise 1.20

Solve for V_1 and V_2 in Figure 1.31.

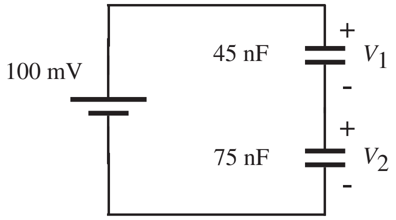


Figure 1.31.

Self-Exercise 1.21

If V_2 = 700 mV, what is the driving terminal voltage V_D in Figure 1.32?

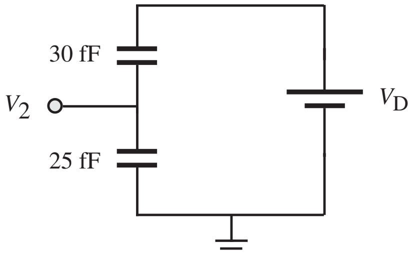


Figure 1.32.

1.3.3 Charging and Discharging Capacitors

So far we have discussed the behavior of circuits with capacitors in the steady state, i.e., when DC sources drive the circuit. In these cases, the analysis of the circuit is done assuming that it reached a stationary state. Conceptually, these cases are different from situations in which the circuit source makes a sudden transition, or a DC source is applied to a discharged capacitor through a switch. In these situations, there is a period of time during which the circuit is not in a stationary state but in a transient state. These cases are important in digital CMOS ICs, since node state changes in ICs are transient states determining the timing and power characteristics of the circuit. We will analyze charge and discharge of capacitors with an example.

■ EXAMPLE 1.5

In the circuit of Figure 1.33, draw the voltage and current evolution at the capacitor with time starting at $t = 0$ when the switch is closed. Assume $V_{\text{in}} = 5$ V and that the capacitor is initially at 0 V.

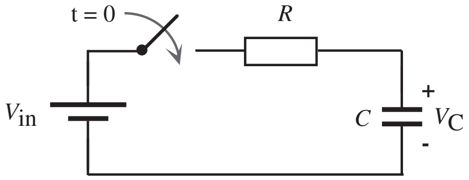


Figure 1.33.

The Kirchoff laws for current and voltage can be applied to circuits with capacitors as we did with resistors. Thus, once the switch is closed, the KVL must follow at any time:

$$V_{\text{in}} = V_R + V_C$$

The Kirchoff current law applied to this circuit states that the current through the resistor must be equal to the current through the capacitor, or

$$\frac{V_R}{R} = C\frac{dV_C}{dt}$$

Using the KVL equation, we can express the voltage across the resistor in terms of the voltage across the capacitor, obtaining

$$\frac{V_{\text{in}} - V_C}{R} = C\frac{dV_C}{dt}$$

This equation relates the input voltage to the voltage at the capacitor. The solution gives the time evolution of the voltage across the capacitor,

$$V_C = V_{\text{in}}(1 - e^{-t/RC})$$

The current through the capacitor is

$$I_C = I_R = \frac{V_{\text{in}} - V_C}{R}$$

$$I_C = \frac{V_{\text{in}}}{R}e^{-t/RC}$$

At $t = 0$, the capacitor voltage is zero (it is discharged) and the current is maximum (the voltage drop at the resistor is maximum), whereas in DC (for $t \to \infty$) the capacitor voltage is equal to the source voltage and the current is zero. This example shows that the voltage evolution is exponential when charging a capaci-

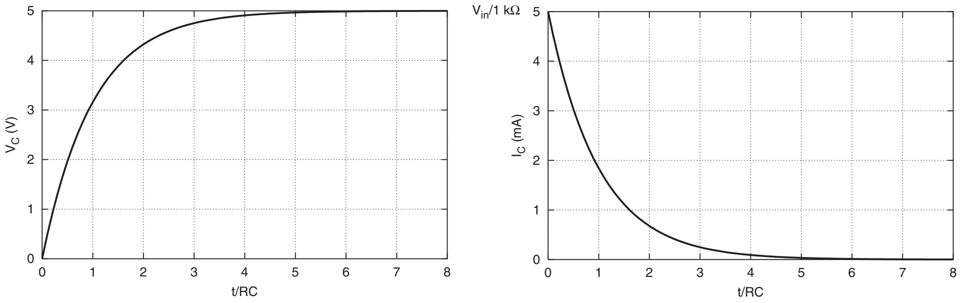


Figure 1.34.

tor through a resistor. The time constant is defined for $t = RC$, that is, the time required to charge the capacitor to $(1 - e^{-1})$ of its final value, or 63%. This means that the larger the value of the resistor or capacitor, the longer it takes to charge/discharge it (Figure 1.34). ■

1.4 DIODES

A circuit analysis of the semiconductor diode is presented below; later chapters discuss its physics and role in transistor construction. Diodes do not act like resistors; they are nonlinear. Diodes pass significant current at one voltage polarity and near zero current for the opposite polarity. A typical diode nonlinear current–voltage relation is shown in Figure 1.35(a) and its circuit symbol in Figure 1.35(b). The positive terminal is called the anode, and the negative one is called the cathode. The diode equation is

$$I_D = I_S(e^{\frac{qV_D}{kT}} - 1) \tag{1.23}$$

where k is the Boltzmann constant ($k = 1.38 \times 10^{-23}$ J/K), q is the charge of the electron ($q = 1.6 \times 10^{-19}$ C), and I_S is the reverse biased current. The quantity kT/q is called the thermal voltage (V_T) whose value is 0.0259 V at T = 300 K; usually, we use $V_T = 26$ mV at that

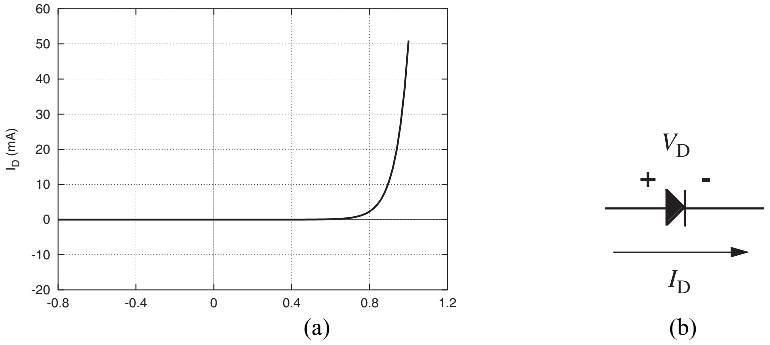


Figure 1.35. (a) Diode I–V characteristics and (b) symbol.

temperature. When the diode applied voltage is positive and well beyond the thermal voltage ($V_D \gg V_T = kT/q$), Equation (1.23) becomes

$$I_D = I_S e^{\frac{qV_D}{kT}} \tag{1.24}$$

The voltage across the diode can be solved from Equation (1.23) as

$$V_D = \frac{kT}{q} \ln\left(\frac{I_D}{I_S} + 1\right) \tag{1.25}$$

For forward bias applications $I_D/I_S \gg 1$ and this reduces to

$$V_D = \frac{kT}{q} \ln \frac{I_D}{I_S} \tag{1.26}$$

Self-Exercise 1.22

(a) Calculate the forward diode voltage if $T = 25°\text{C}$, $I_D = 200$ nA, and $I_S = 1$ nA. Compute from Equation (1.25). (b) At what current will the voltage drop be 400 mV?

Diode Equations (1.23)–(1.26) are useful in their pure form only at the temperature at which I_S was measured. These equations predict that I_D will exponentially drop as temperature rises which is not so. I_S is more temperature-sensitive than the temperature exponential and doubles for about every 10°C rise. The result is that diode current markedly increases as temperature rises.

1.4.1 Diode Resistor Circuits

Figure 1.36 shows a circuit that can be solved for all currents and node element voltages if we know the reverse bias saturation current I_S.

■ **EXAMPLE 1.6**

If $I_S = 10$ nA at room temperature, what is the voltage across the diode in Figure 1.36 and what is I_D? Let $kT/q = 26$ mV.

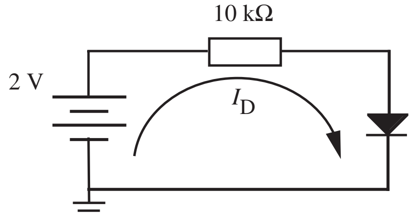


Figure 1.36. Forward-biased diode analysis.

Write KVL using the diode voltage expression:

$$2\text{ V} = I_D(10\text{ k}\Omega) + (26\text{ mV})\ln\left(\frac{I_D}{I_S} + 1\right)$$

This equation has one unknown (I_D), but it is difficult to solve analytically, so an iterative method is easiest. Values of I_D are substituted into the equation, and the value that balances the LHS and RHS is a close approximation. A starting point for I_D can be estimated from the upper bound on I_D. If $V_D = 0$, then $I_D = 2\text{ V}/10\text{ k}\Omega = 200\ \mu\text{A}$. I_D cannot be larger than 200 μA. A close solution is $I_D = 175\ \mu\text{A}$.

The diode voltage is

$$V_D = \frac{kT}{q}\ln\frac{I_D}{I_S}$$

$$= 26\text{ mV} \times \ln\frac{175\text{ uA}}{10\text{ nA}} = 244.2\text{ mV}$$

■

Self-Exercise 1.23

Estimate I_D and V_D in Figure 1.37 for $I_S = 1$ nA.

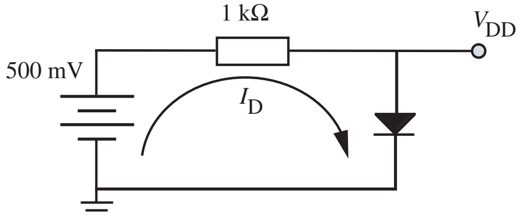


Figure 1.37.

■ EXAMPLE 1.7

Figure 1.38 shows two circuits with the diode cathode connected to the positive terminal of a power supply ($I_S = 100$ nA). What is V_0 in both circuits?

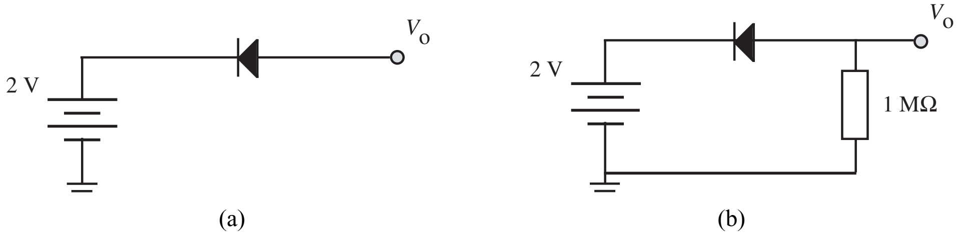


Figure 1.38.

Figure 1.38(a) has a floating node at V_0 so there is no current in the diode. Since $I_D = 0$, the diode voltage drop $V_D = 0$ and

$$V_0 = V_D + 2\ \text{V} = 2\ \text{V}$$

Figure 1.38(b) shows a current path to ground. The diode is reversed-biased and $I_{BB} = -I_D = 100$ nA. Then

$$V_0 = I_{BB} \times 1\ \text{M}\Omega = 100\ \text{nA} \times 1\ \text{M}\Omega = 100\ \text{mV}$$

■

Both problems in Example 1.7 can be analyzed using Equations (1.23) to (1.26) or observing the process in the I–V curve of Figure 1.35. In Figure 1.38(a), the operating point is at the origin. In Figure 1.38(b), it has moved to the left of the origin.

Self-Exercise 1.24

The circuit in Figure 1.39 is similar to IC protection circuits connected to the input pins of an integrated circuit. The diodes protect the logic circuit block when input pin (pad) voltages are accidentally higher than the power supply voltage (V_{DD}) or lower than the ground voltage. If $V_{PAD} > 5$ V, then diode D_2 turns on and bleeds charge away from the input pin. The same process occurs through diode D_1 if the input pad voltage becomes less than ground (0 V). An integrated circuit tester evaluates the diodes by forcing current (100 μA) and measuring the voltage. If the protection circuit is damaged, an abnormal voltage is usually read at the damaged pin.

(a) If diode reverse bias saturation current is $I_S = 100$ nA, what is the expected input voltage measured if the diodes are good and R_1 and R_2 are small? Apply ±100 μA to assess both diodes.

(b) If the upper diode has a dead short across it, what is V_{IN} when the test examines the upper diode?

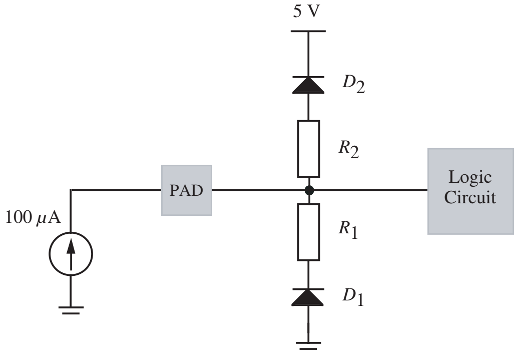


Figure 1.39.

Self-Exercise 1.25

Calculate V_0 and V_{D1} in Figure 1.40, where $I_S = 100$ μA and $T = 25$ °C.

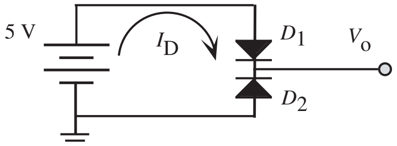


Figure 1.40.

1.4.2 Diode Resistance

Although diodes do not obey Ohm's law, a small signal variation in the forward bias can define a resistance from the slope of the *I–V* curve. The distinction with linear elements is important as we cannot simply divide a diode DC voltage by its DC current. That result is meaningless.

The diode curve is repeated and enlarged in Figure 1.41. If a small signal variation $v(t)$ is applied in addition to the DC operating voltage V_{DC}, then the exponential current/voltage characteristic can be approximated to a line [given that $v(t)$ is small enough] and an equivalent resistance for that bias operation can be defined. Bias point 1 has a current change with voltage that is larger than that of bias point 2; therefore, the dynamic resistance of the diode is smaller at bias point 1. Note that this resistance value depends strongly on the operating voltage bias value. Each point on the curve has a slope in the forward bias. Dividing the DC voltage by the current, V_1/I_1, is not the same as V_2/I_2. Therefore, these DC relations are meaningless. However, dynamic resistance concepts are important in certain transistor applications.

This concept is seen in manipulation of the diode equation, where the forward-biased dynamic resistance r_d is

$$\frac{1}{r_d} = \frac{dI_D}{dV_D} = \frac{d[I_S e^{qV_D/kT}]}{dV_D} = \frac{qI_S e^{qV_D/kT}}{kT} = \frac{qI_D}{kT} \tag{1.27}$$

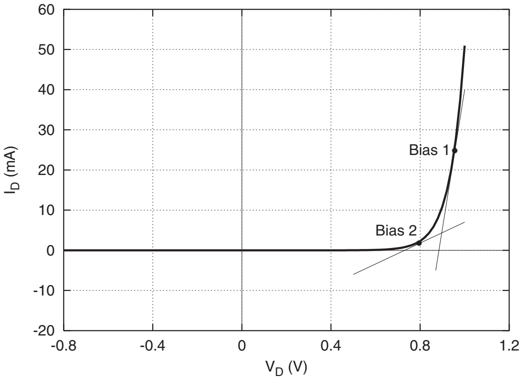


Figure 1.41.

The diode dynamic resistance is

$$r_d = \frac{dV_\mathrm{D}}{dI_\mathrm{D}} = \frac{kT}{qI_\mathrm{D}} \tag{1.28}$$

or at room temperature

$$r_d \approx \frac{26\ \mathrm{mV}}{I_D} \tag{1.29}$$

Self-Exercise 1.26

Find the diode dynamic resistance at room temperature for $I_\mathrm{D} = 1\ \mu\mathrm{A}$, 100 μA, 1 mA, and 10 mA.

How do we use the concept of dynamic diode resistance? A diode can be biased at a DC current, and small changes about that operating point have a resistance. A small sinusoid voltage (v_D) causes a diode current (i_D) change equal to $v_\mathrm{D}/r_\mathrm{D}$. The other important point is that you cannot simply divide DC terminal voltage by DC terminal current to calculate resistance. This is true for diodes and also for transistors, as will be seen later.

1.5 SUMMARY

This chapter introduced the basic analysis of circuits with power supplies, resistors, capacitors, and diodes. Kirchhoff's current and voltage laws were combined with Ohm's law to calculate node voltages and element currents for a variety of circuits. The technique of solving for currents and voltages by inspection is a powerful one because of the rapid insight into the nature of circuits it provides. Finally, the section on diodes illustrated analysis with a nonlinear element. The exercises at the end of the chapter should provide sufficient drill to prepare for subsequent chapters, which will introduce the MOSFET transistor and its simple configurations.

BIBLIOGRAPHY

1. R. C. Dorf and J. A. Svoboda, *Introduction to Electric Circuits,* 4th ed., Wiley, 1998.
2. D. E. Johnson, J. R. Johnson, J. L. Hilburn, and P. Scott, *Electric Circuit Analysis,* 3rd ed., Prentice-Hall, 1989.
3. J. W. Nilsson and S. A. Riedel, *Electric Circuits,* 6th ed., Prentice-Hall, 2000.
4. A. J. Rosa and R. E. Thomas, *The Analysis and Design of Linear Circuits,* 4th ed., Wiley, 2003.

EXERCISES

1.1. Write the shorthand expression for R_eq at the open terminals in Figure 1.42.

1.2. Write the shorthand expression for R_eq at the open terminals in Figure 1.43.

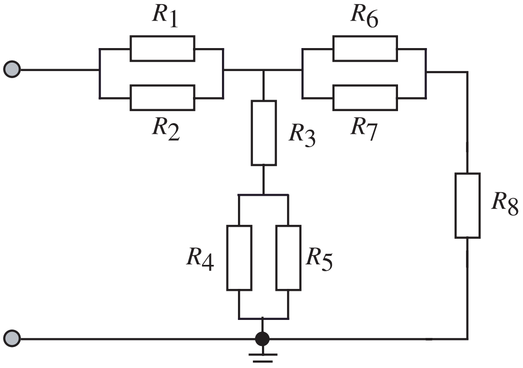


Figure 1.42.

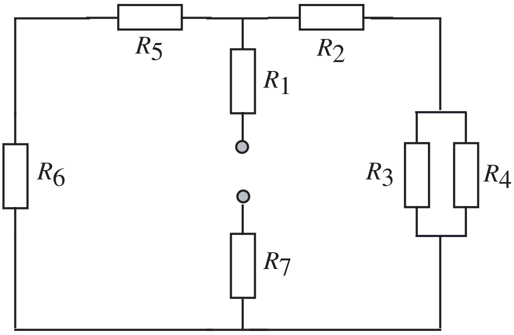


Figure 1.43.

1.3. For the circuit in Figure 1.44, (a) calculate V_0; (b) calculate I_{2M}.

1.4. Calculate V_0 by first writing a voltage divider expression and then solving for V_0 (Figure 1.45a and b).

1.5. Write the shorthand notation for current I_2 in resistor R_2 in Figure 1.46 as a function of driving current I.

1.6. For the circuit in Figure 1.47, (a) solve for V_0 using a voltage divider expression; (b) solve for I_{2K}; (c) solve for I_{900}.

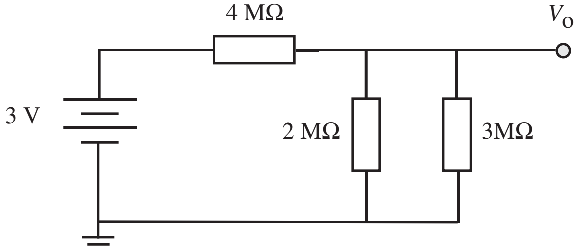


Figure 1.44.

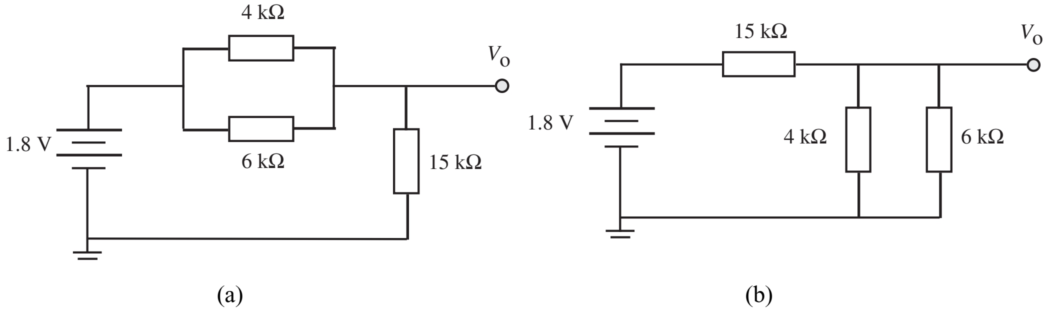


Figure 1.45.

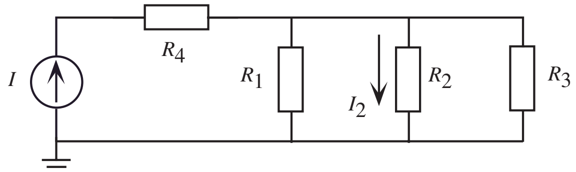


Figure 1.46.

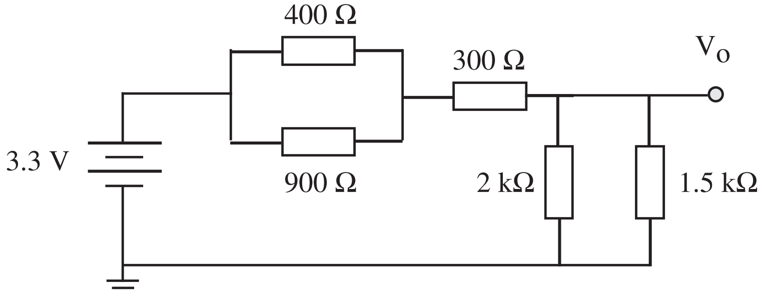


Figure 1.47.

1.7. Use the circuit analysis technique by inspection, and write the shorthand expression to calculate I_{2K} for Figure 1.48.

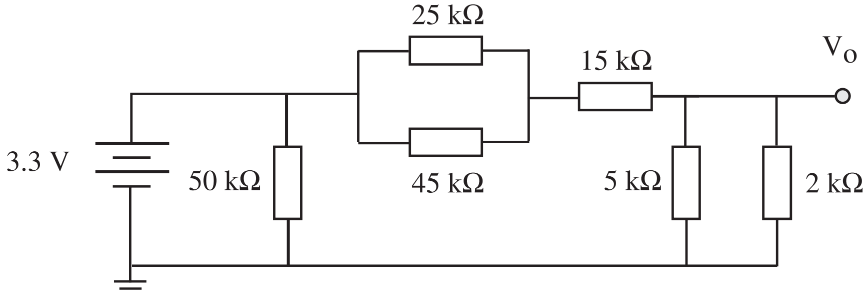


Figure 1.48.

1.8. Given the circuit in Figure 1.49, (a) write the expression for I_{450} and solve; (b) write the expression for V_{800}; (c) show that $I_{800} + I_{400} = 2$ mA.

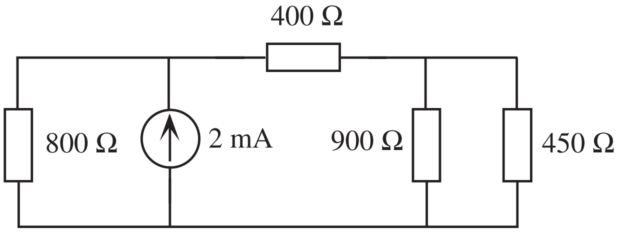


Figure 1.49.

1.9. Find I_{6k} in Figure 1.50. *Hint:* when we have two power supplies and a linear (resistive) network, we solve in three steps.

1. Set one power supply to 0 V and calculate current in the 6 kΩ resistor from the nonzero power supply.
2. Reverse the role and recalculate I_{6k}.
3. The final answer is the sum of the two currents.

This is known as the superposition theorem and can be applied only for linear elements.

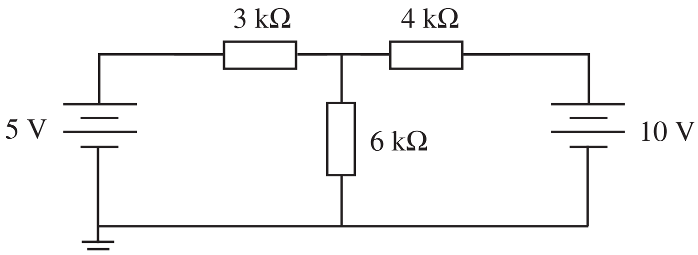


Figure 1.50.

1.10. Find the equivalent capacitance at the input nodes in Figure 1.51.

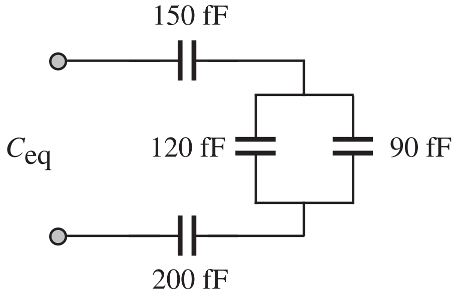


Figure 1.51.

1.11. Find C_1 in Figure 1.52.

1.12. Solve for I_D and V_D in Figure 1.53, where the diode has the value $I_S = 1$ μA.

1.13. Calculate V_0 in Figure 1.54, given that the reverse-bias saturation current $I_S = 1$ nA, and you are at room temperature.

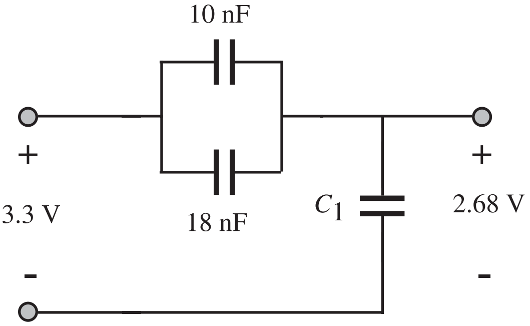


Figure 1.52.

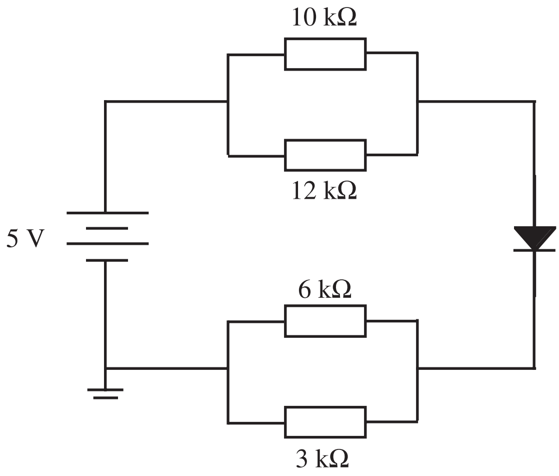


Figure 1.53.

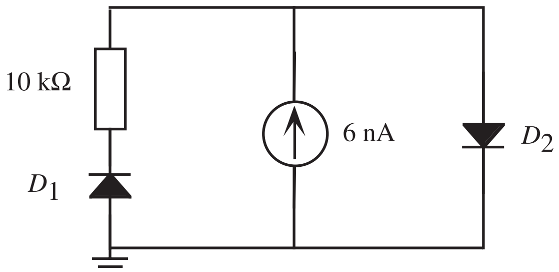


Figure 1.54.

1.14. Diode D_1 in Figure 1.55 has a reverse-bias saturation current of $I_{01} = 1$ nA, and diode D_2 has $I_{02} = 4$ nA. At room temperature, what is V_0?

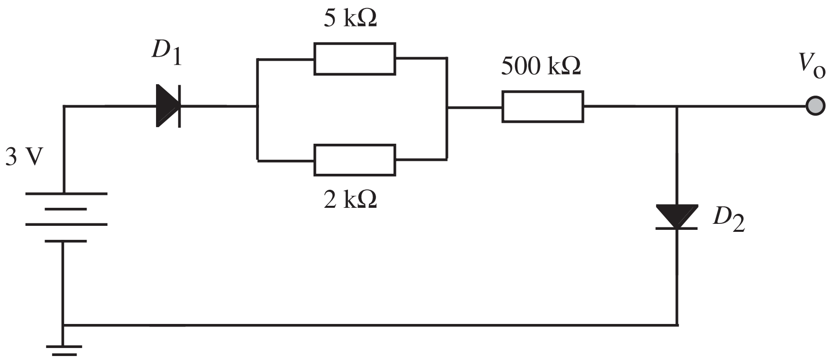


Figure 1.55.

1.15. Calculate the voltage across the diodes in Figure 1.56, given that the reverse-bias saturation current in D_1 is $I_{01} = 175$ nA and $I_{02} = 100$ nA.

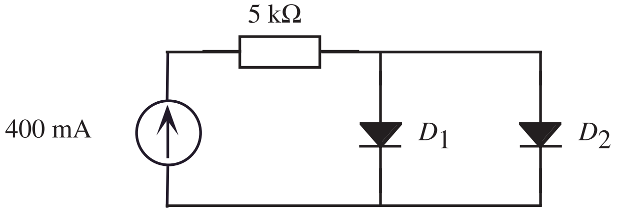


Figure 1.56.

CHAPTER 2

SEMICONDUCTOR PHYSICS

2.1 SEMICONDUCTOR FUNDAMENTALS

2.1.1 Metals, Insulators, and Semiconductors

Materials are classified by their physical properties. Conductivity measures the amount of current through a material when a voltage is applied, and materials are classified into three conductivity groups: metals, insulators, and semiconductors. Metals present practically no resistance to carrier flow, whereas insulators allow virtually no electrical carrier flow under an applied voltage. Semiconductors are unique, and can behave as conductors or insulators. This chapter introduces the principles of semiconductor physics, using only a few equations and numerical examples. Emphasis is on providing a common understanding of these principles and a basic description of how the devices work. The language and visual and mathematical models of semiconductor physics permeate CMOS manufacturing.

The classic explanation for conduction differences between these materials uses the energy band model of solids that derives from quantum mechanics. Neils Bohr found that electrons in an atom could not have arbitrary energy values, but had defined, discrete (quantum) energy values. A basic principle of quantum mechanics is that energy is distributed in quantum packets, and cannot take continuous values. Electrons orbit at discrete distances from the nucleus. Figure 2.1(a) shows the allowed energy levels of a hydrogen atom electron. These energy levels are those that the electron can take that are discrete. The *s* and *p* energy level symbols are taken from quantum mechanical convention.

In a two-atom system, each energy level in the single atom system splits into two sublevels, as shown in Figure 2.1(b). When more atoms are added to construct a crystalline solid, the energy levels successively split, leading to the picture in Figure 2.1(c), where

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

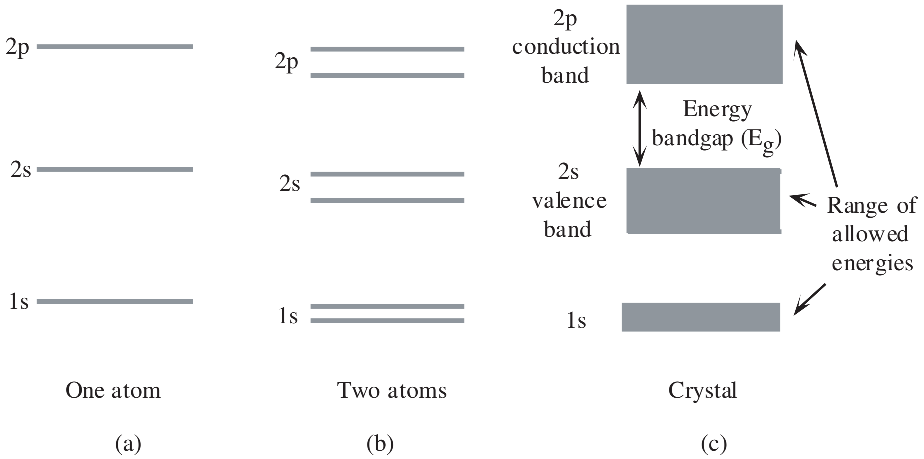


Figure 2.1. (a) Energy levels in a single atom, (b) two atoms, and (c) a solid.

energy bands separated by gaps of forbidden energies (called band gaps) replace single energy levels. The band gap width depends on the type of atom used to build the solid, and it determines the conductive properties of the material.

Energy bands have different conduction properties. The outermost energy band is called the conduction band, and the next-lower one is called the valence band or outer shell. An electron having an energy corresponding to the conduction band is not tied to any atom, and can move "freely" through the solid. Such an electron contributes to current when a voltage is applied. An electron in the valence band has an energy that is attached to an atom of the solid, and is not "free" to move within the solid when a voltage is applied.

Energy bands help us more easily understand the conductive properties of different materials. Figure 2.2(a) shows the energy bands of a metal; the lowest energy value of the conduction band is below the maximum energy of the valence band. This means that the

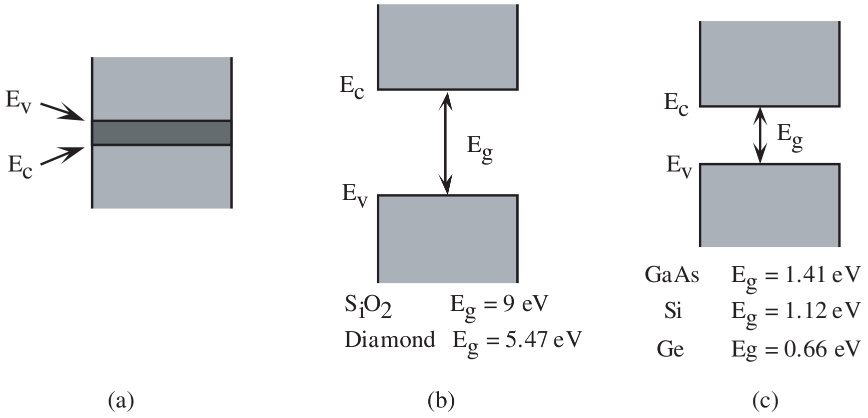


Figure 2.2. Energy bands in solids: (a) metal, (b) insulator, (c) semiconductor.

metal conduction band has an abundance of electrons that are available for conduction. It takes very little additional energy to move an electron from the valence to the conduction band since the bands are merged.

The energy band structure of an insulator is shown in Figure 2.2(b). Insulators have a large energy gap between the valence and conduction bands. The thermal energy needed for an electron to go from the valence to the conduction band is so high that only a few electrons within the material can acquire such an energy and jump over the gap. A voltage applied to the material will cause almost no current since virtually all electrons are tied to atoms in the valence band.

Semiconductors are the third class of conducting material, and show an intermediate behavior. The valence and conduction bands are not merged, but the energy gap is small enough so that some electrons are energetic enough to jump across it. The energy needed for electrons to jump across the gap comes from the ambient temperature or photon energy. We will deal with thermal energy since most integrated circuits are sealed and admit no light from the environment. The process of gaining enough thermal energy and jumping from the valence to the conduction band is inherently statistical. This means that electrons in a solid are continuously moving up and down between the valence and conduction bands. However, at a given temperature there is a population of electrons in the conduction band that contribute to the current.

Since the energy used by an electron to jump the gap is thermal, the population of electrons in the conduction band depends on the temperature. At absolute zero temperature, there is no thermal energy in a pure semiconductor, so that no electron has enough energy to jump across the gap. As temperature increases, the number of conducting electrons increases.

The differences in gap energies between insulators and semiconductors are related to how electrons are arranged within atoms. Electrons are grouped into layers around the atomic nucleus, and electrons in the internal layers cannot be separated from the nucleus. Only electrons from the outside valence layer may jump from their bounded valence state to the conducting state (free from the attractive forces of the atomic nucleus). Atoms of conducting materials have several layers of orbiting electrons. The number of electrons required to fill a given layer remains constant and independent of the atomic element. An atom having all layers completely filled will have all electrons (even those at the outmost layer) strongly “tied” to the nucleus. A large amount of energy is needed to break such a layer and take one electron out of the atom. These atoms are known as noble gases, since they do not react with other elements, as their electrons are closely packed. Atoms in which the external layer is not “closed” (more electrons are required to completely fill such layers) have electrons more lightly attached to the atom. As a result, only a small amount of energy is needed to separate an external electron from the atom. This is the case in metals; the outside electrons belong to the solid instead of being attached to some nucleus.

2.1.2 Carriers in Semiconductors: Electrons and Holes

The only contribution to current in a metal is from electrons in the conducting band. In contrast, semiconductor current has two contributions: one from electrons in the conduction band and the other from electron vacancies in the valence band caused by electrons that jumped into the conduction band. The vacancy of an electron in the valence band leaves an empty local charge space of a value equal to the electron charge, but of opposite sign. Therefore, it has a net charge $+q$ (q being the charge value of an electron; $q = -1.6 \times 10^{-19}$ Coulomb).

The most common semiconductor material for ICs is silicon, that has four electrons in its outer energy band (Figure 2.3(a)). When silicon is crystalline, each electron of the valence layer is shared with one electron of a neighbor atom so that by sharing, each Si atom has eight outer shell electrons. If a valence electron gains enough thermal energy to jump into the conduction band, it leaves a vacancy position. Such a position is available to another valence electron to move into and leave a vacancy at its original site. This process can now be repeated for a third valence electron moving into this last vacancy, and so on. It is important to note that this process does not require the moving valence electron to go to the conduction band to move to such a vacancy. The electron vacancy can be seen as a "particle" of positive charge that moves in the opposite direction to the valence electron. Such a virtual particle with associated charge $+q$ is called a "hole." When silicon is constructed as a crystal, it behaves as a semiconductor.

When the semiconductor is in equilibrium and there is no external electromagnetic field or temperature gradient, then electrons and holes move randomly in space, and no net current is observed. When an electric field is applied, the hole movement is not random, but drifts in the same direction as the field. This gives a net current contribution from holes, in addition to the current contribution from electrons in the conduction band. These dual conduction mechanisms in solids are detailed in the next section.

In a pure silicon material, electrons and holes are created in pairs. The "creation" of a free electron jumping into the conduction band creates a hole in the valence band, whereas an electron dropping from the conduction to the valence band implies that a hole disappears. When an electron jumps from the valence to the conduction band, the process is called an electron–hole pair creation, and when an electron jumps back from the conduction band to the valence band, the process is referred to as electron–hole recombination, or simply recombination. Energy is needed for electron–hole pair formation, but in the opposite process, energy is released when an electron recombines with a hole. This is illustrated in Figure 2.4 for the energy band gap model of a semiconductor and its solid-state physical representation. It emphasizes that mobile carriers are electrons in the conduction band and holes in the valence band. The process of electron–hole creation and

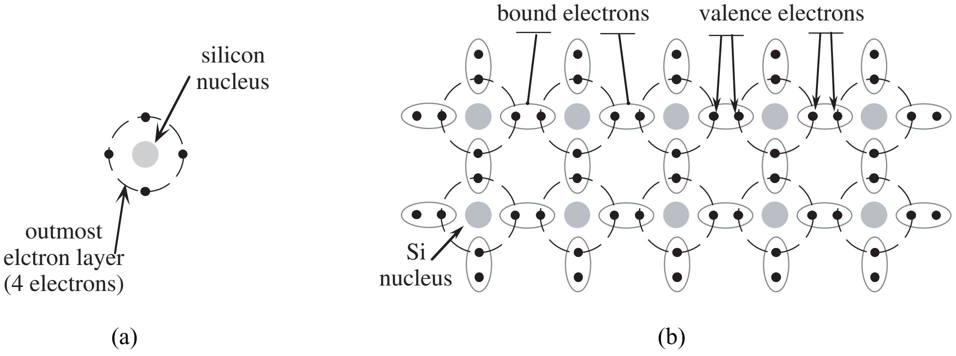


Figure 2.3. (a) Representation of a silicon atom with its four electrons at the outmost layer, and (b) picture of the silicon structure in a crystalline solid.

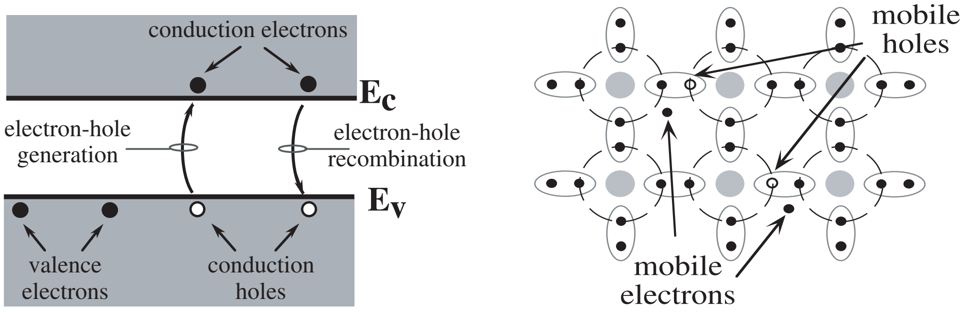


Figure 2.4. Electron-hole pair creation and recombination in the band gap model (left), and its representation in the solid (right).

recombination is inherently statistical—electron–hole pairs are continuously created and recombined in a semiconductor at a given temperature.

2.1.3 Determining Carrier Population*

We will refer to electrons as the carriers in the conduction band and holes as the carriers in the valence band. Electrons and holes are mobile carriers, and we must estimate their populations at each temperature that are statistically available in each band. Fermi–Dirac statistics accounts for the basic properties of electrons in solids, and gives us a mathematical model of the electron statistics for the concentration of carriers. Electrons have a duality property, acting as particles or as indistinguishable waves, and follow the Pauli principle of exclusion.† Since electrons are quantum elements, we describe the probability of an electron being at a given energy for a given temperature. The probability of an electron at a given energy level E is expressed with a probability function $f(E)$, which may be derived from the Fermi–Dirac statistics (see [2] or [1]), obtaining

$$f(E) = \frac{1}{1 + e^{(E - E_F)/kT}} \tag{2.1}$$

where k is the Boltzmann constant (1.38×10^{-23} J/K), T is temperature in Kelvin, and E_F is known as the Fermi energy level, an important parameter for semiconductors. The Fermi energy level is the energy at which the probability function equals 0.5 (this can be easily verified by substituting $E = E_F$ in the equation), or the energy level below which the probability function is 1 for $T = 0$ K. At absolute zero temperature, all possible energy levels are filled. If $f_e(E)$ is the probability function of an electron being at a given energy (E), then holes are the "dual" or complementary particles. The probability of a given hole being at such an energy is $f_h(E) = 1 - f_e(E)$.

Since we are describing electrons in a solid, we must account for the number of available energy states. For example, no state is available within the forbidden energy gap. The

*This subsection can be skipped without loss of continuity.

†The Pauli principle of exclusion states that two quantum objects (electrons in this case) cannot occupy the same energy levels. As a result, the maximum number of electrons per layer in a given atom is fixed.

formulation of the number of states $N(E)$ available for an electron at a given energy in a solid allows us to find the concentration of electrons n having an energy interval dE as

$$n = f(E)N(E)dE \tag{2.2}$$

The electron concentration within the conduction band is found by integrating such a concentration from $E = E_C$ to $E \rightarrow \infty$. Thus,

$$n_0 = \int_{E_C}^{\infty} f(E)N(E)dE \tag{2.3}$$

where n_0 stands for the number of carriers in equilibrium. Similarly, the hole concentration is

$$h_0 = \int_{-\infty}^{E_V} [1 - f(E)]N(E)dE \tag{2.4}$$

A detailed derivation for $N(E)$ and Equations (2.3) and (2.4) are beyond the scope of this work. For a detailed analysis we refer to any of the books cited at the end of the chapter.

2.2 INTRINSIC AND EXTRINSIC SEMICONDUCTORS

The previous concepts applied to "pure" semiconductors, in which all atoms are of the same type. This is referred to as an *intrinsic semiconductor,* implying that the number of electrons is equal to the number of holes, since they are generated in pairs. *Extrinsic semiconductors* are created by intentionally adding impurities* to the semiconductor to increase the concentration of one carrier type (electrons or holes) without increasing the concentration of the other, thus breaking the symmetry between the number of electrons and holes. The intentional substitution of a silicon atom by another element is called *doping.* This process depends on the type of impurity added and the number of impurities introduced in a unit volume. There are *n*-type impurities that increase electron concentration and *p*-type impurities that increase hole concentration. The number of impurities (atoms/unit volume) injected into the solid is much less than the number of silicon atoms, and the crystalline structure of the semiconductor is not globally disturbed.

2.2.1 *n*-Type Semiconductors

Silicon has four electrons in its outer layer that form bonds with neighbor atoms. We can increase the electron concentration without changing the hole concentration by replacing some silicon atoms with Periodic Table Group V atoms having five electrons in their external layer (arsenic, phosphorus, or antimony). Four of the five external electrons of the substituting atom create bonds with the neighboring silicon atoms, while the fifth one is almost free to move within the solid (Figure 2.5). At room temperature, the thermal energy is enough to activate such a fifth electron and move it into the conduction band.

*Any material, no matter what its quality, always has unintended impurities. The effect of such impurities can be neglected if they are kept to a minimum. Additionally, crystalline solids are not perfect crystals and may have some "irregularities" that impact the energy band structure.

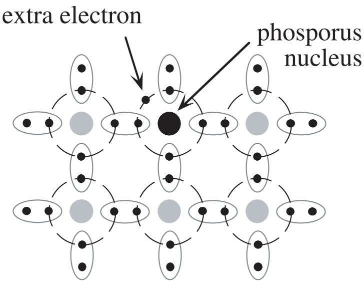


Figure 2.5. Adding a donor atom creates a mobile electron without creating a hole.

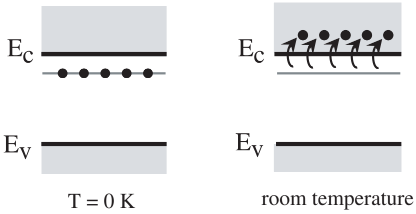


Figure 2.6. Picture of the donor effect in the band-gap energy model with change in temperature.

The electron jumps into the conducting energy band without creating a hole, so no electron–hole pairs are generated, only free electrons. In this case, each impurity atom is called a *donor,* since it implies that an extra electron is donated to the semiconductor. Silicon is a Group IV atom in the Periodic Table and donor atoms come from the Group V atoms.

When extrinsic silicon is doped with donors, the number of conducting electrons is approximately equal to the number of donor atoms injected (N_D) plus some electrons coming from electron–hole pair creation. When donor concentration greatly exceeds the normal intrinsic population of carriers, then the conducting electron concentration is essentially that of the donor concentration. By adding a specific concentration of donors, the population of electrons can be made significantly higher than that of holes. An extrinsic semiconductor doped with donor impurities is called an *n*-type semiconductor.

When a considerable number of donor impurities are added (10^{15}–10^{17} atoms per cm^3), the effect on the bandgap model is creation of an allowed energy level within the gap close to the E_c (Figure 2.6). At zero Kelvin temperature, all electrons are at this energy level and are not mobile within the solid. At room temperature, the energy required to jump into the conduction band is small, and all donor electrons are ionized and remain in the conduction band.

Doping a semiconductor does not increase its net charge, since the negative q charge excess of the fifth electron with respect to the "replaced" silicon atom is balanced by the atomic number* of the donor impurity (5 instead of 4 for silicon) giving an extra positive charge that compensates the electron charge.

*The atomic number is the number of protons of a neutral (nonionized) atom.

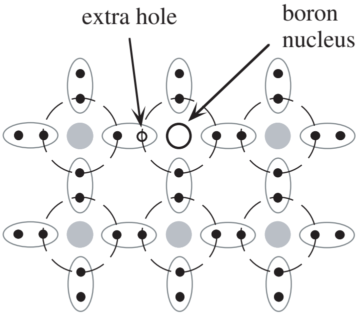


Figure 2.7. Adding acceptor atoms creates a mobile hole without creating an electron.

2.2.2 *p*-Type Semiconductors

We can increase the number of holes without increasing the number of electrons by replacing silicon atoms with elements from Group III in the Periodic Table that have three electrons in the outermost layer. The three electrons of the impurity bond with three of the neighbor silicon atoms, and the fourth bond is missing (Figure 2.7). Impurities having three electrons in the outermost layer are called *acceptors* since a vacancy is created that can accept free electrons. Acceptor doping in the energy-band model creates an energy level close to the silicon valence band, as shown in Figure 2.8. At room temperature, electrons in the valence band have enough energy to jump to this level. This creates a hole in the silicon without injecting electrons into the conduction band. At a given temperature, the concentration of holes in a semiconductor in which a number N_A of acceptor impurities were introduced will be N_A plus some of the holes contributed by the electron–hole pairs created at that temperature. Boron is a Group III atom and is the most popular of acceptor doping atoms.

2.3 CARRIER TRANSPORT IN SEMICONDUCTORS

The movement of electrons and holes in a semiconductor is called carrier transport. Computation of carrier transport requires knowing the carrier concentration calculated from the Fermi function and its subsequent derivations. It also requires the laws governing movement of carriers within the solid. Movement of electrons and holes within a semi-

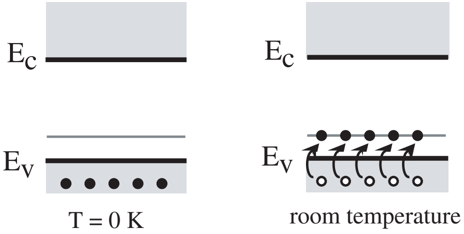


Figure 2.8. Effect of acceptor doping in the band-gap model.

conductor is different from carriers traveling in free space since collisions of carriers with the lattice impact their mobility. Temperature also plays an important role in carrier transport since it determines the carrier population and also affects carrier movement within the solid because of atomic thermal agitation. Although carriers are in constant motion within a solid, an unbiased semiconductor will not register a net current since carrier movement is random and no preferred direction is collectively chosen. A force is required for net movement of a charge to occur. The two main carrier movement mechanisms in solids are drift and diffusion.

2.3.1 Drift Current

Drift is carrier movement due to an electric or magnetic field. An electric field $\mathscr{E}$ applied to a semiconductor causes electrons in the conduction band to move in the direction opposite to the electric field, whereas holes in the valence band move in the same direction of this field. An electric field causes energy band bending, as shown in Figure 2.9(a) and the carrier movement indicated in Figure 2.9(b). Electron and hole current densities (electron hopping in the valence band) have the same direction and both contribute to current.

At relatively low electric fields, the velocity acquired by electrons and holes has a linear dependence on the applied external field. This dependence is maintained until a certain electric field is reached, and then the carrier velocity saturates. Velocity saturation occurs because carriers collide with the lattice atoms of the crystal. For electric fields beyond 10^5 V/cm, electron and hole velocities reach a maximum value in pure silicon at room temperature and no longer depend on the applied field. Modern deep-submicron transistors show velocity saturation during switching.

The relation between drift velocity and electric field for carriers is given by

$$v_d = \mu_0 \mathscr{E}\left[1 + \left(\frac{\mu_0 \mathscr{E}}{v_{\text{sat}}}\right)^{\beta}\right]^{-1/\beta} \tag{2.5}$$

where $\beta \approx 1$ for electrons, $\beta \approx 2$ for holes, μ_0 is the proportionality factor between the electric field $\mathscr{E}$ and the carrier velocity at low electric fields, and v_{sat} is the velocity saturation value reached for high electric fields. Note that this expression leads to $v_d = \mu_0 \mathscr{E}$ for low electric fields and $v_d = v_{\text{sat}}$ for large electric field strengths.

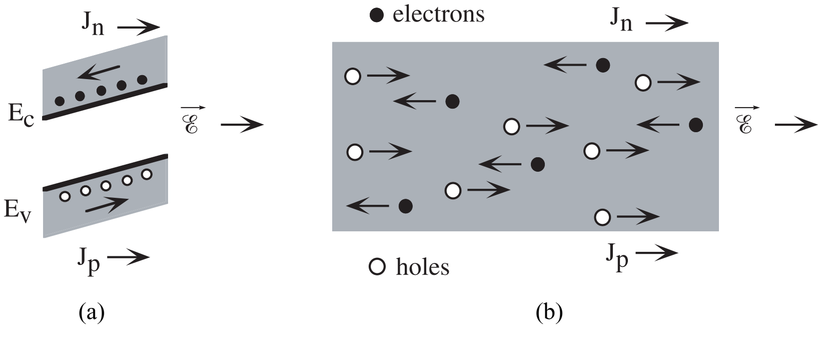


Figure 2.9. Band bending by the effect of an external applied field $\mathscr{E}$.

The current density of charge carriers J_d in an electric field is derived from a series of relations shown below, where v is the velocity of moving charge, N is the moving carrier concentration, q is carrier charge, and $\mathcal{E}$ is the electric field pushing the charge through a solid with mobility μ.

$$J_d = vqN = \mu q \mathcal{E} N \tag{2.6}$$

2.3.2 Diffusion Current

Diffusion is a thermal mechanism that moves particles from high-density macroscopic regions to low-density ones, so that in the final situation, the particle distribution in space is uniform. Electrons and holes diffusing in a solid are moving charged particles that create a diffusion current. Diffusion motion is described by Fick's law, stating that

$$J = -D \nabla N \tag{2.7}$$

where J is flux in particles/cm^2-sec, ∇N is the gradient of particles, and D is the diffusion coefficient. The equation states that the flux of particles is zero if there is a uniform distribution (i.e., ∇N is zero). Electrons and holes diffusing in the same direction give rise to opposite current densities since their charge signs are opposite. Expressions for electron and hole diffusion currents are

$$\begin{aligned} J_{n|\text{diff}} &= qD_n \nabla n \\ J_{p|\text{diff}} &= -qD_p \nabla p \end{aligned} \tag{2.8}$$

D_n and D_p are electron and hole diffusion coefficients that differ because electron and hole mobilities are different. The relation between the diffusion coefficient and the mobility is

$$\frac{D_x}{\mu_x} = \frac{kT}{q} \tag{2.9}$$

where x must be replaced by n for electrons and by p for holes. This equation is known as the Einstein relationship. Drift is the dominant charge transport mechanism in CMOS field-effect transistors, whereas diffusion plays a secondary role. Velocity saturation of electrons and holes occurs in the drift mechanism of all modern CMOS transistors.

2.4 THE *pn* JUNCTION

Diodes are simple semiconductor devices that are the building blocks of MOS transistors. Diodes have a junction formed by joining a p-type and n-type semiconductor. This is referred to as a pn junction. To understand pn junction properties, assume an ideal case in which two pieces of semiconductors with opposite doping are initially separated [Figure 2.10(a)] and then joined [Figure 2.10(b)]. The lattice structure is not lost at the joining surface, and the doping concentration has a sharp change from the left side (n-doped) to the right side (p-doped).

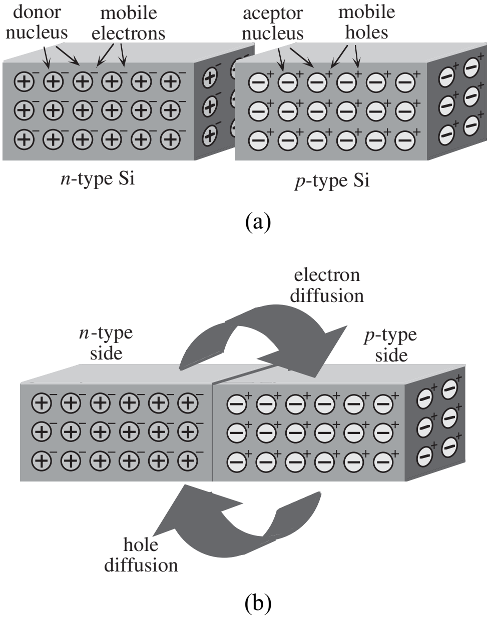


Figure 2.10. Two pieces of semiconductor materials with opposite doping: (a) separated and (b) joined.

Since the *n*-type and *p*-type semiconductor bars are in equilibrium, the total net charge in each bar is zero. At the instant when the semiconductor pieces join, there is a momentary abrupt change in electron and hole concentration at the joining surface. Strong concentration gradients exist for electrons on the *n*-side and holes on the *p*-side. This nonequilibrium condition exists for a short time during which electrons at the junction start to diffuse from the initial *n*-type bar into the *p*-type one, while holes close to the junction move away from the *p*-doped semiconductor into the *n*-type one. If electrons and holes were not charged particles, then this diffusing process would continue until electron and hole concentrations were uniform along the whole piece of joined semiconductors.

The reality is different, since electrons and holes are charged particles and their diffusion creates electric fields in the semiconductor at both sides of the junction. Carriers close to the junction are the first particles to diffuse away and recombine when they meet opposite carrier types on the opposite side of the junction. As a consequence of this carrier migration and recombination, all dopant atoms close to the joining surface are ionized. This creates a zone of net charge (Figure 2.11) around the junction (positive at the *n*-side and negative at the *p*-side) that induces an electric field pointing from the *n*-side to the *p*-side. Carriers moving by diffusion now “feel” the electric field as an opposing force when trying to diffuse. Their final motion depends on which conducting mechanism is stronger. As more carriers move by diffusion, more atoms are ionized, and the electric field strength increases. Finally, the net diffusion mechanism stops when the induced internal electric field (which increases with the number of carriers moving by diffusion) reaches a value such that its force exactly balances the force tending to diffuse carriers across the junction. The result is creation of a depletion region with all donors and acceptors ionized where the net charge and electric field are nonzero.

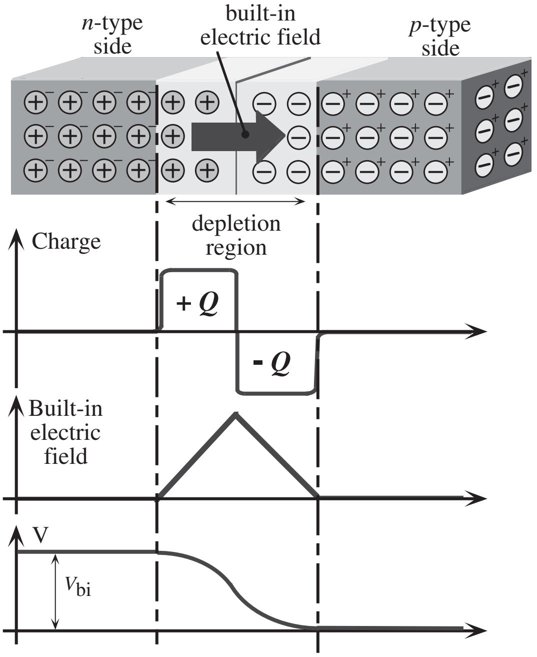


Figure 2.11. A diode in equilibrium, showing the charge, electric field, and potential internal distribution.

Figure 2.11 shows a picture of the diode junction in equilibrium (no external electric field, light, or temperature gradient are present) and the charge, electric field, and potential distribution at each point within the diode. Notice that in the regions where atoms are not ionized (out of the depletion region), there is charge neutrality, so that no electric field or potential drop exist. The net charge in the depletion region is positive in the *n*-type side and negative in the *p*-type side. As a result, the electric field increases while moving from the neutral regions to the junction site. The electric field distribution causes a voltage difference between the two oppositely doped regions that depends mainly on the doping levels. This is called the built-in junction potential (V_{bi}) shown at the bottom of Figure 2.11.

At equilibrium, a charge zone exists on both sides of the junction in which all donors and acceptor atoms are ionized. This zone is known as the depletion region or space-charge zone with a high electric field and a potential that increases when moving from the *p*-type zone to the *n*-type zone. Outside the boundaries of the space charge region within the semiconductor, the electric field, net charge, and potential gradient are all zero.

2.5 BIASING THE *pn* JUNCTION: I–V CHARACTERISTICS

The previous section showed that when two semiconductor bars of opposite doping are joined, a built-in electric field appears, preventing electrons from diffusing too far away from the *n*-side and holes from diffusing away from the *p*-side. Once the system is in equilibrium, no current exists since the junction is isolated with no external conducting path. We will now analyze the behavior of the junction when an external voltage is applied.

2.5.1 The *pn* Junction under Forward Bias

Assume that an external voltage source is connected to a diode, inducing more positive voltage at the *p*-side with respect to the *n*-side (Figure 2.12). The external voltage creates an electric field opposed to the built-in one, decreasing its strength. The built-in electric field is a barrier for electrons diffusing from the *n*-side to the *p*-side and for holes diffusing from the *p*-side to the *n*-side. An external reduction of the built-in field favors the diffusion process. Since many conducting electrons are on the *n*-side and many holes are on the *p*-side, then many electrons will diffuse into the *p*-side and many holes will diffuse into the *n*-side. This process causes a permanent current since electrons must be replaced at the *n*-type side through the voltage source, and holes are required at the *p*-type side. The larger the applied voltage bias, the higher the diode current.

2.5.2 The *pn* Junction under Reverse Bias

When the voltage applied to the diode is positive at the *n*-side with respect to the *p*-side, it induces an external electric field that increases the built-in junction electric field (Figure 2.13). This reduces electron diffusion from the *n*-side silicon into the *p*-side, and hole diffusion in the opposite direction. Only electrons or holes within the junction region itself are accelerated by the junction electric field and contribute to overall current. Since very few electrons are present in a *p*-type material, and few holes are in the *n*-type side, the reverse bias current is very small. In fact, this current, called the reverse-bias saturation current, arises from thermal generation of electron–hole pairs in the depletion region itself. The free electron and hole carriers are then rapidly swept out of the junction, forming the current at the diode terminals.

The device-level current/voltage characteristics of the diode were introduced in Chapter 1. The relation between the applied diode voltage V_D and the obtained current I_D is exponential and reproduced here:

$$I_D = I_S\left(e^{\frac{qV_D}{kT}} - 1\right) \tag{2.10}$$

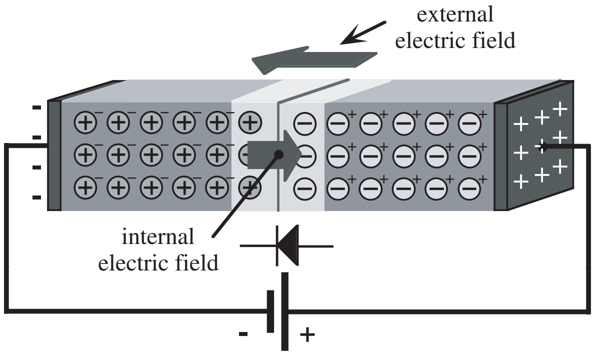


Figure 2.12. Forward-biased diode showing internal junction depletion-field reduction.

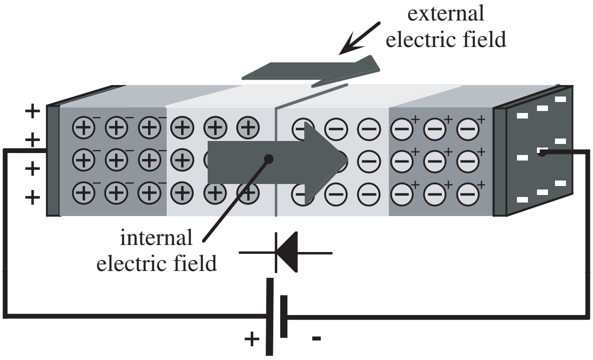


Figure 2.13. Reverse-biased diode showing internal junction depletion-field increase.

Remember that I_0 is called the reverse-bias saturation current. When V_D is large and negative, then $I_D = -I_0$. The I–V characteristic of the diode rapidly increases for positive diode voltages, whereas very little current is obtained for negative diode voltages.

2.6 PARASITICS IN THE DIODE

Diodes are useful as nonsymmetric components that allow current in one direction but not in the opposite. All diodes show a behavior that deviates from their ideality and this deviation can be modeled by so-called *parasitic elements.* Parasitic elements are undesired and can be resistance, capacitance, or inductance. The diode current leakage in the off-state (ideally not conducting) can be modeled by a high parasitic (undesired) resistance.

In CMOS technology, two parasitic diodes exist in each transistor. These diodes are always reverse-biased in ICs, and their main degradation effects at the circuit level are related to reverse current leakage or to delay through parasitic capacitors of the diode.

We saw that when a diode is reverse-biased, the external voltage increases the internal electric field strength. This widens the depletion region. The higher the reverse voltage, the wider the depletion region, and the larger the total net charge in this region. Conversely, a forward bias narrows the depletion region, and reduces the fixed charges across the *pn* junction. Therefore, a diode has an *internal capacitor* since there is a charge variation induced by a voltage variation. From Chapter 1, we know that the term capacitance is defined as the charge variation in a component due to voltage variation at its terminals, i.e.,

$$C = \frac{\partial Q}{\partial V} \tag{2.11}$$

The parasitic capacitor inherent to the *pn* junction is different from the passive or parallel-plate capacitors seen in Chapter 1 because the charge–voltage ratio varies with the applied voltage. Since the Q/V quotient is not constant when the applied voltage changes, a fixed capacitor value cannot be assigned to the parasitic capacitor.

The effect of the diode parasitic capacitor at the circuit level is significant, since it must be charged and discharged when a gate is switching. This contributes to circuit delay

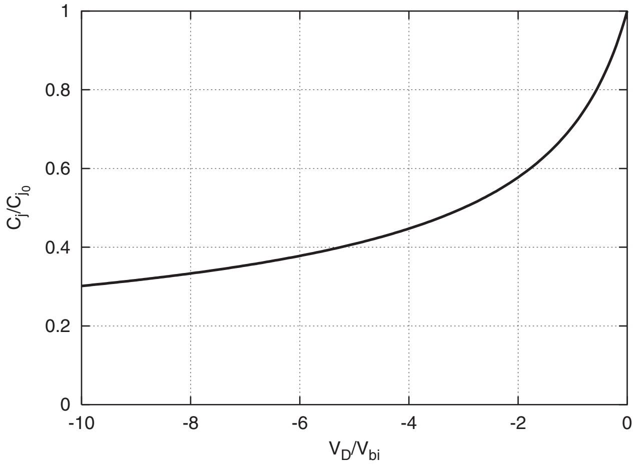


Figure 2.14. Plot of C_j versus applied voltage (V_D) for a diode.

and other secondary effects discussed later. In many cases, the exact dependence of the diode parasitic capacitor with the applied voltage is replaced by an approximate value. The value of the capacitor with the applied voltage is

$$C_j = \frac{C_{j0}}{\left(1 - \frac{V_\mathrm{D}}{V_\mathrm{bi}}\right)^{1/2}} \tag{2.12}$$

where C_{j0} is a constant depending on the *pn* doping values, fundamental constants of the silicon, the area of the surfaces being joined, and the built-in junction potential; V_D is the reverse applied voltage, and V_bi is the built-in junction potential. Figure 2.14 plots the capacitance normalized to C_{j0} with respect to the reverse applied voltage. Equation (2.12) breaks down in the diode forward-bias regions for large positive values of V_D. In the next chapter, it will be shown how this parasitic device affects device operation, and how it is modeled at the circuit level.

2.7 SUMMARY

The construction of integrated circuits ultimately depends on a strong physical base. Chapter 2 provides a brief introduction to these essential concepts. Although a few modeling equations were given, the purpose is qualitative understanding of the language and physical flow of semiconductor conduction, related doping properties, and diode characteristics. These ideas permeate later chapters.

BIBLIOGRAPHY

1. N. W. Ashcroft and N. D. Mermin, *Solid State Physics,* Saunders College, 1976.
2. C. Kittel, *Introduction to Solid State Physics,* 6th ed., Wiley, 1986

3. D. A. Neamen, *Semiconductor Physics and Devices—Basic Principles,* 3rd ed., McGraw-Hill, 2003.

4. R. F. Pierret, *Semiconductor Device Fundamentals,* Addison-Wesley, 1996.

EXERCISES

2.1. Discuss the following:
(a) What is the band-gap voltage E_g in the solid state model?
(b) Use the band-gap voltage to distinguish between metals, semiconductors, and insulators.

2.2. Metal electrical conduction is done by electrons. How does semiconductor conduction differ?

2.3. What energy process occurs when electron–hole pairs are created, and when they recombine?

2.4. What is the Fermi Level?

2.5. Electrical conduction in a metal is done entirely by electrons. The dominant form of conduction in a semiconductor can either be by holes or electrons. How does insertion of a Group III or a Group V element into a host Group IV Si element affect the choice of dominant hole or electron injection.

2.6. Two forces dominate the net motion of carriers in a semiconductor: drift and diffusion. Describe the conditions that promote these two mechanisms.

2.7. Describe how an electric field appears across a *pn* junction and the dynamic relation of this $\mathscr{E}$-field to charge movement across the *pn* junction.

2.8. The diode reverse bias saturation current I_S originates in the depletion region with its high electric field. The sources of this current are the thermal creation of electron–hole pairs that are then swept out of the junction by this high electric field. If the reverse-bias voltage is made larger, what is the impact on I_S? Will I_S increase, decrease, or stay the same? Explain.

2.9. The diode equation (2.10) predicts an exponential increase in diode current I_D as diode voltage V_D increases. Assume that you measured a diode in the forward-bias region and found an exponential relation at the lower V_D, but the curve tended toward a straight line at higher voltages. Explain.

2.10. Chapter 1 described capacitors made of two metal plates separated by a dielectric. Describe how a *pn* junction capacitance differs from the simple metal plate dielectric capacitor.

CHAPTER 3

MOSFET TRANSISTORS

MOSFET transistors are the basic element of today's integrated circuits (ICs). Our goal here is to impart the analytical ability and transistor insights that electrical engineers use in solving IC problems. An abundance of examples and self-exercises are provided to develop intuitive responses to digital transistor circuit operation. We begin with a simple picture of transistors as switches, and evolve to more developed analytical models. We want to smoothly lead the way through these topics, providing knowledge and insight about transistors in the long- and short-channel technologies. The information in this chapter is a foundation for subsequent chapters, and is a basis for understanding the electronic aberrations of defective circuits.

3.1 PRINCIPLES OF OPERATION: LONG-CHANNEL TRANSISTORS

Transistors are the basic blocks for building electronic circuits. A major difference between transistors and passive elements (resistors, capacitors, inductors, and diodes) is that transistor current and voltage characteristics vary with the voltage (or current) on a control terminal. There are two types of transistors with different physical principles: bipolar transistors and field effect transistors (FETs). There is only one type of bipolar transistor—the bipolar junction transistor (BJT)—and two types of FET devices—the junction field effect transistor (JFET) and the metal oxide semiconductor field effect transistor (MOSFET). Today's digital ICs mainly use MOSFETs, whereas bipolars are used in specific digital technologies and more generally in analog circuits. JFETs have specific applications and are not used in digital applications. We will focus on MOSFETs since they appear in more than 90% of today's digital applications.

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

MOSFETs have three signal terminals: gate (G), source (S) and drain (D), plus the bulk terminal (B), to which the gate, drain, and source voltages are referenced. Figure 3.1(a) shows a MOSFET with its four terminals and its thin insulator of SiO_2 (with thickness T_{OX}) between the gate and bulk. Figure 3.1(b) shows symbols commonly used for MOSFETs, where the bulk terminal is labeled (B) or implied (not drawn). There are two types of MOSFET transistors, the *n*MOS transistor and the *p*MOS transistor, depending on the polarity of the carriers responsible for conduction. A simple description of the device will introduce basic concepts and terms.

3.1.1 The MOSFET as a Digital Switch

The simplest view of MOSFET logic operation treats the transistor as a switch. The gate terminal is analogous to the light switch on the wall. When the gate has a high voltage, the transistor closes like a switch, and the drain and source terminals are electrically connected. Just as a light switch requires a certain force level to activate it, the transistor needs a certain voltage level to connect the drain and source terminals. This voltage is called the transistor threshold voltage V_t and is a fixed voltage different for *n*MOS (V_{tn}) and *p*MOS (V_{tp}) devices in a given fabrication process. The *n*MOS threshold voltage V_{tn} is always positive, whereas the *p*MOS threshold voltage V_{tp} is always negative.

Transistors act as switches with two conducting states, on and off, depending on the control (gate) terminal voltage. An ideal transistor has a zero ohm resistance between the drain and source when it is in the on-state, and infinite resistance between these terminals in the off-state. The ideal device should also switch between on- and off-states with a zero delay time as soon as the control variable changes state.

Unfortunately, transistors are not ideal switches. MOS transistors have a small, equivalent drain–source resistance in their conducting state and a high but not infinite resistance in the off-state. Additionally, the delay of a transistor to switch between on- and off-states is not zero. Several parameters (both geometric and technology related) determine the on and off equivalent resistance and capacitance that degrade the time needed to switch between both states.

Figure 3.2 shows switch models for the *n*MOS and *p*MOS transistors. In the *n*MOS, the source is the reference terminal, which is always at the lower voltage, so that $V_{DS} \geq 0$, and $V_{GS} \geq 0$. Since V_{tn} is always positive, the off-state occurs when $V_{GS} < V_{tn}$ so the de-

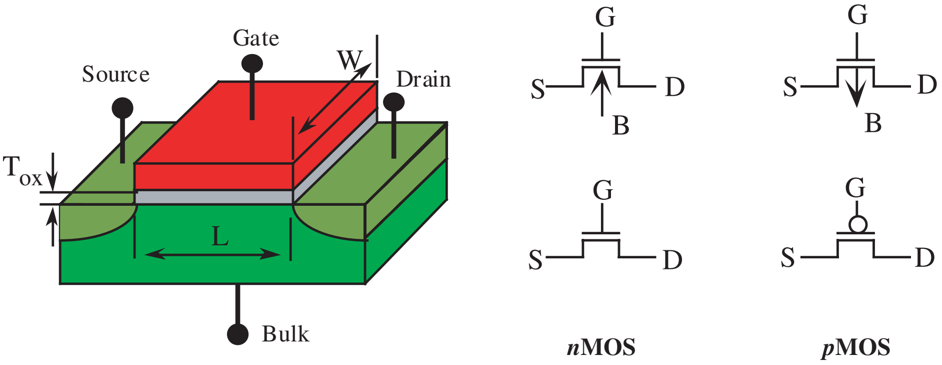
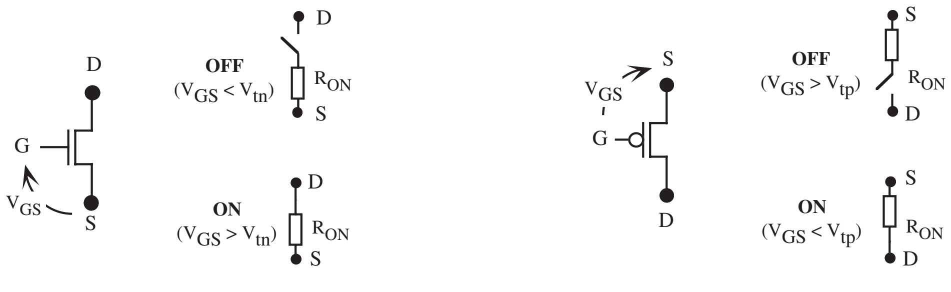


Figure 3.1. (a) MOS structure. (b) Symbols used at the circuit level.



*n*MOS ($V_{tn} > 0$, $V_{\mathrm{DS}} \geq 0$, $V_{\mathrm{GS}} \geq 0$) *p*MOS ($V_{tp} < 0$, $V_{\mathrm{DS}} \leq 0$, $V_{\mathrm{GS}} \leq 0$)

Figure 3.2. Transistor symbols and their equivalent ideal switches.

vice does not conduct, and it is modeled as an open switch. When $V_{\mathrm{GS}} > V_{tn}$, the device is in the on-state and modeled as a closed switch in series with a resistor R_{ON}. The model in Figure 3.2 represents this on resistance as constant, whereas in a real transistor the drain–source current (and therefore its equivalent resistance) depends on the operating state of the device, and must be determined from the V_{GS} and V_{DS} relations.

The *p*MOS transistor model is equivalent, but the signals have an opposite terminal polarity. The source is the reference terminal, which is always at the highest voltage, so that $V_{\mathrm{DS}} \leq 0$, and $V_{\mathrm{GS}} \leq 0$. V_{tp} has negative voltages in *p*MOS transistors. The off-state is defined when $V_{\mathrm{GS}} > V_{tp}$ and the on-state for $V_{\mathrm{GS}} < V_{tp}$. Since V_{tp} is always negative, the *p*MOS transistor turns on when $V_{\mathrm{GS}} < V_{tp}$, where both are negative numbers. This polarity confusion will become clear when we address *p*MOS transistor operation. For now, accept that the *p*MOS transistor has polarity control signals opposite to those of the *n*MOS transistor.

The ideal device characteristics for this simple model are

$$\begin{aligned} I_{\mathrm{DS}} &= 0, \text{ (off-state) when} \begin{cases} V_{\mathrm{GS}} < V_{tn} & n\mathrm{MOS} \\ V_{\mathrm{GS}} > V_{tp} & p\mathrm{MOS} \end{cases} \\ I_{\mathrm{DS}} &= \frac{V_{\mathrm{DS}}}{R_{\mathrm{ON}}}, \text{ (on-state) when} \begin{cases} V_{\mathrm{GS}} \geq V_{tn} & n\mathrm{MOS} \\ V_{\mathrm{GS}} \leq V_{tp} & p\mathrm{MOS} \end{cases} \end{aligned} \tag{3.1}$$

We will next consider the simple switch model and deepen our understanding of MOSFET structure, operating modes, and behavior models.

3.1.2 Physical Structure of MOSFETs

This section will describe transistors as they are today, and subsequent material will develop why they are so. MOSFET transistors are made from a crystalline semiconductor that forms the host structure called the substrate or bulk of the device. Substrates for *n*MOS are constructed from *p*-type silicon, whereas the *p*MOS substrates use *n*-type silicon. The thin oxide of the transistor electrically isolates the gate from the semiconductor crystalline structures underneath. The gate oxide is made of oxidized silicon, forming a noncrystalline, amorphous $\mathrm{SiO_2}$. The gate oxide thickness is typically from near 15 Å to

100 Å (1 Å = 1 angstrom = 10^{-10} m). SiO_2 molecules are about 3.5 Å in diameter, so this vital dimension is a few molecular layers thick. A thinner gate oxide provides more gate terminal control over the device state.

Drain and source regions are made from crystalline silicon by implanting a dopant with polarity opposite to that of the substrate. The region between the drain and source is called the channel. The distance from the drain to the source is a geometrical parameter called the *channel length* (L) of the device, as shown in Figure 3.1(a). Another geometrical parameter of the device is the transistor *channel width* (W) [Figure 3.1(a)]. Transistor length and width are geometrical parameters set by the circuit designer. Other parameters, such as the transistor oxide thickness, threshold voltage, and doping levels, depend on the fabrication process, and cannot be changed by design; they are technology parameters.

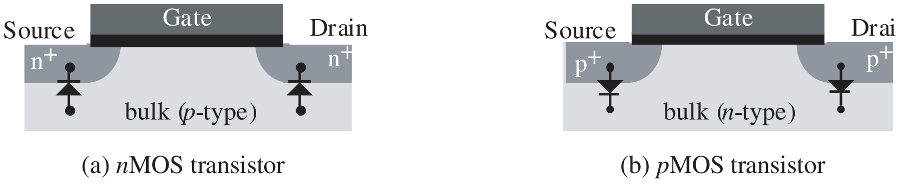
The gate is the control terminal, and the source provides electron or hole carriers that are collected by the drain. Often, the bulk terminals of all transistors are connected to the ground or power rail that is often the source and, therefore, not explicitly drawn in most schematics.

Figure 3.3 shows *n*MOS and *p*MOS transistor structures. The *n*MOS transistor has a *p*-type silicon substrate with opposite doping for the drain and source. *p*MOS transistors have a complementary structure with an *n*-type silicon bulk and *p*-type doped drain and source regions. The gate region in both transistors is constructed with polysilicon, and is isolated from the drain and source by the thin oxide. The region between the drain and source under the gate oxide is called the channel, and is where conduction takes place.

The gate is electrically isolated from the drain, source, and channel by the gate oxide insulator. Since drain and source dopants are opposite in polarity to the substrate (bulk), they form *pn* junction diodes (Figure 3.3) that in normal operation are reverse-biased. CMOS logic circuits typically match one *n*MOS transistor to one *p*MOS transistor.

3.1.3 Understanding MOS Transistor Operation: A Descriptive Approach

Transistor terminals must have proper voltage polarity to operate correctly (Figure 3.4). The bulk or substrate of *n*MOS (*p*MOS) transistors must always be connected to the lower (higher) voltage that is the reference terminal. We will assume that the bulk and source terminals are connected, to simplify the description. The positive convention current in an *n*MOS (*p*MOS) device is from the drain (source) to the source (drain), and is referred to as I_{DS} or just I_D, since drain and source current are equal. When a positive (negative) voltage is applied to the drain terminal, the drain current depends on the voltage applied to the gate control terminal. Note that for *p*MOS transistors, V_{GS}, V_{DS}, and I_{DS} are negative.



(a) *n*MOS transistor

(b) *p*MOS transistor

Figure 3.3. Relative doping and equivalent electrical connections between device terminals for (a) *n*MOS and (b) *p*MOS transistors.

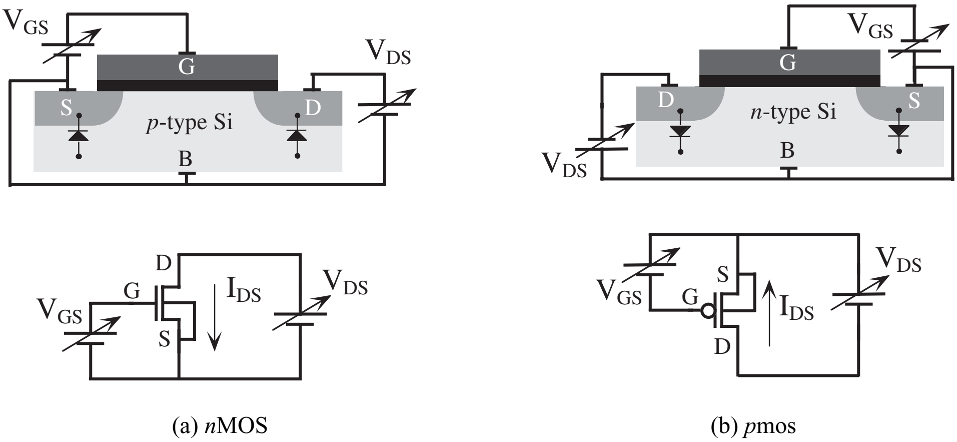


Figure 3.4. Normal transistor biasing (a) *n*MOS and (b) *p*MOS.

If V_{GS} is zero, then an applied drain voltage reverse-biases the drain–bulk diode (Figure 3.5), and there are no free charges between the drain and source. As a result, there is no current when $V_{GS} = 0$ for *n*MOS devices (the same hold for *p*MOS devices). This is the off, or nonconducting, state of the transistor.

We will first analyze transistor operation when the source and substrate are at the same voltage. When the gate terminal voltage of an *n*MOS (*p*MOS) transistor is slightly increased (decreased), a vertical electric field exists between the gate and the substrate across the oxide. In *n*MOS (*p*MOS) transistors, the holes (electrons) of the *p*-type (*n*-type) substrate close to the silicon–oxide interface initially "feel" this electrical field, and move away from the interface. As a result, a depletion region forms beneath the oxide interface for this small gate voltage (Figure 3.6). The depletion region contains no mobile carriers, so the application of a drain voltage provides no drain current, since free carriers still do not exist in the channel.

If the gate voltage of the *n*MOS (*p*MOS) device is further increased (decreased), then the vertical electric field is strong enough to attract minority carriers (electrons in the *n*MOS device and holes in the *p*MOS device) from the bulk toward the gate. These minority carriers are attracted to the gate, but the silicon dioxide insulator stops them, and the

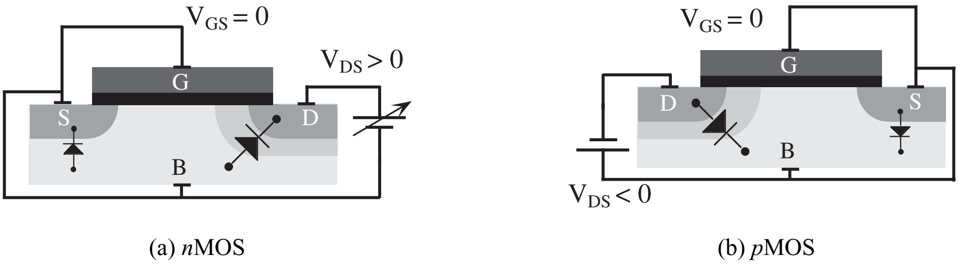
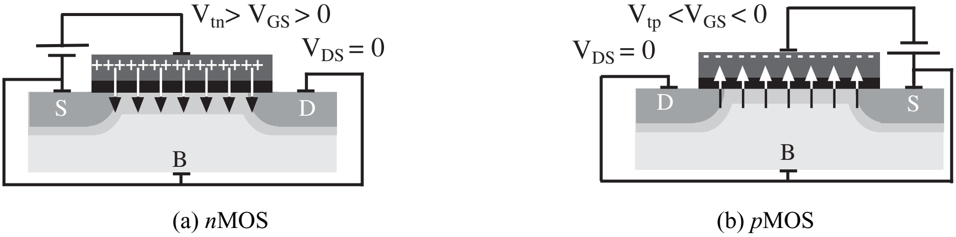


Figure 3.5. When the gate–source voltage is zero, the drain–source voltage reverses the built-in drain–bulk diode, preventing current from flowing from the drain to the source.



(a) *n*MOS (b) *p*MOS

Figure 3.6. (a) Depleting the *n*MOS channel of holes with small positive values of gate–source voltage, and (b) depleting the *p*MOS channel of electrons with small negative values of gate–source voltage.

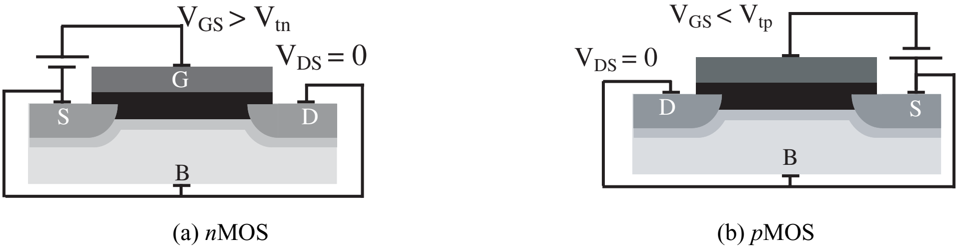
electrons (holes) accumulate at the silicon–oxide interface. They form a conducting plate of minority mobile carriers (electrons in the *p*-type bulk of the *n*MOS device, and holes in the *n*-type bulk of the *p*-MOS device). These carriers form the inversion region or conducting channel, which can be viewed as a "short circuit" to the drain/source-bulk diodes. This connection is shown in Figure 3.7.

Since the drain and source are at the same voltage, the channel carrier distribution is uniform along the device. The gate voltage for which the conducting channels respond is an intrinsic parameter of the transistor called the *threshold voltage,* referred to as V_t. As a first approximation, V_t can be considered constant for a given technology. The threshold voltage of a *n*MOS transistor is positive, while for a *p*MOS transistor it is negative. Since *n*MOS and *p*MOS transistors have a different threshold voltages, V_{tn} refers to the *n*MOS transistor threshold voltage, and V_{tp} to the *p*MOS transistor.

An *n*MOS (*p*MOS) transistor has a conducting channel when the gate–source voltage is greater than (less than) the threshold voltage, i.e., $V_{\mathrm{GS}} > V_{tn}$ ($V_{\mathrm{GS}} < V_{tp}$).

When the channel forms in the *n*MOS (*p*MOS) transistor, a positive (negative) drain voltage with respect to the source creates a horizontal electric field, moving the channel carriers toward the drain and forming a positive (negative) drain current. If the horizontal electric field is of the same order or smaller than the vertical one, the inversion channel remains almost uniform along the device length. This happens when

$$\begin{aligned} V_{\mathrm{DS}} &< (V_{\mathrm{GS}} - V_{tn}) \quad n\text{MOS transistor} \\ V_{\mathrm{DS}} &> (V_{\mathrm{GS}} - V_{tp}) \quad p\text{MOS transistor} \end{aligned} \tag{3.2}$$



(a) *n*MOS (b) *p*MOS

Figure 3.7. Creating the conducting channel for (a) *n*MOS and (b) *p*MOS transistors.

This condition will be explained below. It states that the vertical electric field dominates the horizontal one. The transistor is in its linear region, also called the ohmic or nonsaturated region.

If the drain voltage increases beyond the limit of Equation (3.2), the horizontal electric field becomes stronger than the vertical field at the drain end, creating an asymmetry of the channel carrier inversion distribution. The drain electric field is strong enough so that carrier inversion is not supported in this local drain region. The conducting channel retracts from the drain, and no longer "touches" this terminal. When this happens, the inversion channel is said to be "pinched off" and the device is in the *saturation region.* The pinch-off point is the location that separates the channel inversion region from the drain depletion region. It varies with changes in bias voltages. The channel distribution in this bias is shown in Figure 3.8.

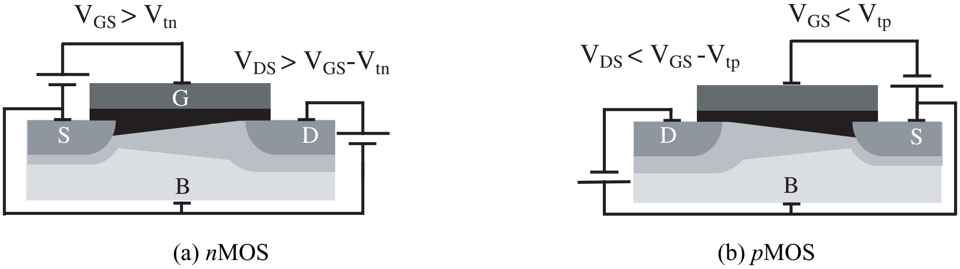
Although there are no inversion charges at the drain end of the channel, the drain region is still electrically active. Carriers depart from the source and move under the effect of the horizontal field. Once they arrive at the pinch-off point of the channel, they travel from that point to the drain, driven by the high electric field of the depletion region.

CMOS ICs use all three states described here: off-state, saturated state, and the linear state. We will next look at real curves of MOS parameters, and learn how to use the analytical equations that predict and analyze transistor behavior in normal and defective environments. It is important to work through all examples and exercises. The examples will analyze MOS long- and short-channel transistor circuits.

3.1.4 MOSFET Input Characteristics

MOS transistors cannot be described with a single current–voltage curve, as can diodes and resistors, since they have four terminals. Transistors require that two sets of current–voltage curves be characterized: the input characteristic and the output characteristic. The input characteristic is a single curve that relates drain current to the input gate–source driving voltage. Since the gate terminal is electrically isolated from the remaining terminals (drain, source, and bulk), the current through the gate is essentially zero, so that gate current is not part of the device characteristics. The input characteristic curve can locate the voltage on the control terminal (gate) at which the transistor leaves the off-state.

Figure 3.9 shows measured input characteristics for an *n*MOS and *p*MOS transistor with a small, 0.1 V potential across their drain-to-source terminals. As V_{GS} increases for



(a) *n*MOS (b) *p*MOS

Figure 3.8. Channel pinch-off for (a) *n*MOS and (b) *p*MOS transistor devices.

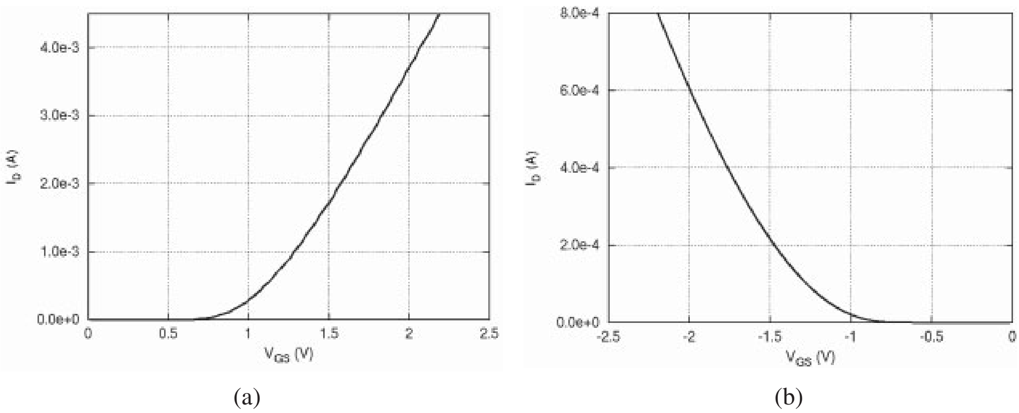


Figure 3.9. Measured input characteristics (I_D vs. V_{GS}) for (a) an *n*MOS, and (b) a *p*MOS transistor.

the *n*MOS transistor in Figure 3.9(a), a voltage is reached at which drain current begins. When the *n*MOS device conducts, the drain current is positive, since the current enters the drain. For V_{GS} between 0 V and 0.7 V, the drain current is nearly zero, indicating that the equivalent resistance between the drain and source terminals is extremely high. Once the gate–source voltage reaches 0.7 V, the current increases rapidly with V_{GS}, indicating that the equivalent resistance at the drain decreases with increasing gate–source voltage. Therefore, the threshold voltage of this device is about $V_{tn} \approx 0.7$ V. When a transistor turns on and current moves through a load, then voltage changes occur that translate into logic levels.

The *p*MOS transistor input characteristic in Figure 3.9(b) is analogous to the *n*MOS transistor except that the I_{DS} and V_{GS} polarities are reversed. V_{DS} is negative ($V_{DS} \approx -0.1$ V) and the drain current in a *p*MOS transistor is negative, indicating that it exits the drain terminal. Additionally, the gate is at a voltage lower than the source terminal voltage to attract holes to the channel surface. The threshold voltage of the *p*MOS device in Figure 3.9(b) can be seen as approximately $V_{tp} \approx -0.8$ V.

3.1.5 *n*MOS Transistor Output Characteristics

MOS transistor output characteristics plot I_D versus V_{DS} for several values of V_{GS}. Figure 3.10 shows such a measurement of an *n*MOS transistor. When the transistor is in the off-state ($V_{GS} < V_{tn}$), then I_D is near zero for any V_{DS} value. When the device is in the on-state, transistor conduction properties vary with the bias state of the device. Two states, described as the ohmic (or nonsaturated) state and the saturated state, are distinguished when the device is in the on-state. The boundary of the two bias states is seen for each curve in Figure 3.10 as the intersection of the straight line of the saturated region with the curving line of the ohmic region. This intersection point occurs at V_{Dsat}. In the ohmic state, the drain current initially increases almost linearly with the drain voltage before bending and flattening out. The drain current in saturation is virtually independent of V_{DS} and the transistor acts as a current source. A near-constant current is driven from the transistor no matter what the drain-to-source voltage is.

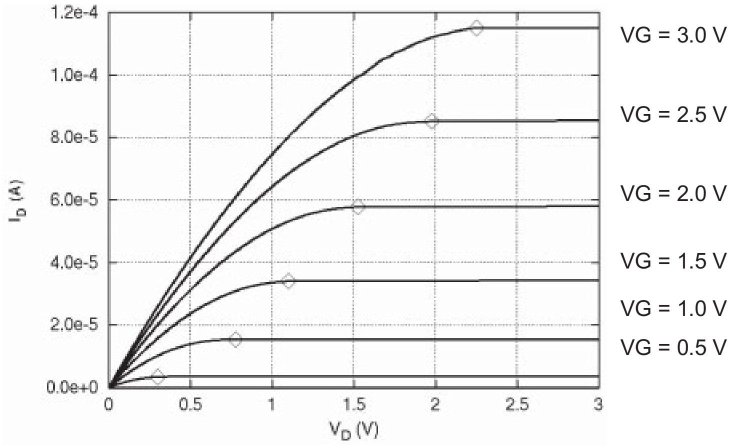


Figure 3.10. *n*MOS transistor output characteristics as a family of curves. The diamond symbol marks the pinch-off voltage, V_{DSAT}.

This family of curves is rapidly measured with a digital curve tracer called a parameter analyzer. Many useful parameters can be measured from data in the family of curves, and deviations in the curves are also a good indicator of a damaged transistor. Later, we will deepen our understanding of transistor operation expressed by Figure 3.10. *p*MOS transistor I_{D} versus V_{DS} curves have shapes similar to those in Figure 3.10, but the voltage and current polarities are negative to account for hole inversion and drain current that enters the transistor (*p*MOS device curves are shown later).

We next develop skills with the equations that predict voltages and currents in a transistor for any point in the family of curves in Figure 3.10. This capability is needed to analyze electronic behavior of CMOS circuits with bridge, open circuit, or parametric defects.

MOS equations can be derived by calculating the amount of charge in the channel at each point, and integrating such an expression from the drain to the source. This procedure is found in several books [3, 6, 7], and leads to expressions for the drain current in the linear and saturated states. Equations (3.3) and (3.4) are these equations for the *n*MOS transistor in the saturated and ohmic states.

$$I_{\mathrm{D}} = \frac{\mu \varepsilon_{\mathrm{ox}}}{2T_{\mathrm{ox}}} \frac{W}{L} (V_{\mathrm{GS}} - V_{tn})^2 \qquad \text{(saturated state)} \tag{3.3}$$

$$I_{\mathrm{D}} = \frac{\mu \varepsilon_{\mathrm{ox}}}{2T_{\mathrm{ox}}} \frac{W}{L} [2(V_{\mathrm{GS}} - V_{tn})V_{\mathrm{DS}} - V_{\mathrm{DS}}^2] \qquad \text{(ohmic state)} \tag{3.4}$$

where μ is the electron mobility, $\varepsilon_{\mathrm{ox}}$ is thin oxide (SiO$_2$) dielectric constant, T_{ox} is the transistor oxide thickness, and W and L are transistor effective gate width and length. A constant, K, is introduced to indicate the drive strength of the transistor as

$$K = \frac{\mu \varepsilon_{\mathrm{ox}}}{2T_{\mathrm{ox}}} \tag{3.5}$$

If these constants are known, then Equation (3.3) can predict I_D for any value of V_{GS} in the saturated region, and Equation (3.4) can predict any I_D in the ohmic region if V_{GS} and V_{DS} are specified. Equation (3.3) is a square-law relation between I_D and V_{GS} that is independent of V_{DS}. Equation (3.3) is a flat line for a given V_{GS}, whereas Equation (3.4) is a parabola. For any V_{GS}, the two equations have an intersect point that is seen in Figure 3.10. The intersection point occurs at a parameter called V_{Dsat}, for which either equation describes the current and voltage relations.

We can solve for this important bias condition at which the saturated and ohmic states intersect (V_{Dsat}), and this knowledge is essential for solving problems that follow. Figure 3.11 plots three parabolas of Equation (3.4) at V_{GS} = 2.0 V, 1.6 V, and 1.2 V. Only the left-hand sides of the parabolas are used to predict the curves in Figure 3.10, but the parabolas also have a right-hand side. The dotted lines on the right-hand side of the curves are part of the continuous solution to the parabolas, but are electronically invalid, as examples will show.

The midpoint at zero slope defines the useful upper region of Equation (3.4), and also defines the boundary between the saturated and ohmic bias states. We can define the boundary bias condition by differentiating Equation (3.4) with respect to V_{DS}, setting the expression to zero, and then solving for the conditions. Equation (3.6) shows the derivative of Equation (3.4) set to zero:

$$\frac{dI_D}{dV_{DS}} = \frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}[2(V_{GS} - V_{tn}) - 2V_{DS}] = 0 \tag{3.6}$$

Terms cancel, giving the bias condition at the transition between saturation and nonsaturation states as

$$V_{GS} = V_{DS} + V_{tn} \tag{3.7}$$

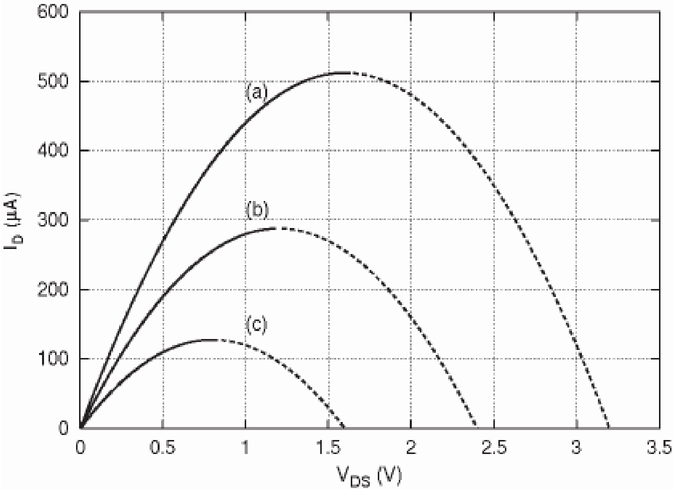


Figure 3.11. Plot of parabola of the nonsaturation state equation [Equation (3.4)]. K_n = 100 μA/V^2, V_{tn} = 0.4 V, and W/L = 2. Solid lines indicate valid regions, but dotted lines do not. (a) V_{GS} = 2.0 V, (b) V_{GS} = 1.6 V, (c) V_{GS} = 1.2 V.

This equation holds for each of the intersection points in Figure 3.11 denoted at the peak of each curve. Equation (3.7) can be extended to define the *n*MOS saturated bias condition:

$$V_{\mathrm{DS}} > V_{\mathrm{GS}} - V_t \qquad \text{or} \qquad V_{\mathrm{GS}} < V_{\mathrm{DS}} + V_t \tag{3.8}$$

and the *n*MOS ohmic condition:

$$V_{\mathrm{DS}} < V_{\mathrm{GS}} - V_t \qquad \text{or} \qquad V_{\mathrm{GS}} > V_{\mathrm{DS}} + V_t \tag{3.9}$$

We use these relations to analyze the effect of defects on CMOS circuits. A series of examples and exercises will illustrate their use. We emphasize that drill imparts the intuition that experienced failure analysts and test, reliability, and product engineers use in CMOS IC manufacturing.

■ EXAMPLE 3.1

Determine the bias state for the three conditions in Figure 3.12 if $V_{tn} = 0.4$ V.

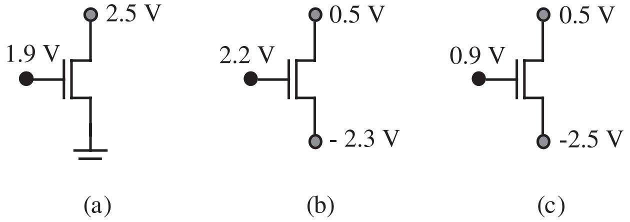


Figure 3.12. Transistor bias-state examples.

(a) $V_{\mathrm{GS}} = 1.9$ V, $V_{\mathrm{DS}} = 2.5$ V, and $V_{tn} = 0.4$ V, therefore $V_{\mathrm{GS}} = 1.9\text{ V} < 2.5\text{ V} + 0.4\text{ V} = 2.9$ V. Equation (3.8) is satisfied, and the transistor is in the saturated state described by Equation (3.3).

(b) $V_{\mathrm{GS}} = V_{\mathrm{G}} - V_{\mathrm{S}} = 2.2\text{ V} - (-2.3\text{ V}) = 4.6$ V. $V_{\mathrm{DS}} = V_{\mathrm{D}} - V_{\mathrm{S}} = 0.5\text{ V} - (-2.3) = 2.8$ V. Therefore, $V_{\mathrm{GS}} = 4.6\text{ V} > 2.2\text{ V} + 0.4\text{ V} = 2.6$ V. Equation (3.9) is satisfied, and the transistor is in the nonsaturated state.

(c) $V_{\mathrm{GS}} = V_{\mathrm{G}} - V_{\mathrm{S}} = 0.9\text{ V} - (-2.5\text{ V}) = 3.4$ V. $V_{\mathrm{DS}} = V_{\mathrm{D}} - V_{\mathrm{S}} = 0.5\text{ V} - (-2.5\text{ V}) = 3$ V. Therefore, $V_{\mathrm{GS}} = 3.4\text{ V} = V_{\mathrm{DS}} + V_t = 3\text{ V} + 0.4\text{ V} = 3.4$ V, and the transistor is at the boundary of the saturated and ohmic regions. Either Equation (3.3) or (3.4) can be used to calculate I_{D}. ■

Self-Exercise 3.1

Determine the bias state for the three conditions in Figure 3.13 if $V_{tn} = 0.4$ V.

After solving bias Example 3.1 and Self-Exercise 3.1 with the proper bias-state equations, you may check your work by referring to the *n*MOS transistor family of curves in Figure

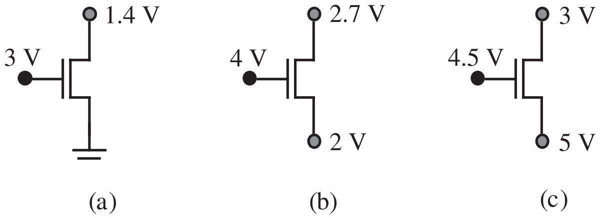


Figure 3.13. Bias-state exercises.

3.10. Find the coordinates in the example and exercise, and verify that the bias state is correct. A series of examples and exercises with the *n*MOS transistor will reinforce these important relations.

■ EXAMPLE 3.2

Calculate I_D and V_{DS} if $K_n = 100\ \mu A/V^2$, $V_{tn} = 0.6$ V, and $W/L = 3$ for transistor M1 in the circuit in Figure 3.14.

The bias state of M1 is not known, so we must initially assume one of the two states, solve for bias voltages, and then check for consistency against that transistor's bias condition. Initially, assume that the transistor is in the saturated state so that

$$I_D = \frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}(V_{GS} - V_{tn})^2 = K_n\frac{W}{L}(V_{GS} - V_{tn})^2$$
$$= (100\ \mu A)\,(3)\,(1.5 - 0.6)^2$$
$$= 243\ \mu A$$

Using Kirchhoff's voltage law (KVL):

$$V_{DS} = V_{DD} - I_D R$$
$$= 5 - (243\ \mu A)(15\ k\Omega)$$
$$= 1.355\ V$$

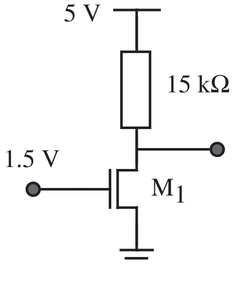


Figure 3.14.

We assumed that the transistor was in saturation, so we must check the result to see if that is true. For saturation,

$$V_{GS} < V_{DS} + V_{tn}$$
$$1.5\ V < 1.355\ V + 0.6\ V$$

so the transistor is in saturation, and our assumption and answers are correct. ■

■ EXAMPLE 3.3

Repeat Example 3.2, finding I_D and V_{DS} if $V_G = 1.8$ V.

Assume a transistor saturated state and

$$\begin{aligned} I_D &= (100\ \mu\text{A})(3)(1.8 - 0.6)^2 \\ &= 432\ \mu\text{A} \end{aligned}$$

$$\begin{aligned} V_{DS} &= 5 - (432\ \mu\text{A})(15\ \text{k}\Omega) \\ &= -1.48\ \text{V} \end{aligned}$$

This value for V_{DS} is clearly not reasonable since there are no negative potentials in the circuit. Also, the bias check gives

$$V_{GS} > V_{DS} + V_{tn}$$
$$1.8V > -1.48\ \text{V} + 0.6\ \text{V}$$

The initial saturated state assumption was wrong, so we repeat the analysis using the ohmic state assumption:

$$I_D = \frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}[2(V_{GS} - V_{tn})V_{DS} - V_{DS}^2]$$

$$\begin{aligned} I_D &= K_n\frac{W}{L}[2(V_{GS} - V_{tn})V_{DS} - V_{DS}^2] \\ &= (100\ \mu\text{A})\ (3)\ [2(1.8 - 0.6)\ V_{DS} - V_{DS}^2] \\ &= 300\ \mu\text{A}\ [2.4\ V_{DS} - V_{DS}^2] \end{aligned}$$

This equation has two unknowns, so another equation must be found. We will use the KVL statement,

$$V_{DD} = I_D R + V_{DS}$$

$$\begin{aligned} I_D &= (V_{DD} - V_{DS})/R \\ &= (5 - V_{DS})/15\ \text{k}\Omega \end{aligned}$$

The two equations can be equated to their I_D solution, giving

$$\frac{(5 - V_{DS})}{15\ \text{k}\Omega} = 300\ \mu\text{A}(2.4\ V_{DS} - V_{DS}^2)$$

After some algebra, this reduces to

$$V_{\mathrm{DS}}^2 - 2.622V_{\mathrm{DS}} + 1.11 = 0$$

The two quadratic solutions are

$$V_{\mathrm{DS}} = 0.531\ \mathrm{V},\ 2.09\ \mathrm{V}$$

The valid solution is $V_{\mathrm{DS}} = 0.531$ V, since this satisfies the nonsaturation condition that was used in its solution:

$$V_{\mathrm{GS}} > V_{\mathrm{DS}} + V_{tn}$$
$$1.8\ \mathrm{V} > 0.531\ \mathrm{V} + 0.6\ \mathrm{V}$$

and

$$\begin{aligned} I_{\mathrm{D}} &= (V_{\mathrm{DD}} - V_{\mathrm{DS}})/15\ \mathrm{k\Omega} \\ &= (5\ \mathrm{V} - 0.531\ \mathrm{V})/15\ \mathrm{k\Omega} \\ &= 298\ \mu\mathrm{A} \end{aligned}$$

■

■ EXAMPLE 3.4

What value of R_{d} will drive transistor M1 in Figure 3.15 just into nonsaturation if $K_n = 50\ \mu\mathrm{A/V}^2$, $V_{tn} = 0.4$ V, and $W/L = 10$?

Since the bias state is at the boundary, either Equation (3.3) or (3.4) can be used. Equation (3.3) is simpler so

$$\begin{aligned} I_{\mathrm{D}} &= \frac{\mu\varepsilon_{\mathrm{ox}}}{2T_{\mathrm{ox}}}\frac{W}{L}(V_{\mathrm{GS}} - V_{tn})^2 = K_{\mathrm{n}}\frac{W}{L}(V_{\mathrm{GS}} - V_{tn})^2 \\ &= (50\ \mu\mathrm{A})(10)(1.0 - 0.4)^2 \\ &= 180\ \mu\mathrm{A} \end{aligned}$$

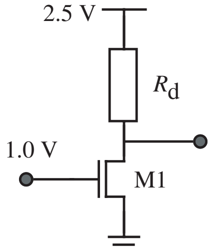


Figure 3.15.

The bias boundary condition

$$V_{\mathrm{GS}} = V_{\mathrm{DS}} + V_{tn}$$

becomes

$$\begin{aligned} V_{\mathrm{G}} - V_{\mathrm{S}} &= V_{\mathrm{D}} - V_{\mathrm{S}} + V_{tn} \\ V_{\mathrm{D}} &= V_{\mathrm{G}} - V_{tn} \\ V_{\mathrm{D}} &= 1.0\ \mathrm{V} - 0.4\ \mathrm{V} = 0.6\ \mathrm{V} \end{aligned}$$

Then

$$\begin{aligned} R_{\mathrm{d}} &= \frac{V_{\mathrm{DD}} - V_{\mathrm{D}}}{I_{\mathrm{D}}} \\ &= \frac{2.5 - 0.6}{180\ \mu\mathrm{A}} \\ &= 10.56\ \mathrm{k}\Omega \end{aligned}$$

■

■ EXAMPLE 3.5

Transistors emit light from the drain depletion region when they are in the saturated bias state.

(a) Show whether this useful failure analysis technique will work for the circuit in Figure 3.16. $V_{tn} = 0.6$ V, $K_n = 75\ \mu\mathrm{A/V}^2$, and $W/L = 2$.

(b) Find I_{D}, V_{GS}, and V_{DS}.

The saturated bias condition is

$$V_{\mathrm{GS}} < V_{\mathrm{DS}} + V_{tn}$$

or

$$\begin{aligned} V_{\mathrm{G}} &< V_{\mathrm{D}} + V_{tn} \\ 1.2\ \mathrm{V} &< 3.3\ \mathrm{V} + 0.6\ \mathrm{V} \end{aligned}$$

so, transistor M1 is saturated and emitting visible light from its drain-channel region.

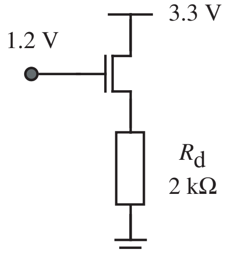


Figure 3.16.

Since M1 is in saturation,

$$
\begin{aligned}
I_D &= \frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}(V_{GS} - V_{tn})^2 = K_n\frac{W}{L}(V_{GS} - V_{tn})^2 \\
&= (75\ \mu\text{A})(2)[(V_G - V_S) - 0.6]^2 \\
&= (150\ \mu\text{A})[(1.2 - V_S) - 0.6]^2 \\
&= (150\ \mu\text{A})(0.6 - V_S)^2
\end{aligned}
$$

Also,

$$I_D = \frac{V_S}{R_d} = \frac{V_S}{2\ \text{k}\Omega}$$

Then,

$$\frac{V_S}{2\ \text{k}\Omega} = (150\ \mu\text{A})(0.6\ \text{V} - V_S)^2$$

This reduces to

$$V_S^2 - 4.533\ V_S + 0.36 = 0$$

whose two solutions are $V_S = 80.85$ mV and 4.452 V. The valid solution is

$$V_S = 80.85\ \text{mV}$$

and

$$
\begin{aligned}
I_D = I_S = V_S/R_d &= 80.85\ \text{mV}/2\ \text{k}\Omega = 40.43\ \mu\text{A} \\
V_{GS} &= 1.2 - 0.08085 = 1.119\ \text{V} \\
V_{DS} &= 3.2 - 0.08085 = 3.119\ \text{V}
\end{aligned}
$$

■

Self-Exercise 3.2

Find I_D and V_D in Figure 3.17. Verify the bias state consistency of your choice of MOS drain-current model for $V_{tn} = 1.0$ V, $K_n = 25\ \mu\text{A/V}^2$, and $W/L = 2$.

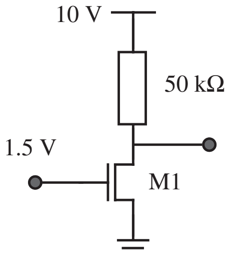


Figure 3.17.

Self-Exercise 3.3

Repeat Self-Exercise 3.2 (Figure 3.17) if $V_G = 3.0$ V and $W/L = 3$.

Self-Exercise 3.4

Calculate V_{GS} and give the correct bias state for transistor M1 in Figure 3.18. $V_{tn} = 0.5$ V.

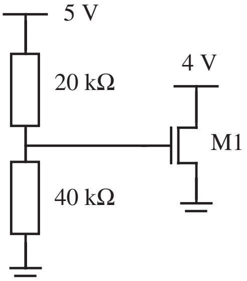


Figure 3.18.

Self-Exercise 3.5

Adjust R_1 in Figure 3.19 so that M1 is on the saturated/ohmic border where $V_{tn} = 0.5$ V.

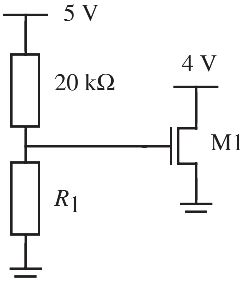


Figure 3.19.

Self-Exercise 3.6

Calculate R_0 in Figure 3.20 so that $V_0 = 2.5$ V. Given: $K_n = 300\ \mu A/V^2$, $V_{tn} = 0.7$ V, and $W/L = 2$.

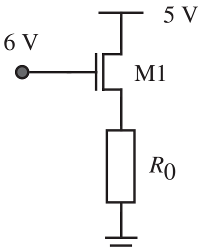


Figure 3.20.

3.1.6 *p*MOS Transistor Output Characteristics

*p*MOS transistor analysis is similar to that for the *n*MOS transistor with a major exception: care must be taken with the polarities of the current and voltages. The *p*MOS transistor major carrier is the hole that emanates from the source, enters the channel, and exits the drain terminal as a negative current. The gate-to-source threshold voltage V_{tp} needed to invert an *n*-substrate is negative to attract holes to the channel surface. The equations to model the *p*MOS transistor in saturation and nonsaturation conditions have a form similar to those for the *n*MOS device, but modified for these polarity considerations. We will choose a *p*MOS transistor equation form that is close to the *n*MOS transistor equations. Equations (3.10) and (3.11) describe the terminal behavior of a *p*MOS device. Remember that V_{GS}, V_{DS}, V_{tp}, and I_{D} are negative.

$$I_{\mathrm{D}} = -\frac{\mu\varepsilon_{\mathrm{ox}}}{2T_{\mathrm{ox}}}\frac{W}{L}(V_{\mathrm{GS}} - V_{tp})^2 \qquad \text{(saturated state)} \tag{3.10}$$

$$I_{\mathrm{D}} = -\frac{\mu\varepsilon_{\mathrm{ox}}}{2T_{\mathrm{ox}}}\frac{W}{L}[2(V_{\mathrm{GS}} - V_{tp})V_{\mathrm{DS}} - V_{\mathrm{DS}}^2] \qquad \text{(ohmic state)} \tag{3.11}$$

Figure 3.21 shows a measured *p*MOS transistor family of curves with all voltages given with respect to the source. The plot is shown in quadrant I, even though the drain current and voltage are negative. This is an author's choice, made to to retain similarity to the *n*MOS transistor family of curves.

The boundary of the bias states can again be found by differentiating Equation (3.11), setting the result to zero, and solving for the conditions to get

$$V_{\mathrm{DS}} = V_{\mathrm{GS}} - V_{tp} \tag{3.12}$$

The condition for transistor saturation is

$$V_{\mathrm{DS}} < V_{\mathrm{GS}} - V_{tp} \qquad \text{or} \qquad V_{\mathrm{GS}} > V_{\mathrm{DS}} + V_{tp} \tag{3.13}$$

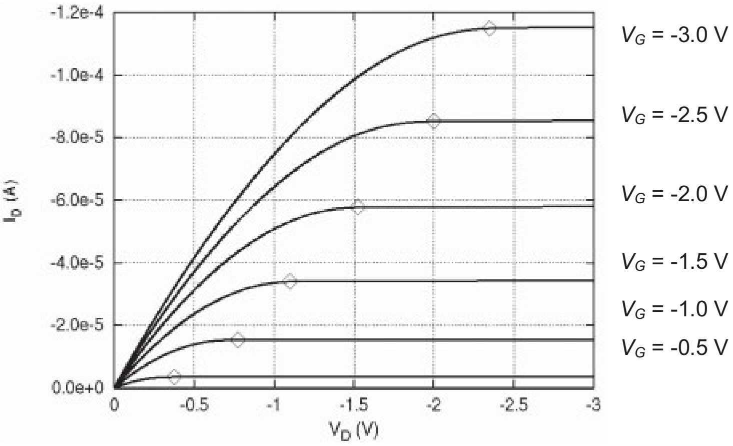


Figure 3.21. *p*MOS transistor family of curves.

and the condition for transistor nonsaturation is

$$V_{DS} > V_{GS} - V_{tp} \qquad \text{or} \qquad V_{GS} < V_{DS} + V_{tp} \tag{3.14}$$

An example is given using these equations.

■ EXAMPLE 3.6

Determine the bias state for the *p*MOS transistors in Figure 3.22, where V_{tp} = –0.4 V. The gate terminal has its most negative voltage with respect to the source terminal.

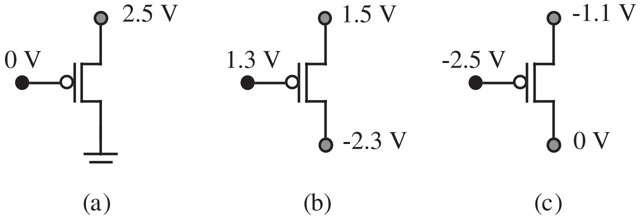


Figure 3.22. Transistors for Example 3.6, with $V_{tp} = -1.2$ V.

(a) $V_{GS} = -2.5$ V and $V_{DS} = -2.5$ V, therefore $V_{GS} > V_{DS} + V_{tp}$, or – 2.5 V > – 2.5 + (–0.4) V, so the transistor is in saturation.

(b) The gate voltage is not sufficiently more negative than either the drain or source terminal so that the transistor is in the off-state.

(c) $V_{GS} = -2.5 - (-1.1) = -1.4$ V and $V_{DS} = 0 - (-1.1) = 1.1$ V. What is wrong? The gate voltage is sufficiently negative to turn on the transistor, but the source-to-drain voltage is negative. Holes must leave the source and flow to the drain, but they can't under this condition. The answer is that the drain terminal is on the top and the source on the bottom so that $V_{GS} = -2.5 - 0 = -2.5$ V and $V_{DS} = -1.1 - 0 = -1.1$ V. Therefore $V_{GS} < V_{DS} + V_{tp}$, or – 2.5 V < –1.1 + (–1.2) V, so the transistor is in nonsaturation. The source terminal always has a higher or equal voltage than the drain terminal in a *p*MOS transistor. ■

Self-Exercise 3.7

Give the correct bias state for the three *p*MOS's shown in Figure 3.23, where V_{tp} = –0.4 V.

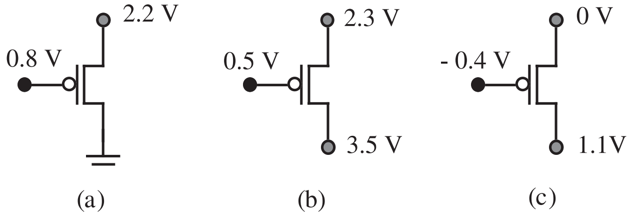


Figure 3.23.

After solving Example 3.6 and Self-Exercise 3.7 with the proper bias state equations, you may check your work by referring to the *p*MOS transistor family of curves in Figure 3.21. Find the coordinates in the example and exercise, and verify that the bias state is correct. A series of examples and exercises with the *p*MOS transistor will reinforce these important relations.

■ EXAMPLE 3.7

Calculate I_D and V_{DS} for circuit in Figure 3.24. $V_{tp} = -1.0$ V, $K_p = 100\ \mu A/V^2$, and $W/L = 4$.

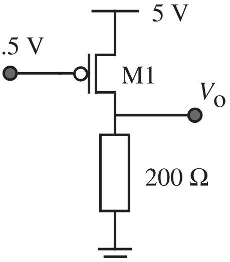


Figure 3.24.

Assume a saturated bias state:

$$I_D = -\frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}(V_{GS} - V_{tp})^2 = -100\ \mu A(4)[-3.5 - (-1)]^2$$
$$= -2.5\ \text{mA}$$
$$V_0 = -I_D(200\ \Omega) = (2.5\ \text{mA})(200\ \Omega) = 0.5\ \text{V}$$

then

$$V_{DS} = 0.5\ \text{V} - 5\ \text{V} = -4.5\ \text{V}$$

The bias state consistency is

$$V_{GS} > V_{DS} + V_{tp}$$
$$-3.5 > -4.5 + (-1.0)$$

so the transistor is in the saturated bias state and the solutions are correct. ■

■ EXAMPLE 3.8

Calculate I_D and V_{DS} in Figure 3.25. $V_{tp} = -0.6$ V, $K_p = 80\ \mu A/V^2$, and $W/L = 10$.

Assume a saturated bias state:

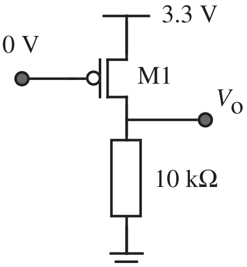


Figure 3.25.

$$I_D = -\frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}(V_{GS} - V_{tp})^2 = -80\ \mu A(10)[-3.3 - (-0.6)]^2$$
$$= -5.832\ mA$$

Then

$$V_0 = -I_D(10\ k\Omega) = -(-5.832\ mA)(10\ k\Omega)$$
$$= 58.32\ V$$

This voltage is beyond the power supply value, and is not possible. The saturated state assumption was wrong, so we must start again using the ohmic state equation:

$$I_D = -\frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}[2(V_{GS} - V_{tp})V_{DS} - V_{DS}^2] = -80\ \mu A(10)[2(-3.3 + 0.6)V_{DS} - V_{DS}^2]$$

Another equation is required to solve the problem, so using the KVL (Ohm's law, here)

$$-I_D = \frac{V_D}{R_d} = \frac{V_{DD} + V_{DS}}{10\ k\Omega}$$
$$= \frac{3.3 + V_{DS}}{10\ k\Omega} = 80\ \mu A(10)[2(-3.3 + 0.6)V_{DS} - V_{DS}^2]$$

The two quadratic solutions are: $V_{DS} = -75.70$ mV and –5.450 V. The correct solution is $V_{DS} = -75.70$ mV. Therefore

$$V_0 = 3.3\ V + (-75.70\ mV) = 3.224\ V$$

■ EXAMPLE 3.9

Calculate I_D and V_{SD}, and verify the assumed bias state of transistor M1 for the circuit in Figure 3.26. $V_{tp} = -0.4$ V, $K_p = 60\ \mu A/V^2$, and $W/L = 2$.

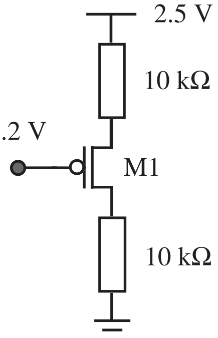


Figure 3.26.

Assume a saturated bias state and

$$I_D = -\frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}(V_{GS} - V_{tp})^2$$

Since V_{GS} is not known, we must search for another expression to supplement this equation. We can use the KVL statement:

$$\begin{aligned} V_{GS} &= 1.2 - [V_{DD} - (-I_D R_S)] \\ &= 1.2 - 2.5 + I_D R_S \\ &= -1.3 + (10\ \text{k}\Omega)I_D \end{aligned}$$

We equate this expression to the saturated current expression to get

$$\begin{aligned} I_D &= -60\ \mu\text{A}(2)[-1.3 + (10\ \text{k}\Omega)I_D + 0.4]^2 \\ &= -120\ \mu\text{A}[-0.9 + (10\ \text{k}\Omega)I_D]^2 \end{aligned}$$

This quadratic equation in I_D gives solutions

$$I_D = -35.56\ \mu\text{A} \qquad \text{and} \qquad -227.8\ \mu\text{A}$$

The valid solution is $I_D = -35.56\ \mu\text{A}$, since the other solution for I_D, when multiplied by the sum of the two resistors, gives a voltage greater than the power supply. V_{SD} is then

$$\begin{aligned} V_{SD} &= V_{DD} - I_D(20\ \text{k}\Omega) \\ V_{SD} &= I_D(20\ \text{k}\Omega) - V_{DD} \\ &= 2.5\ \text{V} - (35.56\ \mu\text{A})(20\ \text{k}\Omega) \\ &= 1.789\ \text{V} \end{aligned}$$

and

$$V_S = V_{DD} - I_D R_S = 2.5 + (-35.56\ \mu A)(10\ k\Omega)$$
$$= 2.144\ V$$

so that

$$V_{GS} = V_G - V_S = 1.2 - 2.144 = -0.944\ V$$

Transistor M1 is in saturation since

$$V_{GS} > V_{DS} + V_{tp}$$
$$-0.944\ V > -1.789\ V - 0.4\ V$$

■

■ EXAMPLE 3.10

What value of R_d in Figure 3.27 will raise V_0 to half of the power supply voltage (i.e., $V_0 = 0.5\ V_{DD}$). $V_{tp} = -0.7$ V, $K_p = 80\ \mu A/V^2$, and $W/L = 5$.

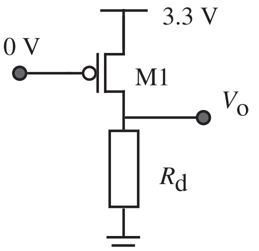


Figure 3.27.

Check for bias state consistency:

$$V_{GS} < V_{DS} + V_{tp}$$
$$-3.3\ V < -1.65\ V - 0.7\ V$$

So M1 is in ohmic bias state. Therefore, we use the ohmic state equation, where $V_{DS} = -1.65$ V:

$$I_D = -\frac{\mu\varepsilon_{ox}}{2T_{ox}}\frac{W}{L}[2(V_{GS} - V_{tp})V_{DS} - V_{DS}^2]$$
$$= -80\ \mu A(5)\{2[-3.3 - (-0.7)](-1.65) - 1.65^2\}$$
$$= -2.343\ mA$$

Then

$$R_d = \frac{V_0}{-I_D} = \frac{1.65}{2.343\text{ mA}} = 704.2\ \Omega$$

■

Self-Exercise 3.8

Calculate I_D and V_0 for circuit in Figure 3.28. $V_{tp} = -0.8$ V, $K_p = 30\ \mu\text{A/V}^2$, and $W/L = 2$.

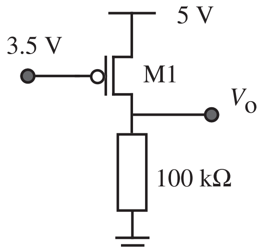


Figure 3.28.

Self-Exercise 3.9

Repeat Self-Exercise 3.8, but let $V_G = 1.5$ V.

Self-Exercise 3.10

Find I_D and V_0 for the circuit in Figure 3.29. $V_{tp} = -0.6$ V, $K_p = 20\ \mu\text{A/V}^2$, and $W/L = 3$.

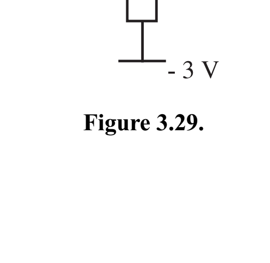


Figure 3.29.

Self-Exercise 3.11

The voltage drop across each of the two identical resistors and V_{DS} are equal for the circuit in Figure 3.30. $V_{tp} = -0.5$ V, $K_p = 100\ \mu A/V^2$, and $W/L = 2$. Find the value of the resistors.

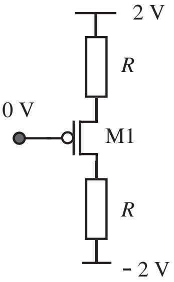


Figure 3.30.

Self-Exercise 3.12

For the circuit in Figure 3.31, $V_{tp} = -0.8$ V and $K_p = 100\ \mu A/V^2$. What is the required W/L ratio if M1 is to pass 0.5 A and keep $V_{SD} < 0.1$ V.

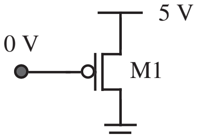


Figure 3.31.

These many examples and exercises with MOS transistors have a purpose. These problems, when combined with transistor family of curves plots, should now allow you to think in terms of a transistor's reaction to its voltage environment. This is basic to electronics engineering instruction. It should allow you to quickly anticipate and recognize aberrations caused by defective circuits, and later to predict what category of defect exists. The techniques needed to solve these problems should become reflexive.

3.2 THRESHOLD VOLTAGE IN MOS TRANSISTORS

Until now, we assumed that the transistor source and substrate terminals were connected to the same voltage. This is valid for isolated transistors, but when transistors are connected in CMOS circuits, this condition may not hold for all devices.

Figure 3.32 is a circuit cross section of two *n*MOS and one *p*MOS transistors fabricated in a CMOS process. All devices are constructed on the same *p*-type silicon substrate. Since *p*MOS transistors are formed on *n*-type substrates, there must be a region of the circuit that is oppositely doped to the initial bulk, forming what is called a *well.*

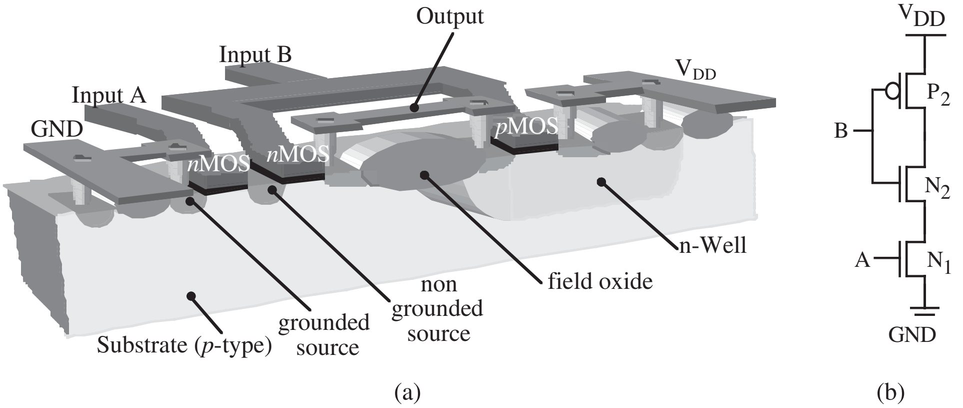


Figure 3.32. (a) Structure for two series-connected *n*MOS transistors and one *p*MOS transistor in a CMOS technology. (b) Circuit schematic.

The *p*-type substrate for *n*MOS transistors is connected to zero (or ground, GND), whereas the *n*-type well is connected to V_{DD}, since it forms the bulk of the *p*MOS transistors. The source of the *n*MOS device N_1 is connected to ground, so that previous equations are valid for this device. The transistor N_2 source is connected to the drain of N_1 to make a series connection of both devices, required to implement the operation of the logic gate (a detailed analysis of transistor interconnection to form gates is given in Chapter 5). As a result, the source of transistor N_2 is not grounded, and it can acquire voltages close to V_{DD}, whereas its substrate is connected to ground through the polarizing contact. Therefore, the condition $V_{SB} = 0$ will not hold in some bias cases for transistor N_2.

When the source and substrate voltages differ, the gate–source voltage is not fully related to the vertical electric field responsible for creating the channel. The effect of the higher source voltage above (below) the substrate for an *n*MOS (*p*MOS) transistor is to lower the electric field induced from the gate to attract carriers to channel. The result is an effective raising of the transistor threshold voltage. The threshold voltage can be estimated as [8]

$$V_t = V_{t0} \pm \gamma\sqrt{V_{SB}} \qquad (3.15)$$

where V_{t0} is the threshold voltage when the source and the substrate are at the same voltage, and γ is a parameter dependent on the technology. The parameter γ is called the body effect constant. The positive sign is used for *n*MOS transistors, and the negative sign for *p*MOS transistors. When the source and substrate are tied together, $V_{SB} = 0$, and the threshold voltage is constant.

The significance of the threshold body effect lies with certain circuit configurations whose transistor thresholds will be altered, generally being higher than expected. This can lead to conduction states and changes in transistor delay time. We will return to this topic when we discuss pass transistor properties, particularly in memories, and circuits such as that in Figure 3.32.

3.3 PARASITIC CAPACITORS IN MOS TRANSISTORS

We have learned the equations that describe the static operation of the transistor, i.e., the current into the device when voltage nodes remain stable with time, but the dynamic operation requires knowledge of other aspects of the devices.

One limitation of high-speed digital ICs is the time required to switch a transistor between the on- and off-states. This delay mechanism is primarily due to transistor parasitic capacitors that fall into two types: voltage-dependent and non-voltage-dependent capacitors. Non-voltage-dependent capacitors are characterized by physical overlap of the gate terminal with the drain and source areas. The voltage-dependent capacitors are the reverse-biased drain–substrate and source–substrate diodes, plus those characterized by the creation of depletion regions and conducting channels [3]. Another significant cause of delay in modern ICs is the capacitance of the interconnect wires between transistors.

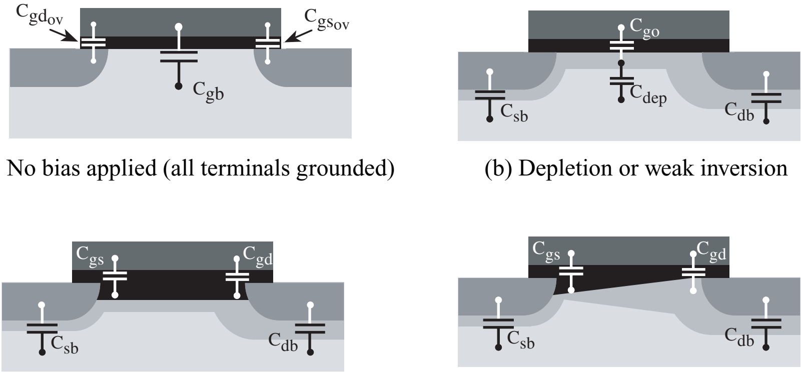
3.3.1 Non-Voltage-Dependent Internal Capacitors

Non-voltage-dependent capacitor values are calculated from device dimensions using a parallel plate model. A parallel plate capacitor has two conductors of area (A) separated by a distance (d). The space between the conductors can be empty or filled with an insulator to increase the capacitance value. The capacitance is

$$C_p = \left(\frac{\varepsilon_0 \varepsilon_{\mathrm{m}}}{d}\right) A \tag{3.16}$$

where ε_0 is the permittivity of free space, and ε_{m} is the relative permittivity of the material filling the capacitor. The non-voltage-dependent capacitors in a MOSFET device are C_{GDov} and C_{GSov} shown in Figure 3.33. Their value is found by applying Equation (3.16) to the overlap region between the gate and the drain.

$$C_{\mathrm{ov}} = \left(\frac{\varepsilon_0 \varepsilon_{\mathrm{Si}}}{T_{\mathrm{ox}}}\right) W \cdot L_{\mathrm{D}} \tag{3.17}$$



(a) No bias applied (all terminals grounded)

(b) Depletion or weak inversion

(c) Nonsaturation

(d) Saturation

Figure 3.33. Parasitic capacitors of MOS transistors for four operating regions.

where $\varepsilon_{\mathrm{Si}}$ is the relative permittivity of the silicon dioxide, W is the transistor width, L_{D} is the overlap distance between the gate and the drain or source, and T_{ox} is the gate oxide thickness.

3.3.2 Voltage-Dependent Internal Capacitors

Once the capacitor is biased, the depletion and inversion regions change the effective internal parasitic capacitors. These capacitors are voltage-dependent since their value is related to charge redistribution within the device. We will provide approximate expressions for such capacitors.

In all operation regions, consider the following three terminal capacitors:

C_{gb} Gate–substrate (gate–bulk) capacitor. It has a strong dependence on the transistor biasing.

C_{gs}, C_{gd} Gate–channel capacitors. They are divided into gate–drain and gate–source capacitors, since the channel distribution is not uniform along the device in saturation.

C_{sb}, C_{db} Source and drain-to-bulk capacitors. Their capacitance is due to the reverse-bias, built-in diodes.

The total voltage-dependent gate capacitance of a MOS is found by summing the capacitors:

$$C_{\mathrm{g}} = C_{\mathrm{gb}} + C_{\mathrm{gs}} + C_{\mathrm{gd}} \tag{3.18}$$

The drain/source-to-bulk capacitors do not impact the gate voltage. These capacitors can be calculated from the reverse diode capacitor expression in Equation 2.12, so that

$$C_{\mathrm{xb}} = \frac{C_{\mathrm{xb0}}}{\left(1 - \dfrac{V_{\mathrm{xb}}}{V_{\mathrm{bi}}}\right)^{1/2}} \tag{3.19}$$

where x must be replaced by d or s to refer to the drain and source terminals respectively. The value of C_{g} must be computed for each transistor operation region.

No Biasing (All Terminals at the Same Voltage). The gate–substrate capacitance is due to the MOS structure, and is calculated as a parallel plate capacitor filled by the gate oxide. This "physical" capacitor is referred to as C_{g0} and has the expression

$$C_{\mathrm{gb}} = C_{\mathrm{g0}} = \frac{\varepsilon_0 \varepsilon_{\mathrm{Si}}}{T_{\mathrm{ox}}} W \cdot L_{\mathrm{eff}} \tag{3.20}$$

where L_{eff} is the transistor effective length. It differs from the physical length L drawn in the design because the drain and source regions diffuse under the gate making $L_{\mathrm{eff}} = -2L_{\mathrm{overlap}}$. The only difference between Equations (3.17) and (3.20) is the capacitor area. In Equation (3.20) the area is that of the whole device, whereas in Equation (3.17) the area was only the overlap region between the gate and the drain or source.

Depletion or Weak Inversion. Although conceptually different, depletion and weak inversion will be treated as equivalent. In these regions, the gate voltage is not sufficient to create a channel, and the electric field from the gate induces a depletion region within the bulk (Figure 3.33(b)). The gate capacitor is the series connection of C_{g0} and the depletion capacitor C_{dep}. The computation of the depletion capacitance is complicated, and beyond the scope of this book.

Nonsaturation. When the transistor is in nonsaturation, the conducting channel "touches" the drain and source terminals (Figure 3.33(c)). Thus, C_{gs} and G_{gd} dominate and the gate–bulk capacitor is now negligible, since the conducting layer "disconnects" the gate from the bulk. As a result, the gate–bulk capacitor C_{g0} is shared between C_{gd} and C_{gs} so

$$C_{gs} = C_{gd} = \frac{\varepsilon_0 \varepsilon_{Si}}{2T_{ox}} W \cdot L_{eff} \tag{3.21}$$

Saturation. Once in saturation, the conduction channel no longer touches the drain end. The gate–drain channel capacitance is now negligible, and all gate capacitances are connected to the source terminal. The gate–source capacitance can be approximated as [8]

$$C_{gs} = \frac{2\varepsilon_0 \varepsilon_{Si}}{3T_{ox}} W \cdot L_{eff} \tag{3.22}$$

The simplifications made in the computation of the gate capacitance are summarized in Table 3.1.

Several forms of transistor capacitances have been described. Although we did not stress numerical work, you should assimilate the locations and different properties of each. Dynamic performance of an IC depends upon these capacitances, and also upon parasitic resistances of the transistors and interconnect wires. The latter elements are discussed in Chapter 9. We will next explore the influence on the basic transistor properties described thus far of the extreme scaling of dimensions of modern ICs.

3.4 DEVICE SCALING: SHORT-CHANNEL MOS TRANSISTORS

Device scaling is relentless in the microelectronics industry since transistor miniaturization allows faster device operation, yield improvement, and cost savings resulting from

Table 3.1. Simplified Intrinsic MOS Capacitor Expressions for Each Operating Region

	C_{gb}	C_{gs}	C_{gd}
Cutoff	$\frac{\varepsilon_0 \varepsilon_{Si}}{T_{ox}} W \cdot L_{eff}$	0	0
Ohmic	0	$\frac{\varepsilon_0 \varepsilon_{Si}}{2T_{ox}} W \cdot L_{eff}$	$\frac{\varepsilon_0 \varepsilon_{Si}}{2T_{ox}} W \cdot L_{eff}$
Saturation	0	$\frac{2\varepsilon_0 \varepsilon_{Si}}{3T_{ox}} W \cdot L_{eff}$	0

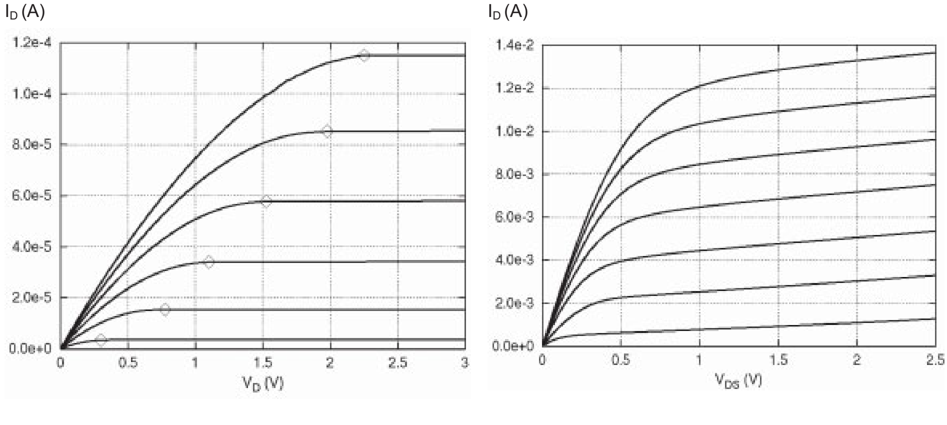
more dies per wafer. There is an important barrier encountered when minimum dimensions go below about 0.5 μm channel lengths since new device physical properties appear. Devices with dimensions more than 0.5 μm are usually called long-channel devices, whereas smaller transistors are called deep submicron or short-channel devices. The division at 0.5 μm is slightly arbitrary, but gives an approximate dimension for entering the short channel and deep submicron regions of electronics.

Figure 3.34 compares the output current characteristics for a long-channel transistor (1.2 μm technology) and a submicron technology device (0.25 μm technology). The major differences are

- The spacing between equally incremented drain–current versus gate–voltage curves is constant in short-channel transistors in the saturation region, whereas the long-channel transistor spacing between these curves increases nonlinearly for increasing gate voltages.
- Once the device reaches saturation, the current remains nearly flat for the long-channel device (it has small dependence on V_{DS} for this region), whereas for the short-channel transistor the saturated current shows more slope.
- The third difference relates to the total amount of current. If two transistors have equal gate size and are at the same gate and drain voltage, then the short-channel transistor passes more drain current than the long-channel one. This is largely due to the thinner oxides used in short-channel transistors.

We will identify the mechanisms causing these differences, and present equations describing these behaviors for short-channel transistors. Short-channel transistor analysis has more complexity than long-channel analysis. It requires detailed concentration on your part, but we hope that you will obtain an understanding of the difficulty of manual analysis of the sort similar to that for long-channel transistors. We acknowledge that we build on the impressive work of others, such as Foty [2], Tsividis [7], and Weste and Eshraghian [8].

There are many submicron effects, but we will focus on channel length modulation, velocity saturation, subthreshold current, DIBL, and hot-carrier effects, since they may have



(a) (b)

Figure 3.34. Current characteristics for (a) long-channel, and (b) short-channel transistors.

a significant impact on IC testing and reliability. Our goal is to understand short-channel effects, and develop tools that allow manual analysis as we did for long-channel transistors.

3.4.1 Channel Length Modulation

Channel length modulation appears when the device is in saturation and is related to the pinch-off effect described earlier. When the device reaches saturation, the channel no longer “touches” the drain and acquires an asymmetric shape that is thinner at the drain end (Figure 3.33(d)). As the drain–source voltage increases, the channel saturation or depletion region moves further away from the drain end because the drain electric field “pushes” it back. The reverse-bias depletion region widens, and the effective channel length decreases by an amount ΔL (Figure 3.35) for increasing V_{DS}. In large devices, the relative change of the effective channel length ΔL with respect to the total channel length with V_{DS} is negligible, but for shorter devices $\Delta L/L$ becomes important.

The effective length of the device varies with V_{DS} once the device is in saturation, and, as a result, the curves are no longer flat in this region compared to long-channel devices (Figure 3.34(b)).

The small drain current dependence on the drain–source voltage in the saturation region can be modeled by multiplying the saturation current by a slope factor that depends on this voltage. The resulting equation is

$$I_{\mathrm{D}} = I_{\mathrm{Dsat}}(1 + \lambda V_{\mathrm{DS}}) \tag{3.23}$$

where λ is a parameter that can be measured for each particular technology.

3.4.2 Velocity Saturation

When charged carriers move in a solid under the force of an electric field, they acquire a velocity proportional to the magnitude of this electric field. The applied electric field ($\mathscr{E}$) and the carrier velocity (v) are related through the mobility parameter μ, introduced in Chapter 2 as

$$v = \mu\mathscr{E} \tag{3.24}$$

For small electric fields, μ is constant and independent of the applied electric field.

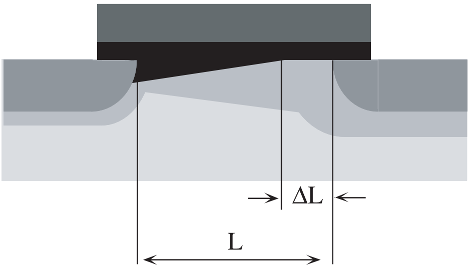


Figure 3.35. The channel length modulation effect.

$$\mu = \mu_0 \text{ (constant)} \tag{3.25}$$

As a result when carrier velocity (v) is plotted versus the applied electric field ($\mathscr{E}$), the result is a straight line (low electric field region of Figure 3.36).

The reason for this linear dependence between velocity and small electric fields is that electrons moving in semiconductors collide with silicon atoms, an effect known as scattering. Electron scattering is linear with small electric fields. If the electric field further increases, the carrier velocity enters a region in which it is said to move at *velocity saturation.* As device dimensions scale down, the electric fields within the transistor increase, making the velocity saturation more important. Velocity saturation due to mobility reduction is important in submicron devices.

If Equation (3.24) holds for carriers moving in small electric fields and carriers moving at the velocity saturation, then the mobility μ must change with the electric field (Figure 3.36). Two effects are combined in transistors to account for this mobility: reduction due to the horizontal electric field and mobility reduction due to the vertical electric field.

Horizontal Electric Field Mobility Reduction. Carriers in short-channel devices reach velocity saturation at lower values of V_{DS} than for long-channel transistors. Figure 3.37(a) illustrates the effect of V_{Dsat} moving toward small values, and shows that saturation current is smaller in short-channel transistors. This effect is due to channel length reduction that implies higher horizontal electric fields for equivalent drain–source voltages. The horizontal electric field within the channel is due to the voltage applied to the drain terminal. Horizontal mobility (μ_H) reduction can be related to the drain voltage by

$$\mu_H = \frac{\mu_0}{1 + \dfrac{V_{DS}}{L_{eff}\mathscr{E}_{crit}}} = \frac{\mu_0}{1 + \theta_2 V_{DS}} \tag{3.26}$$

$1/(L_{eff}\mathscr{E}_{crit})$ is a parameter called *drain bias mobility reduction* that in some texts is referred to as θ_2. $\mathscr{E}_{crit}$ is the electric field shown in Figure 3.36, and depends on the technology. For large transistors, $\theta_2 V_{DS}$ is smaller than 1, and, therefore, the mobility is constant ($\mu_H \approx \mu_0$), giving the linear relation between velocity and electric field for this region. When L_{eff} decreases ($L_{eff} \rightarrow 0$), θ_2 increases, and $\theta_2 V_{DS}$ becomes important, lowering the mobility below μ_0. Even for short-channel transistors, when V_{DS} is small, the denominator

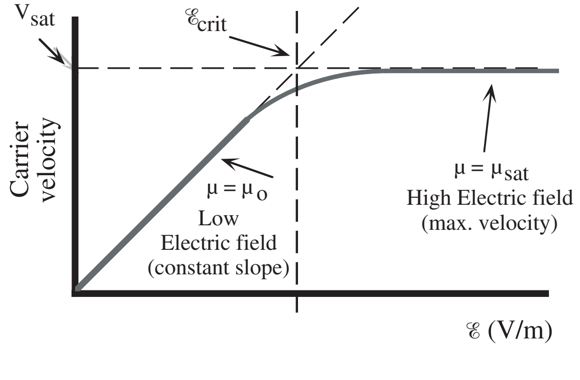
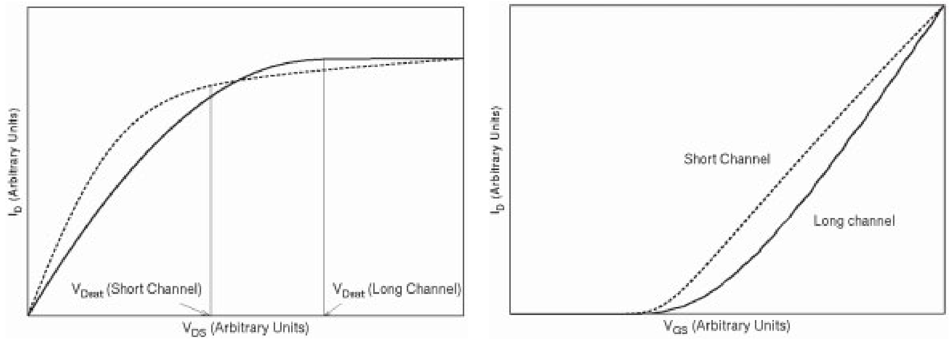


Figure 3.36. Electric field to carrier velocity dependence in a solid.



(a) (b)

Figure 3.37. Two effects of velocity saturation on submicron devices. (a) The saturation voltage V_{Dsat} moves to lower values, and (b) the I_{D} versus V_{GS} relationship becomes linear.

of Equation (3.26) is almost equal to 1, and the mobility is constant. When V_{DS} increases, the denominator of Equation (3.26) becomes much larger than one, and the mobility is reduced. Equation (3.26) is a simple model of mobility reduction in submicron transistors caused by the lateral (horizontal) electric field.

Vertical Electric Field Mobility Reduction. There is also a vertical electric field due to the gate voltage that creates the conduction channel. When carriers move within the channel under the effect of the horizontal field, they "feel" the effect of this gate–substrate-induced electrical vertical field, pushing carriers toward the gate oxide. This provokes carrier collisions with the oxide–channel interface, reducing their mobility. The interface of Si and $\mathrm{SiO_2}$ is rough and imperfect, so that carriers move with more difficulty.

The mobility reduction from this effect can be described in a similar way to that in the horizontal field mobility reduction [Equation (3.26)]. Now, the expression of the vertical mobility (μ_{V}) reduction will contain the gate–source voltage instead of the drain–source voltage:

$$\mu_{\mathrm{V}} = \frac{\mu_0}{1 + \theta_1(V_{\mathrm{GS}} - V_t)} \tag{3.27}$$

Similar to Equation (3.26), θ_1 is a parameter called *gate bias mobility reduction* that depends on the technology.

3.4.3 Putting it All Together: A Physically Based Model

We discussed two mechanisms that impact transistor behavior when technology scales down. Although other effects appear in submicron technologies, channel length modulation and mobility reduction are the two main effects used in the device model equations in CAD simulators. We now introduce model equations to describe the drain current of submicron devices. The objective is to give the reader basic and accurate expressions to account for the most relevant submicron effects. Device modeling is an active research area. For deeper discussions refer to [2] and [7].

The long-channel transistor drain-current square law dependence on the gate–source voltage for ohmic and saturation conditions are repeated from Equations (3.3) and (3.4):

$$\begin{aligned} I_D &= \mu_0 C_{ox}\left[\frac{W}{L}(V_{GS}-V_{tn})V_{DS}-\frac{V_{DS}^2}{2}\right] \qquad \text{(ohmic)} \\ I_D &= \mu_0 C_{ox}\frac{W}{L}(V_{GS}-V_{tn})^2 \qquad \text{(saturation)} \end{aligned} \tag{3.28}$$

where $C_{ox} = \varepsilon_{ox}/T_{ox}$. These expressions can be modified to incorporate the effects described previously for short-channel transistors.

Ohmic State. The ohmic region expression for short-channel devices is similar to Equation (3.28), including the velocity saturation effects:

$$I_D = \mu_0 C_{ox}\frac{W}{L_{eff}}\,\frac{(V_{GS}-V_t)V_{DS}-\frac{V_{DS}^2}{2}(1+\delta)}{[1+\theta_1(V_{GS}-V_t)]\left(1+\frac{V_{DS}}{L_{eff}\mathscr{E}_{crit}}\right)} \tag{3.29}$$

Equation (3.29) combines the horizontal and vertical mobility reduction [the two expressions in brackets of the denominator taken from Equations (3.26) and (3.27)]. The parameter δ relates the charge at the inversion channel to the surface potential and the oxide capacitance, and will not be described in detail.

In addition to the drain current expression, we need an expression for the saturation voltage V_{Dsat} that describes carrier saturation at smaller drain voltages (Figure 3.37(a)). This saturation voltage depends on the gate voltage, and is calculated by differentiating the drain current in the ohmic region [Equation (3.29) in this case] and setting the expression equal to zero. This leads to

$$V_{Dsat} = \frac{2(V_{GS}-V_t)}{(1+\delta)\left(1+\sqrt{1+\frac{2(V_{GS}-V_t)}{L_{eff}\mathscr{E}_{crit}(1+\delta)}}\right)} \tag{3.30}$$

The short-channel drain saturation voltage depends on the square root of the gate voltage, whereas for long-channel devices the term in the square root tends to 1, and the saturation voltage depends on $(V_{GS} - V_t)$.

Saturated State ($V_{DS} > V_{Dsat}$). The short-channel drain current expression in saturation is obtained from Equation (3.29), substituting V_{DS} by V_{Dsat} and including the channel modulation effect from Equation (3.23):

$$I_D = \mu_0 C_{ox}\frac{W}{L_{eff}}\,\frac{(V_{GS}-V_t)V_{Dsat}-\frac{V_{Dsat}^2}{2}(1+\delta)}{[1+\theta_1(V_{GS}-V_t)]\left(1+\frac{V_{Dsat}}{L_{ef}\mathscr{E}_{crit}}\right)}[1+\lambda(V_{DS}-V_{Dsat})] \tag{3.31}$$

Equations (3.29) and (3.31) describe the drain current of submicron MOS transistors for the ohmic and saturated states, respectively. The transition of the drain current between states described by these equations occurs sharply and can lead to computation errors from derivative discontinuities. Equation (3.32) [7] describes the drain current for a MOS transistor using only one expression, which is valid for ohmic and saturated states. It is taken from Equation (3.31), where V_{Dsat} is substituted by V_{DS1}. V_{DS1} is an internal drain voltage that makes a smooth transition from V_{DS} in Equation (3.29) to V_{Dsat} in Equation (3.31) when changing from the ohmic state to the saturated one, as shown in Figure 3.38:

$$I_{\text{D}} = \mu_0 C_{\text{ox}} \frac{W}{L_{\text{eff}}} \frac{(V_{\text{GS}} - V_t)V_{\text{DS1}} - \dfrac{V_{\text{DS1}}^2}{2}(1 + \delta)}{[1 + \theta_1(V_{\text{GS}} - V_t)]\left(1 + \dfrac{V}{L_{\text{eff}}\mathscr{E}_{\text{crit}}}\right)} [1 + \lambda(V_{\text{DS}} - V_{\text{DS1}})] \tag{3.32}$$

Figure 3.38 shows that the internal voltage V_{DS1} is mainly the drain–source voltage for small applied V_{DS}. Once the device enters saturation, the internal drain–source voltage becomes V_{Dsat}.

Figure 3.39 compares measured drain current for *n*MOS and *p*MOS transistors versus the drain current equation [Equation (3.32)], showing that the model fits the measured transistor curves.

3.4.4 An Empirical Short-Channel Model for Manual Calculations

Equations (3.29) and (3.31) have a physical basis. They provide understanding of how these physical effects impact the device characteristics, but lack simplicity for hand calculations. Another approach derives mathematical expressions that fit transistor curves over both bias ranges. These expressions use empirical parameters to specifically match measured device characteristics. These parameters typically do not have a physical relation with the underlying device mechanism, and are called fitting parameters (they fit a curve with a mathematical expression).

Empirically based models afford the possibility of deriving a mathematical expression for hand calculations. This section introduces an empirical model for short-channel tran-

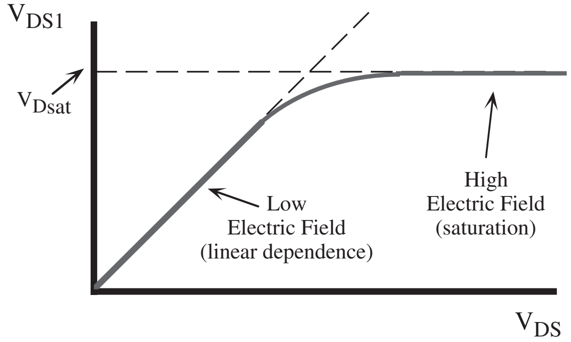


Figure 3.38. Internal drain–source voltage versus applied drain source voltage [V_{DS1} in Equation (3.29)].

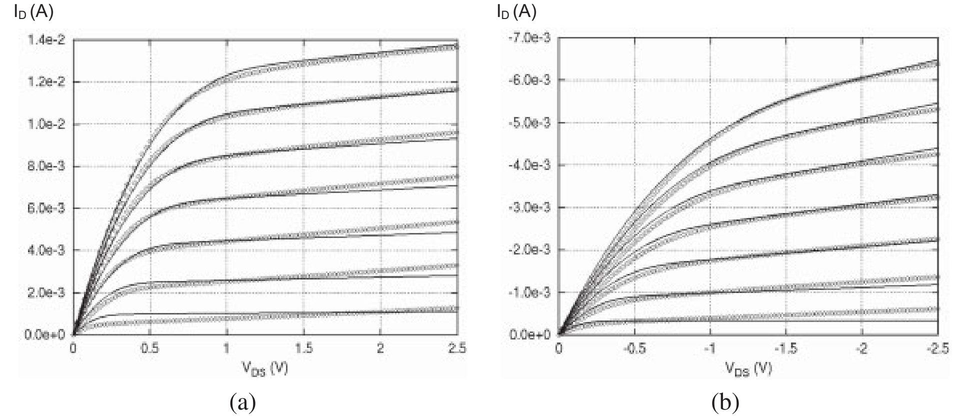


Figure 3.39. Comparison of experimental data (diamonds) with the drain current model of Equation (3.32) (lines) for submicron (a) n-type and (b) p-type transistors.

sistors applicable for manual calculations. These equations can compute voltage and current in simple circuits with just a few transistors. This is valuable when computing electrical parameters in small circuits in which fabrication defects are present. This empirical model is based on the Sakurai model [5].

We begin with the transistor in saturation. The model introduces a maximum current parameter I_{D0} that is the current when $V_{DS} = V_{GS} = V_{DD}$ (V_{DD} is the maximum voltage in the circuit). The drain current in saturation has a small linear dependence with the drain voltage. This dependency is described with a second parameter, λ, similar to Equation (3.23).

Finally, we know that the drain current has a quadratic dependence with the gate voltage for long-channel devices and a linear dependence for submicron devices [Figure 3.37(b)]. To account for this variation, we use a third fitting parameter called α (taken from Sakurai [5]). This parameter is equal to 1 for short-channel devices and equal to 2 for long-channel transistors, and can be somewhat related to carrier velocity saturation. A mathematical expression for the drain current in saturation is

$$I_D = I_{D0}\left(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}}\right)^{\alpha}[1 + \lambda(V_{DS} - V_{DD})] \qquad \text{for} \qquad V_{DS} > V_{Dsat} \tag{3.33}$$

We "constructed" this equation to fulfill these conditions:

- The drain current is equal to I_{D0} when $V_{GS} = V_{DD}$ and $V_{DS} = V_{DD}$.
- The drain current dependence on the drain–source voltage is linear with slope λ.
- When $\alpha = 1$ (submicron transistors), the drain current dependence with the gate voltage is linear, and for large devices the dependence is quadratic ($\alpha = 2$).

Equation (3.33) was constructed to fit experimental results. None of the parameters introduced are founded on physical phenomena in the transistor, although we could provide a physical meaning for some of them.

Equation (3.33) holds when the drain voltage is beyond the saturation voltage V_{Dsat} that is empirically defined as

$$V_{\text{Dsat}} = V_{\text{D0}}\left(\frac{V_{\text{GS}} - V_{th}}{V_{\text{DD}} - V_{th}}\right)^{\alpha/2} \tag{3.34}$$

where V_{D0} is the saturation voltage when $V_{\text{DS}} = V_{\text{GS}} = V_{\text{DD}}$, and the exponential dependence with $\alpha/2$ was observed experimentally.

We next need an expression for the drain current in nonsaturation. In this region, drain current has a quadratic dependence on the drain voltage, and when $V_{\text{DS}} = V_{\text{Dsat}}$ its value must match the current value for saturation from Equations (3.33) and (3.34). This gives

$$I_{\text{DS}} = \left(2 - \frac{V_{\text{DS}}}{V_{\text{Dsat}}}\right)\frac{V_{\text{DS}}}{V_{\text{Dsat}}} I_{\text{D0}}\left(\frac{V_{\text{GS}} - V_{th}}{V_{\text{DD}} - V_{th}}\right)^{\alpha}[1 + \lambda(V_{\text{Dsat}} - V_{\text{DD}})] \quad \text{for} \quad V_{\text{DS}} < V_{\text{Dsat}} \tag{3.35}$$

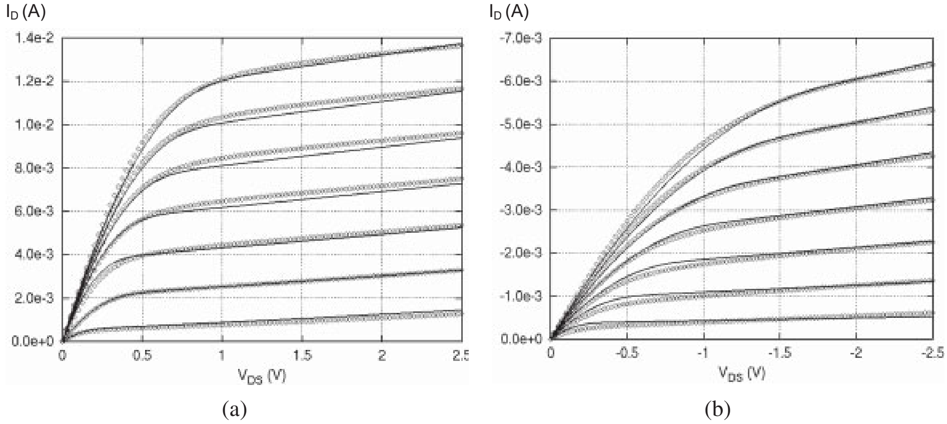
We introduced no new parameters for this region. The reader can easily verify that Equations (3.35) and (3.33) have the same expression for $V_{\text{DS}} = V_{\text{Dsat}}$. It is also easy to verify that for $\lambda \neq 0$ there is a slope discontinuity at this point. A discontinuity is not desirable for models used in CAD tools or simulators because convergence problems must be avoided. In our case, the aim of this model is hand calculations in which this problem does not arise.

Combining both expressions and neglecting subthreshold leakage:

$$I_{\text{D}} = \begin{cases} 0 & \text{when } V_{\text{GS}} < V_{th} \\ \left(2 - \dfrac{V_{\text{DS}}}{V_{\text{Dsat}}}\right)\dfrac{V_{\text{DS}}}{V_{\text{Dsat}}} I_{\text{D0}}\left(\dfrac{V_{\text{GS}} - V_{th}}{V_{\text{DD}} - V_{th}}\right)^{\alpha}[1 + \lambda(V_{\text{Dsat}} - V_{\text{DD}})] & \text{when } V_{\text{DS}} < V_{\text{Dsat}} \\ I_{\text{D0}}\left(\dfrac{V_{\text{GS}} - V_{th}}{V_{\text{DD}} - V_{th}}\right)^{\alpha}[1 + \lambda(V_{\text{DS}} - V_{\text{DD}})] & \text{when } V_{\text{DS}} \geq V_{\text{Dsat}} \end{cases} \tag{3.36}$$

with V_{Dsat} given by Equation (3.34).

Figure 3.40 compares experimental data for a submicron process (0.25 μm) and the



(a) (b)

Figure 3.40. Comparison of experimental data for (a) *n*MOS, and (b) *p*MOS short-channel transistors with the empirical model of Equation (3.36).

empirical model. The fitting parameters are $\alpha = 1.12$, $I_{D0} = 13.46$ mA, $V_{D0} = 1$ V, $\lambda = 0.08$, $V_{DD} = 2.25$ V, and $V_{th} = 0.64$ V.

■ EXAMPLE 3.11

Calculate I_D and V_{DS} for the circuit in Figure 3.41 using the empirical model with the parameters $\alpha = 1.12$, $I_{D0} = 13.46$ mA, $V_{D0} = 1$ V, $\lambda = 0.08$, and $V_{th} = 0.64$ V.

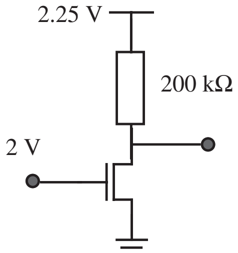


Figure 3.41.

First, find the region of operation for the device. We initially assume that the device is in saturation. From Equation (3.33), the drain current would be

$$\begin{aligned} I_D &= I_{D0}\left(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}}\right)^{\alpha}[1 + \lambda(V_{DS} - V_{DD})] \\ &= 13.46 \text{ mA}\left(\frac{2 - 0.64}{2.25 - 0.64}\right)^{1.12}[1 + 0.08(V_{DS} - 2.25)] \\ &= (9.136 + 0.8914 V_{DS}) \text{ mA} \end{aligned}$$

Applying KVL to the circuit,

$$\begin{gathered} V_{DD} = V_{DS} + I_D R \\ V_{DS} = V_{DD} - I_D R \\ V_{DS} = 2 \text{ V} - I_D \, 200 \ \Omega \end{gathered}$$

Solving both equations:

$$V_{DS} = 0.147 \text{ V}$$

We must verify that the device is in saturation, so we calculate V_{Dsat} and make sure that $V_{DS} > V_{Dsat}$. Using Equation (1.34),

$$V_{Dsat} = V_{D0}\left(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}}\right)^{\alpha/2}$$

$$= 1\left(\frac{2 - 0.65}{2.25 - 0.65}\right)^{0.56}$$

$$= 0.91\ \text{V}$$

since $V_{\text{DS}} < V_{\text{Dsat}}$, the device is in the ohmic region. We must recalculate the drain current using Equation (3.35):

$$I_{\text{D}} = \left(2 - \frac{V_{\text{DS}}}{V_{\text{DSsat}}}\right)\frac{V_{\text{DS}}}{V_{\text{DSsat}}}\, I_{\text{D0}}\left(\frac{V_{\text{GS}} - V_{th}}{V_{\text{DD}} - V_{th}}\right)^{\alpha}[1 + \lambda(V_{\text{DS}} - V_{\text{DD}})]$$

$$\text{Combining with } I_{\text{D}} = \frac{2.25 - V_{\text{DS}}}{200}$$

$$12.01V_{\text{DS}}^2 - 26.86V_{\text{DS}} + 11.5 = 0$$

Solving with the KVL equation derived above we get

$$V_{\text{DS}} = 0.56\ \text{V} \qquad \text{and} \qquad V_{\text{DS}} = 1.68\ \text{V}$$

This solution is valid for the device to be in ohmic state. Therefore, the solution of the problem is

$$I_{\text{D}} = 8.47\ \text{mA}$$

$$V_{\text{DS}} = 0.56\ \text{V}$$

■

Self-Exercise 3.13

Repeat the problem of Example 3.11 with V_{GS} =1.5 V.

Self-Exercise 3.14

Repeat the problem of Example 3.11 with V_{GS} = 0.75 V.

Short-channel transistors require specific constants for each technology, and they are not intuitive. As a result, many engineers still use long-channel equations for the "back of the envelope" estimations, acknowledging the increased error. Computers use complicated models to obtain accurate results [2], but do not give a feel for the underlying electronic physics. This conflict of rapid, more inaccurate hand calculations versus accurate computer calculation is unavoidable. We need an accurate approach, and we need an approach that rapidly gives us insight into physical behavior.

3.4.5 Other Submicron Effects

MOSFET devices show other effects when scaled down. We will not give analytical descriptions for these effects, but will briefly introduce them. The effects are: *subthreshold current, drain-induced barrier lowering (DIBL),* and *hot carrier effects.*

Subthreshold Current. Subthreshold current is the drain–source current when the gate–source voltage is below the transistor threshold voltage. The threshold voltage distinguishes the conduction from the nonconduction states of a MOS transistor. This gives the threshold voltage a vague definition. The transition from the conducting to the nonconducting state is not sharp, but continuous. This means that when the gate–substrate voltage increases, the charge in the channel is not created abruptly, but appears gradually with V_{GS}. There is a range of gate voltages lower than V_t for which there are carriers in the inversion layer that contribute to the drain current. In this region, the number of carriers that constitute the channel varies exponentially with the gate voltage.

We want no current when the transistor is in the off-state, i.e., when the gate voltage is below the threshold voltage. If V_t is large so that $|V_{GS}| < |V_t|$, then the number of carriers into the channel approaches zero. However, a high V_t increases the time required to switch between the on and off conducting states, resulting in slower devices.

Smaller transistors require lower operating voltages to restrict the internal electric field within reasonable margins. This in turn requires lowering the threshold voltage to maintain the operating speed of the device. This tendency is used in today's processes to maintain circuit performance at the cost of power increase. Modern devices show considerable current leakage even at $V_{GS} = 0$.

Threshold voltage reduction increases transistor leakage since an appreciable subthreshold current occurs during the off-state of the transistor. The subthreshold current can be easily observed from the device characteristics by plotting the logarithm of I_D versus V_{GS} for the subthreshold region (Figure 3.42). The subthreshold slope (S_t) is the amount of gate voltage required to increase the drain current one decade and is measured from Figure 3.42.

Subthreshold current has impact at the circuit level, since it is a fixed current contribution from all devices in the off-state. A subthreshold current of 10 nA at $V_{GS} = 0$ is insignificant for a single device, but in a 100 million transistor circuit the impact on the overall power consumption can be significant. Some technologies have MOSFETs with two different V_t's to reduce this problem. This separates the design into high-speed and low-speed transistors. High speed devices with lower V_t contribute higher leakage, and are

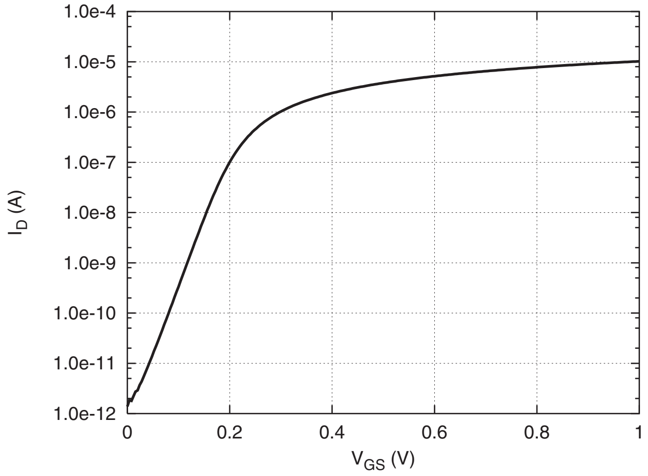


Figure 3.42. Effect of the subthreshold current in submicron devices.

used in critical delay paths or circuit blocks that must operate at high speed. Circuit blocks that do not require high speed are designed with high V_t transistors, and contribute less to the overall leakage.

Drain-Induced Barrier Lowering (DIBL). The population of channel carriers in long-channel devices is controlled by the gate voltage through the vertical electric field, whereas the horizontal field controls the current between the drain and the source. In large-channel devices, the horizontal and vertical electrical fields can be treated as having separate effects on the device characteristics. When the device is scaled down, the drain region moves closer to the source, and its electric field influences the whole channel. In this situation, the drain-induced electric field also plays a role in attracting carriers to the channel without control from the gate terminal. This effect is called *drain-induced barrier lowering* (DIBL), since the drain lowers the potential barrier for the source carriers to form the channel. The threshold voltage "feels" the impact of this effect on the transistor curves. Since DIBL attracts carriers into the channel with a loss of gate control, DIBL lowers the threshold voltage and increases off-state leakage.

An expression of the subthreshold leakage of a MOSFET including DIBL is given by Chandrakasan et al. [1]:

$$I_{\text{subth}} = A \cdot e^{(1/nV_{\text{T}})V_{\text{GS}}-V_{t0}-\gamma \cdot V_{\text{S}}+\eta \cdot V_{\text{DS}})} \cdot (1 - e^{-V_{\text{DS}}/V_{\text{T}}}) \tag{3.37}$$

with

$$A = \mu_0 \cdot C_{\text{ox}} \frac{W}{L} (V_{\text{T}})^2 e^{1.8} e^{-\Delta V_{\text{TH}}/\eta V_{\text{T}}} \tag{3.38}$$

where μ_0 is the zero bias mobility, $V_{\text{T}} = kT/q$ is the thermal voltage, γ is the linearized body effect coefficient, η is the DIBL coefficient, and n is the subthreshold swing coefficient. The term ΔV_{TH} is introduced to account for transistor-to-transistor leakage variations.

Hot Carrier Effects. In saturation, carriers crossing from the inverted channel pinch-off point to the drain travel at their maximum saturated speed, and so gain their maximum kinetic energy. These carriers collide with atoms of the bulk, causing a weak avalanche effect that creates electron–hole pairs. These carriers have high energy, and are called hot carriers.

Some carriers interact with the bulk, giving rise to a substrate current. Hot carriers can significantly affect reliability since some of these carriers generated by impact ionization have enough energy to enter the gate oxide and cause damage (traps) in the SiO_2 region. The accumulation of such traps can gradually degrade device performance, change the device V_t, and can increase conduction through the oxide, giving rise to oxide wearout and breakdown. This phenomena and its effects are detailed in Chapter 6.

Very Short Channel Devices. When the channel length is drastically reduced, the horizontal electric field increases and the effects of velocity saturation become much stronger. Horizontal mobility reduction becomes more important than vertical mobility reduction, and the drain saturation voltage is reduced. In these cases, Equation (3.32) can be rewritten neglecting the quadratic term in the drain internal voltage (V_{DS1}) since it is very small, and considering only the horizontal mobility effect (since it becomes predominant), leading to

$$I_D \approx \mu_0 C_{ox} \frac{W}{L_{eff}} \frac{(V_{GS} - V_t)V_{DS1}}{1 + \dfrac{V_{DS1}}{L_{eff}\mathscr{E}_{crit}}} \tag{3.39}$$

and for $V_{DS1}/(L_{eff}\mathscr{E}_{crit}) \gg 1$ gives

$$I_D \approx WC_{ox}(V_{GS} - V_t)\mu\mathscr{E}_{crit} \tag{3.40}$$

This assumption and the result states that for very short channel devices, carriers are always velocity saturated, and the drain current does not depend on the transistor length. Equations (3.32) and (3.40) apply to transistors saturated and ohmic states under the assumption that $V_{DS1}/(L_{eff}\mathscr{E}_{crit}) \gg 1$.

3.5 SUMMARY

This chapter examined MOSFET transistors using the physical description of semiconductors and diodes in Chapter 2. Transistor operation was explained with figures showing the interaction of gate, drain, source, and bulk regions with external bias, minority carrier inversion, and diodes. Abundant examples with *n*MOS and *p*MOS transistors model equations emphasized reflexive approaches to analyze circuits with transistors and resistors. The chapter closed with descriptions of the body effect and short-channel transistors. A modeling approach was given for short-channel transistors that stretches the outer limit of manual calculations for transistors.

REFERENCES

1. A. Chandrakasan, W. Bowhill, and F. Fox (Eds.), *Design of High-Performance Microprocessor Circuits,* IEEE Press, 2001.
2. D. Foty, *MOSFET Modeling with SPICE,* Prentice-Hall, 1997.
3. R. F. Pierret, *Semiconductor Device Fundamentals,* Addison-Wesley, 1996.
4. J. Rosselló and J. Segura, “Charge-based analytical model for the evaluation of power consumption in submircrometer CMOS buffers,” *IEEE Transactions on Computer Aided Design, 21,* 4, 433–448, April 2002.
5. T. Sakurai and R. Newton, “Delay analsysis of series-connected MOSFET circuits,” *IEEE Journal of Solid-State Circuits, 26,* 122–131, February 1991.
6. Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices,* Cambridge University Press, 1998.
7. Y. Tsividis, *Operation and Modeling of The MOS Transistor,* 2nd ed., McGraw-Hill, 1999.
8. N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design,* Addison-Wesley, 1993.

EXERCISES

3.1. For the three circuits in Figure 3.43, (a) give the transistor bias state, (b) write the appropriate model equation, and (c) calculate I_D, where $V_{tp} = -0.4$ V and $K_p = 100$ μA/V^2.

3.2. Repeat the same steps as in the previous exercise if $V_{tn} = 0.4$ V and $K_n = 200$ μA/V^2 (Figure 3.44).

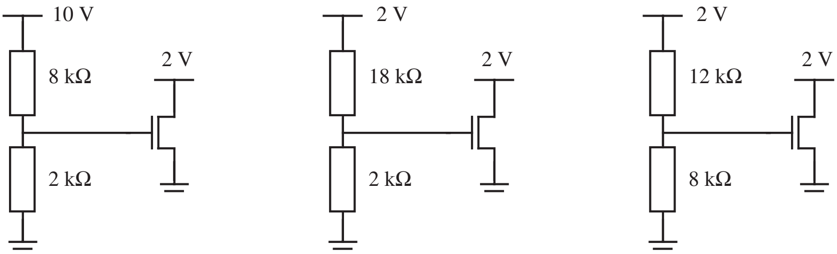


Figure 3.43.

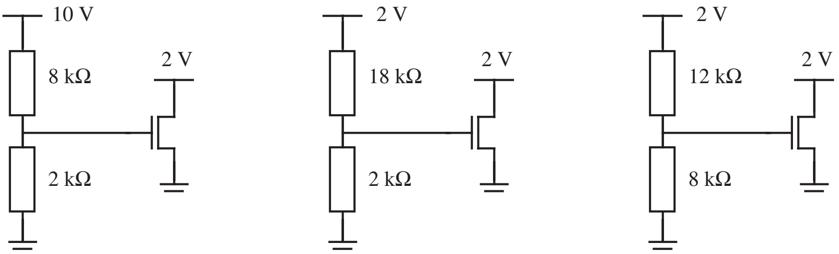


Figure 3.44.

3.3. Given $V_{tp} = -0.6$ V and $K_p = 75$ μA/V^2, and $W/L = 5$ (Figure 3.45), (a) solve for source voltage V_S, (b) solve for drain voltage.

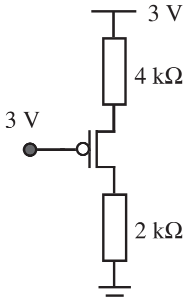


Figure 3.45.

3.4. Given the circuit in Figure 3.46 and the transistor parameters of Problem 3.2, (a) find the value of R_D to satisfy $V_0 = 1.2$ V. (b) As V_{DD} drops, find the value of V_D at the transition point where the transistor enters saturation and the new value of V_{DD}.

3.5. The transistor parameters in the circuit in Figure 3.47 are: $V_{tp} = -0.5$ V, $K_p = 75$ μA/V^2, and $W/L = 4$. If $V_0 = 1.2$ V, what is V_{IN}?

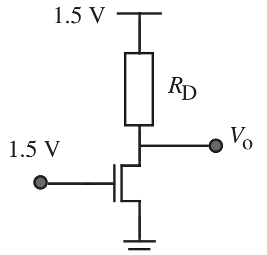


Figure 3.46.

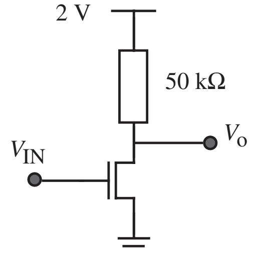


Figure 3.47.

3.6. Calculate V_G in Figure 3.48 so that $I_D = 200$ μA, given that $V_{tn} = 0.8$ V and $K_n = 100$ μA/V^2, and $W/L = 4$.

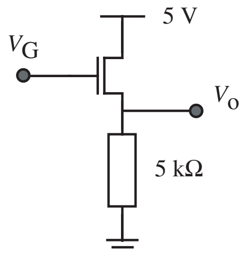


Figure 3.48.

3.7. Given that $R_1 = R_2$ and $K_n = 200$ μA/V^2, determine the resistance values in Figure 3.49 so that $V_D = 1$ V and $V_S = -1$ V.

3.8. Given $V_{tp} = -0.6$ V, $K_p = 50$ μA/V^2, $W/L = 3$, and $V_D = 0.8$ V. If $V_0 = 1.2$ V, what is R in Figure 3.50?

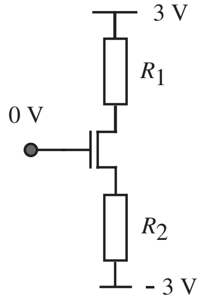


Figure 3.49.

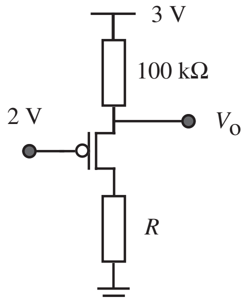


Figure 3.50.

3.9. In the circuit in Figure 3.51 $V_{tp} = -0.6$ V, $K_p = 75$ μA/V^2, and $W/L = 2$. (a) What value of R will place the transistor on the boundary between saturation and ohmic? (b) If R doubles its value, what are I_D and V_0?

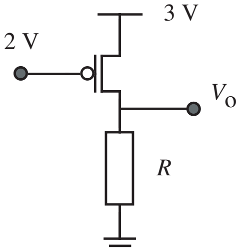


Figure 3.51.

3.10. The *p*MOSFET in Figure 3.52 has $V_{tp} = -0.5$ V, $K_p = 150$ μA/V^2, and $W/L = 3$, and a body effect constant $\gamma = 0.1$. The bulk voltage is at –0.3 V. (a) Calculate I_D. (b) If $V_{IN} = -1$ V, find I_D.

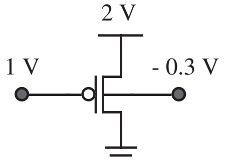


Figure 3.52.

3.11. Repeat Exercise 3.7 using the empirical model of Equation (3.36) with $\alpha = 1.14$, $I_{D0} = 14$ mA, $V_{D0} = 1$ V, $\lambda = 0.08$, and $V_{th} = 0.64$ V.

3.12. MOS transistors have two forms of capacitance. Describe each type and where you find them in the device.

3.13. What are the major differences between a short-channel transistor and a long-channel transistor?

3.14. Determine V_0 in the circuit of Figure 3.53, assuming that the topmost device is long channel, while the bottom one is short channel and needs the empirical model of Equation (3.36). Use: $V_{\mathrm{IN}} = 1.5$ V, $V_{tn(\mathrm{up})} = 0.4$ V, $V_{tn(\mathrm{dwn})} = 0.3$ V, $L_{\mathrm{up}} = 2$ μm, $W_{\mathrm{up}} = 30$ μm, $K_{\mathrm{up}} = 0.9$ mA/V^2, $I_{\mathrm{D0}} = 5$ mA, $\alpha = 1.2$, $\lambda = 0.09$, and $V_{\mathrm{D0}} = 1$V.

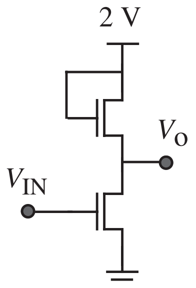


Figure 3.53.

3.15. Observe the log I_{D} versus V_{GS} curve in Fig. 3.22.
(a) Define subthreshold current for short channel transistors
(b) Why is subthreshold current in short channel transistors a problem in advanced CMOS technologies

3.16. Describe how DIBL lowers V_t in short channel transistors.

CHAPTER 4

CMOS BASIC GATES

4.1 INTRODUCTION

This chapter describes the electronics of basic logic gates, starting with the CMOS inverter whose simple appearance hides its complexity. Knowledge of inverter properties leads to knowledge of larger gates, such as NAND and NOR gates and their complicated properties. We will relate CMOS digital circuits to logic behavior and to CMOS failure mechanisms that typically involve small defects that alter normal inverter properties.

CMOS logic gates are digital cells, meaning that they perform Boolean algebra and their input and output voltages take one of the two possible logic states (high/low or 1/0). The output logic states have terminal voltages that respond to a range of input voltages but map into one of the two logic states. For example, a 1 V power supply technology has nominal output logic levels of 1 V (high) and 0 V (low). However, the input may range from 0 V to 0.3 V and the output still remains at a logic high of about 1 V.

This mapping of an input voltage range to a logic state gives noise immunity to digital circuits that is a major difference between analog and digital circuits. A small voltage fluctuation in an analog circuit node can cause significant error in the output signal. The same fluctuation in digital ICs is tolerated if it remains within the assigned range, and no error occurs.

There is a third range of digital voltage levels that is not mapped to any logic state. These voltages are between the logic levels, and they occur during an input/output voltage transition. None of the circuit nodes take these voltages in a normal or quiescent operation state since they have no logic meaning.

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

4.2 THE CMOS INVERTER

An inverter circuit converts a logic high-input voltage, such as 1 V, to a low logic voltage of 0 V (or 0 V to 1 V). The electronic symbol and truth table are shown in Figures 4.1(a) and (b). The Boolean statement is $V_{\text{out}} = V'_{\text{in}}$. The nMOS and pMOS transistors of the CMOS inverter (Figure 4.1(c)) act as complimentary switches. A logic high-input voltage turns on the nMOS transistor, driving the output node to ground, and also turns off the pMOS transistor. A low input voltage turns on the pMOS transistor and the nMOS off driving the output node to a high voltage.

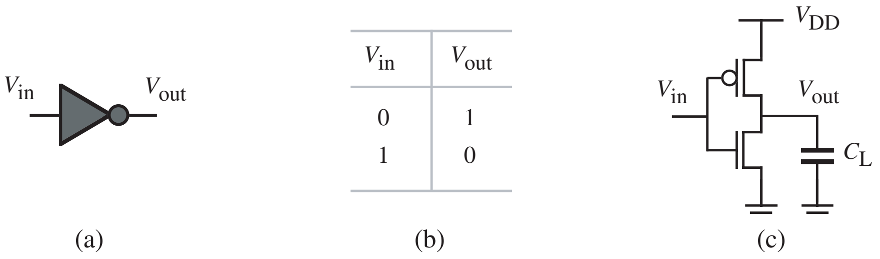
Boolean values are read in the quiescent state that occurs when all signal nodes settle to their steady-state values. Only one inverter transistor is on connecting the output terminal V_{out} to one of the power rails, and there is no current in the circuit since the other transistor is off, thus eliminating a DC path between the rails. A capacitor load C_{L} is shown in Figure 4.1(c) as it is unavoidable in any circuit. The capacitance is from transistor node and wiring capacitances and does not affect static properties, but hinders the speed of logic transitions. We will analyze the static and dynamic operation.

4.2.1 Inverter Static Operation

Voltage Characteristic. The static voltage characteristic measures the logic gate input and output voltage over the whole voltage range. This curve defines the voltage levels mapped to each logic state. Figure 4.2 shows an inverter static voltage transfer curve (V_{out} versus V_{in}). Noise margin refers to the amount of input signal variation allowed before the output voltage shows significant change. Noise margins are often simplistically defined at the points of the curve where the slope is –1. There are five bias state regions corresponding to the transistor operating regions.

These five regions are:

1. *Region I. nMOS off, pMOS ohmic.* This voltage range exists for $V_{\text{in}} < V_{tn}$. The nMOS transistor is off, and the pMOS transistor is driven into nonsaturation since $V_{\text{GS}} \approx -V_{\text{DD}} < V_{\text{DS}} + V_{tp}$ (see Equation 3.14). The pMOS drain node at V_{out} is pulled up to a logic high V_{DD} through the low impedance of the pMOS channel.
2. *Region II. nMOS saturated, pMOS ohmic.* V_{in} goes just above the nMOS threshold voltage ($V_{\text{in}} > V_{tn}$), and the nMOS transistor is barely turned on and in saturation



V_{in}	V_{out}
0	1
1	0

(a) (b) (c)

Figure 4.1. Inverter (a) symbol, (b) truth table, and (c) schematic.

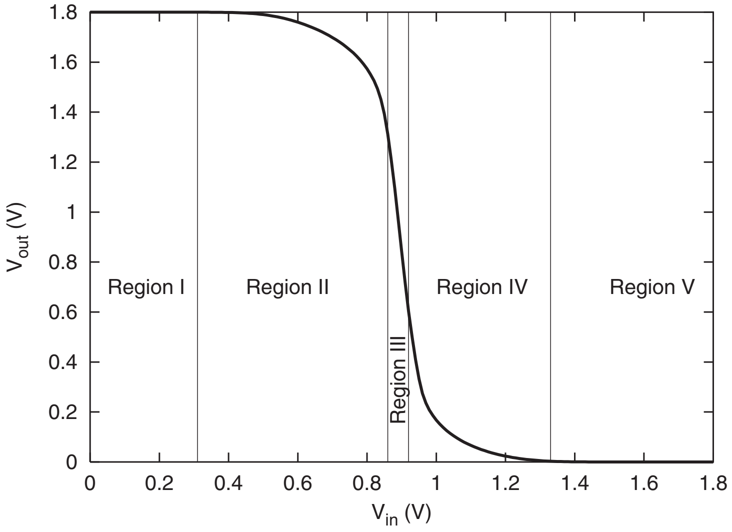


Figure 4.2. Inverter V_{in} versus V_{out} current transfer curve with five bias states.

($V_{\text{DS}} = V_{\text{out}} > V_{\text{in}} - V_{tn}$). Current now passes through both transistors and V_{out} drops as V_{in} is increased. The *p*MOS transistor remains in the ohmic state, but with decreasing gate drive.

3. *Region III. nMOS saturated, pMOS saturated.* When the output voltage goes below $V_{\text{in}} - V_{tp}$ and remains above $V_{\text{in}} - V_{tn}$, the *n*MOS and *p*MOS transistors are both in saturation, and the region has a straight line. Since V_{out} and V_{in} are linearly related, analog amplification occurs here. The drain voltage is a faithful replica of small changes in the input waveform, but amplified by a value equal to the slope of the straight line. MOS analog circuit designs use this property. It is also good for digital circuits that demand rapid V_{out} change during logic transitions of V_{in}. A digital goal is to get through the transition region as quickly as possible, and what better way than to have the circuit behave as an amplifier.
4. *Region IV. nMOS ohmic, pMOS saturated.* As V_{in} further increases, it approaches a value such that the difference between V_{in} and V_{DD} is close to the *p*MOS transistor threshold voltage. This is similar to Region II, but the roles of the transistors are reversed. The *p*MOS transistor is in saturation and the *n*MOS enters nonsaturation.
5. *Region V. nMOS ohmic, pMOS off.* When V_{in} goes to a logic high voltage, then $V_{\text{in}} \gg V_{\text{out}} + V_{tn}$. The *p*MOS transistor is turned off, and the *n*MOS transistor is in its ohmic state pulling the drain voltage V_{out} down to the source ground.

Inverter logic threshold voltage (V_{thr}) is the point at which $V_{\text{in}} = V_{\text{out}}$. V_{thr} is typically near $V_{\text{DD}}/2$. This is a unique condition since $V_{\text{in}} = V_{\text{out}}$ occurs only once in the inverter voltage range. The logic state changes as V_{in} moves through V_{thr}. V_{thr} is important when analyzing defect properties in CMOS circuits, since many defects cause intermediate voltages at some circuit nodes. Whether these defects cause a logic malfunction depends on the logic threshold voltage and the input voltage. Voltages slightly less than the logic high voltages and slightly more than logic low voltages are called weak logic voltages. Weak logic states are read correctly, but noise margins and gate driving voltages are compromised.

Except for Region I and Region V, the point at which transistors change from one zone to another depends on the inverter input and output voltages (Regions I and V depend only on the input). The input voltage at which these changes occur depends on the relative sizing of the devices, since the transistor width-to-length dimension (W/L) determines the current for a given gate–source voltage and, therefore, the effective equivalent resistance between drain and source.

In practice, input and output voltage ranges differ partly due to design and partly due to electrical noise. Figure 4.3 shows the voltage levels for the logic high and low values at a gate input and output. The following terms are defined

V_{IL} = input low voltage: maximum input voltage recognized as a logic low.

V_{IH} = input high voltage: minimum input voltage recognized as a logic high.

V_{OL} = output low voltage: maximum voltage at a gate output for a logic low for a specified load current.

V_{OH} = output high voltage: minimum voltage at a gate output for a logic high for a specified load current.

These voltage-based logic levels define the noise margin or immunity needed when connecting logic gates. *Noise Margin* (*NM*) is a parameter obtained from these levels and is defined for each logic value. NM_{H} and NM_{L} for the high and low logic values are

$$NM_{\mathrm{H}} = V_{\mathrm{OH}} - V_{\mathrm{IH}}$$

$$NM_{\mathrm{L}} = V_{\mathrm{IL}} - V_{\mathrm{OL}}$$

Noise margins must be positive for proper logic operation and the higher these values, the better the circuit noise immunity. These parameters are an essential measurement during production testing of ICs. Board designers must know that ICs that connect to each other are within specification and interface properly.

The logic threshold voltage of a CMOS inverter is set by the *p*MOS and *n*MOS transistor width and length ratios in the aspect ratio $(W_p/L_p)/(W_n/L_n)$ [see Figure 3.1(a)]. Usually, both transistors have the same channel length set by the minimum value of the technology.

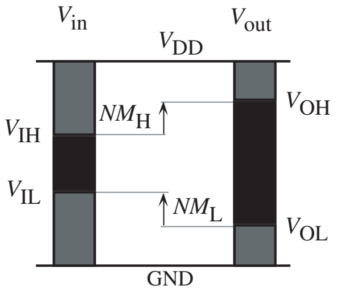


Figure 4.3. Voltage ranges mapped to logic Boolean values.

Therefore the W_p/W_n ratio determines the inverter logic threshold voltage. Inverters are often designed for a symmetric static transfer characteristic, so that the V_{out} versus V_{in} curve intersects the unity–gain line $V_{\text{in}} = V_{\text{out}}$ at about $V_{\text{DD}}/2$. A symmetrical transfer curve denotes a circuit with equal pullup and pulldown current drive strength. The aspect ratio giving a symmetric transfer characteristic for constant gate oxide thickness is

$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}\left(\frac{1-\dfrac{2V_{tn}}{V_{\text{DD}}}}{1-\dfrac{2|V_{tp}|}{V_{\text{DD}}}}\right)^2 \qquad (4.1)$$

Equation (4.1) is found by equating the saturation current for *n*MOS and *p*MOS transistors and setting the input voltage equal to $V_{\text{DD}}/2$. If it is possible for the technology to have $\mu_n = \mu_p$, and $V_{tn} = |V_{tp}|$, then a symmetric inverter requires $W_n = W_p$. In practice, electron and hole mobility are never equal, and the threshold voltages of *n*MOS and *p*MOS transistors have slightly different absolute values. Symmetrical CMOS inverter designs often have $(W_p/W_n) \approx 2$ to compensate for the smaller hole mobility in *p*MOS transistors.

Self-Exercise 4.1

Compute the ratio between *n*MOS and *p*MOS transistor width to obtain a symmetric inverter for a 0.18 μm technology in which $\mu_n = 360$ cm^2/V · s, $\mu_p = 109$ cm^2/V · s, $V_{tn} = 0.35$ V, and $V_{tp} = -0.36$ V with $V_{\text{DD}} = 1.8$ V.

The inverter logic threshold voltage can be made smaller or larger within the range $V_{tn} < V_{\text{thr}} < V_{\text{DD}}+V_{tp}$ by setting the ratio W_p/W_n above or below the value of Equation (4.1). $V_{\text{thr}} = 0.5V_{\text{DD}}$ when the pull-up and pull-down transistor current drive strength are equal. If $V_{\text{thr}} < 0.5V_{\text{DD}}$ then the *n*MOS pull-down transistor is stronger than the pull-up. If $V_{\text{thr}} > 0.5V_{\text{DD}}$ then the *p*MOS pull-up is stronger.

Current Characteristic. The DC power supply current transfer curve is equally important. Figure 4.4 shows the I_{DD} versus V_{in} characteristic. At $V_{\text{in}} < V_{tn}$ in Region I, the *n*MOS transistor is off and no current passes through the circuit. When $V_{\text{in}} = V_{\text{DD}}$ in Region V, the *p*MOS transistor is off and, again, no current passes from the power supply to ground. Typical inverter current at these quiescent logic levels is in the low pA's and is mostly drain–substrate reverse-bias saturation current or subthreshold current (Chapters 2–3). Virtually no power is dissipated in the quiescent logic states, giving CMOS ICs their traditional technology advantage. The peak current near $V_{\text{DD}}/2$ depends upon transistor drive strength (the size of the width-to-length ratio). Both transistors are in the saturated state and the current peaks. When an inverter changes logic state, the transient current is wasted power. Peak currents in large microprocessor designs are many Amperes, as this is the sum of the transient currents of millions of logic gates.

Transistor off-state leakage in deep submicron technologies is a concern and the major cause is the intentional threshold voltage reduction needed to maintain circuit performance. Deep submicron ICs must reduce V_{DD} to contain the internal electric fields and power dissipation, but V_t is kept at about 15%–25% of V_{DD} to ensure strong gate overdrive voltage. The design tradeoff is stronger gate overdrive (faster IC) with lower V_t versus significantly higher off-state leakage (higher off-state power). Total IC off-

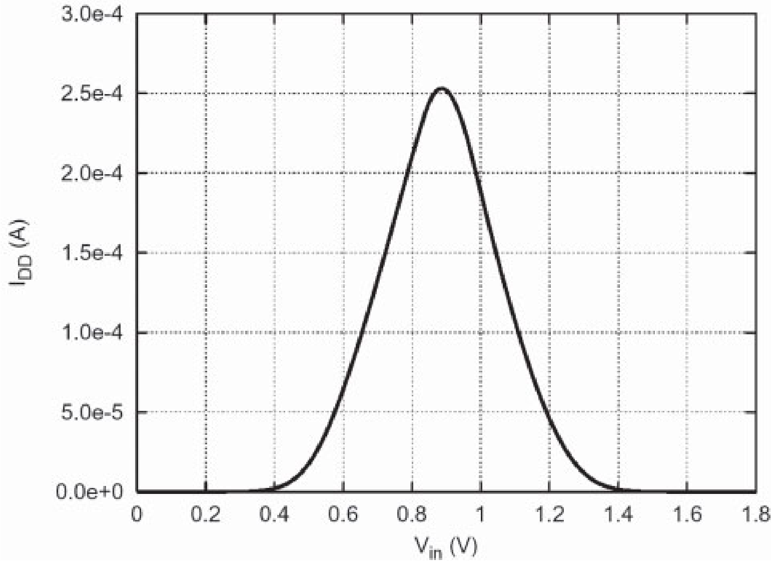


Figure 4.4. Inverter power supply current transfer curve.

state leakage can approach Ampere levels when high-speed performance is the dominant issue. Subthreshold current rises rapidly when V_t is lowered. One solution reduces off-state leakage by using transistors with different threshold voltages in the same circuit. This technique uses low V_t transistors in the speed-critical paths and high V_t devices in the noncritical paths. Another approach uses high V_t transistors to disconnect logic gates (designed with low V_t devices) from the power supply when they are inactive. Low V_t transistors have higher performance and higher leakage, whereas high V_t devices are slower with smaller leakage current.

Graphical Analysis of Bias Regions. Figure 4.5 distinguishes the bias state regions with the voltage transfer function and two 45° lines. The lower unity–slope line separates the bias condition, mutually satisfying saturation and nonsaturation for an *n*MOS:

$$V_{GS} = V_{DS} + V_{tn} \tag{4.2}$$

or

$$V_{in} = V_{out} + V_{tn} \Rightarrow V_{out} = V_{in} - V_{tn} \tag{4.3}$$

Equation (4.3) is a straight line superimposed on the voltage transfer curve in Figure 4.5 and labeled as (a). All points on the transfer curve lying above line (a) represent the *n*MOS transistor in saturation or the off-state. All points below line (a) represent the *n*MOS transistor in the ohmic state. A similar derivation leads to the *p*MOS transistor bias boundary line labeled (b) in Figure 4.5. The *p*MOS transistor is saturated or off for all points on the curve below line (b) and in the ohmic state above line (b). Both transistors are saturated in the region between the two lines.

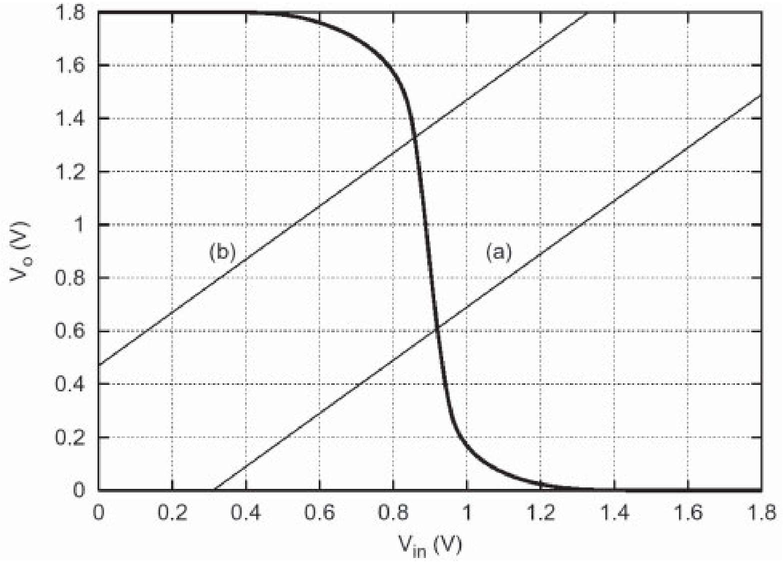


Figure 4.5. Inverter transfer curves and transistor state.

■ EXAMPLE 4.1

Figure 4.6 is an inverter transfer curve. Estimate V_{tn} and V_{tp} using bias line concepts.

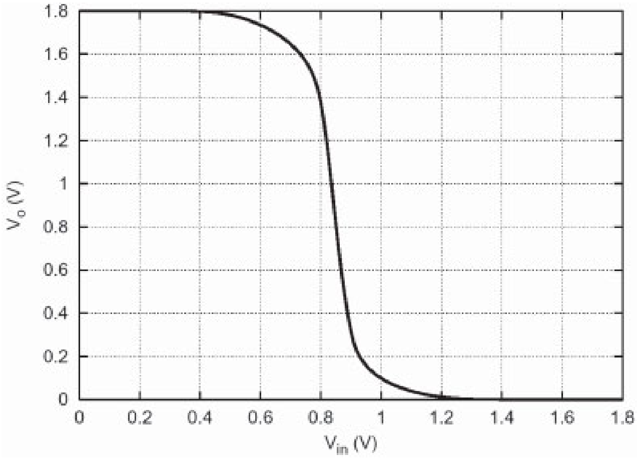


Figure 4.6. CMOS inverter voltage transfer curve.

Put a small mark on the ends of the linear region in Figure 4.6 and draw 45° lines. The threshold values are the intercepts. $V_{tn} = 0.42$ V and $V_{tp} = –0.44$ V. ■

Graphical analysis allows visualization of transistor states during logic transitions. The maximum gain region is where both transistors are saturated, as seen between the two dotted bias lines (a, b) in Figure 4.5. An example emphasizes this thinking.

■ EXAMPLE 4.2

When V_{out} switches in an inverter from V_{DD} to 0 V, estimate the fraction of this voltage range (V_{DD}) for which the nMOS transistor is in saturation. Let $V_{tn} = 0.2V_{\text{DD}}$, $V_{tp} = -0.2V_{\text{DD}}$, and $K'_n = K'_p$, where $K'_n = K_n\,(W/L)_n$, and $K'_p = K_p\,(W/L)_p$.

We know $I_{Dn} = -I_{Dp}$ for all the points in the static curve and also the point on line (a) in Figure 4.5 where the nMOS transistor leaves saturation. At this point, both transistors can be treated as in the saturation state. This is the transition between Regions III and IV in Figure 4.2. So,

$$K'_n(V_{\text{GS}} - V_{tn})^2 = K'_p(V_{\text{GS}} - V_{tp})^2$$

$$K'_n(V_{\text{in}} - V_{tn})^2 = K'_p(V_{\text{DD}} - V_{\text{in}} - V_{tp})^2$$

Solve for

$$V_{\text{in}} = \frac{V_{\text{DD}} + V_{tn}\sqrt{\dfrac{K'_n}{K'_p}} - V_{tp}}{1 + \sqrt{\dfrac{K'_n}{K'_p}}}$$

Substituting

$$V_{\text{in}} = V_{\text{out}} + V_{tn}$$

into the above V_{in} equation:

$$V_{\text{out}} = \frac{V_{\text{DD}} - V_{tn} - V_{tp}}{1 + \sqrt{\dfrac{K'_n}{K'_p}}}$$

Substituting

$$K'_n = K'_p, \qquad V_{tn} = 0.2V_{\text{DD}}, \qquad \text{and } V_{tp} = -0.2V_{\text{DD}}$$

then

$$V_{\text{out}} = 0.5V_{\text{DD}}$$

at the transition point.

The fraction is

$$\frac{V_{\text{DD}} - V_{\text{out}}}{V_{\text{DD}}} = 0.7 = 70\%$$

The point is made that for these conditions, the nMOS and pMOS transistors are individually in saturation for about 70% of the transition and jointly in saturation for about 40%. This has strong implications for failure analysis and design debug tools.■

CMOS inverter theory underlies failure analysis, test, and reliability electronics. For example, Chapter 3 described drain region light (photon) emission when a transistor was saturated. The high electric field of the drain depletion region accelerates channel charge, causing impact ionization and subsequent photon emission. Figure 4.5 shows where and when light is expected in the inverter. During normal logic transitions, light is emitted from the *n*MOS transistor in the saturated state region above line (a). The *p*MOS transistor emits light in the region below line (b). Both transistors emit light in the region between lines (a) and (b). The IBM PICA timing path analyzer with 15 ps timing resolution is based on these principles [7]. Failure analysts must know when defects force a transistor into an inadvertent saturated bias state, causing light emission [4]. Visible light from a defect location is an efficient failure analysis tool that, in addition to locating a defect region, says that a transistor is in its saturated state (or a diode is breaking down) with a weak gate and drain voltage. Failure analysts must describe defect properties consistent with all electrical clues.

Self-Exercise 4.2

A CMOS inverter has transistor parameters: $K_n\ (W/L)_n = 265\ \mu\text{A/V}^2$, $K_p\ (W/L)_p = 200\ \mu\text{A/V}^2$, $V_{tn} = 0.55$ V, $V_{tp} = 0.63$ V, and $V_{\text{DD}} = 2.5$ V. (a) For what fraction of the total output voltage swing will the *n*MOS transistor be in saturation. (b) Same question for the *p*MOS transistor. (c) What fraction will both be in saturation at the same time.

Self-Exercise 4.3

Estimate the voltage analog gain for the inverter whose transfer curve is in Figure 4.5

4.2.2 Dynamic Operation

The transfer curve of Figure 4.2 gives essential inverter information, but does not represent the circuit behavior during its rapid transition. The switching time for modern inverters can be a few tens of picoseconds, and parasitic capacitance of the transistors and the external wiring load alter the phase relation between V_{in} and V_{out}. An inverter and its transfer curve phase relation are drawn in Figure 4.7 for different input speed transitions, showing that for rapid transitions, the drain voltage lags input gate voltage.

The circuit model for the inverter dynamic analysis in Figure 4.7(a) shows two parasitic capacitors that are important during the transition. The input–output capacitor (called the coupling capacitor C_{coup}) comes from the gate–channel device capacitance that is strongly bias-dependent and the overlapping gate–drain capacitors from both the *n*MOS and *p*MOS devices. For high-speed transitions, the coupling capacitor tries to maintain its initial voltage difference between the input and the output ($-V_{\text{DD}}$ for a low–high input transition and V_{DD} for a high–low one). This temporarily drives the output voltage beyond V_{DD} (overshoot) for an input rising transition and below ground for a falling edge (undershoot). This induces noise in the supply (ground) node while increasing transition delay since the output voltage swing is higher than V_{DD}.

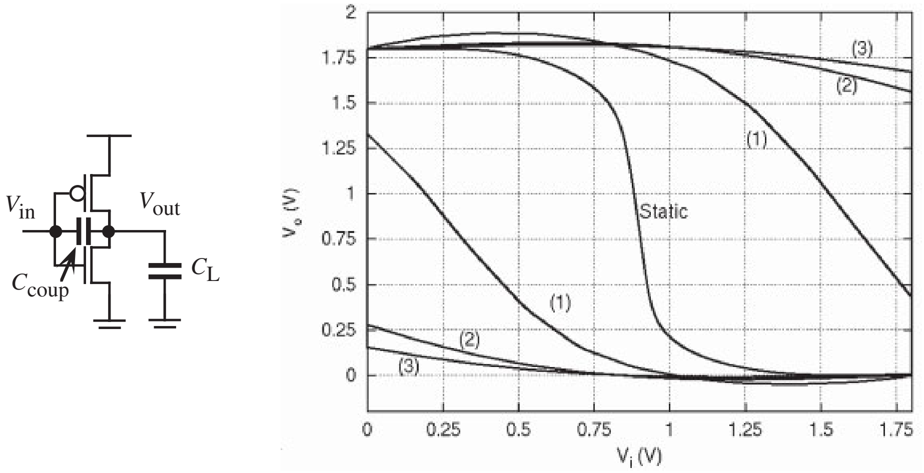


Figure 4.7. (a) Dynamic CMOS inverter circuit model, and (b) transfer curves for high-to-low output and low-to-high output for different input ramp speeds.

Curves (2) and (3) in Figure 4.7(b) correspond to very high speed input transistors and show that the output drain node remains at a relatively high voltage when the gate input has almost completed its transition. The same phenomena holds when the gate input switches rapidly from high to low. The drain remains in a low-voltage state until the input has almost completed its transition. The circuit parasitic capacitances cause this phase relation. The slow transition response of the static curve allows time for the drain nodes to exactly follow the input gate voltage in time. Curve (1) is an intermediate case for a medium speed input transition, in which the output is beyond $V_{DD}/2$ when the input reaches its final value, although it is far from the almost-static transfer curve.

This effect is more important at the beginning of the transition since for a rising (falling) input swing the *n*MOS (*p*MOS) device is off until $V_{in} = V_{tn}$ ($V_{in} = V_{DD} + V_{tp}$). During this period, the output voltage is disconnected from ground (supply) and there is no pull-down (pull-up) element to compensate for the charge injected from the input. This effect is shown in Figure 4.8 for the input and output voltage evolution for high–low and low–high input transitions.

Three parameters determine the duration and magnitude of overshoot (undershoot):

1. *The value of the input–output coupling capacitance.* The larger the capacitance, the more charge is transferred to the output during the input rising (falling) transition and the higher (lower) the output voltage overshoot (undershoot).
2. *The input transition time.* The charge injection from the input to the output through the coupling capacitance depends on the time derivative of the input voltage ($i = C\ dV/dt$). The shorter the input transition times, the higher the overshoot or undershoot.
3. *The width of the nMOS (pMOS) transistor.* Overshoot (undershoot) occurs at the beginning of the input transition, since the *n*MOS (*p*MOS) device is off and does not pull down (pull up) the output voltage. Once the input voltage goes beyond (below) V_{tn} ($V_{DD} + V_{tp}$) the *n*MOS (*p*MOS) device turns on and pulls the output voltage

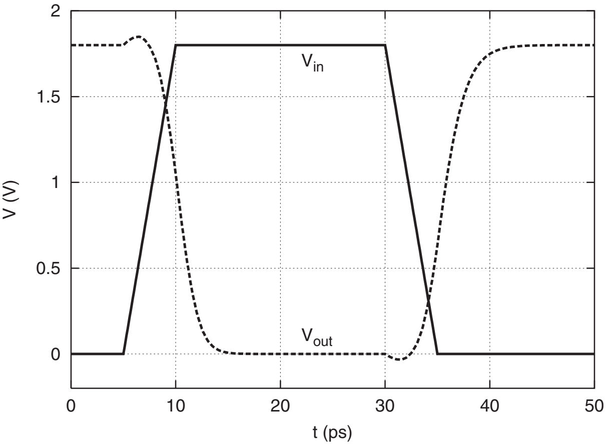


Figure 4.8. Time evolution of the input and output voltage of an inverter for an input high–low and low–high transitions. The output voltage exceeds the power and ground levels during the transition.

down (up). The larger the *W*/*L*, the larger its current drive and the smaller the time required to pull down (pull up) the output.

The inverter output capacitance is the sum of the drain diffusion capacitance, interconnect wiring, and the input capacitance of the load gates. The transition time is sensitive to the load capacitance since the output must be charged–discharged during the transitions. Figure 4.8 also illustrates the definition of logic gate propagation delay time T_{PD}, which is the time between the input and output waveform measured at the 50% amplitude points.

4.2.3 Inverter Speed Property

The dynamic relation between the input and the output voltages defines the different operating regions, but in the dynamic case the input voltage transition time does not uniquely determine the output voltage transition time. Sometimes, the input reaches its final value while the output voltage is still close to its initial value because the transistor that must sink or source current is too small compared to the output load. Different situations appear depending on transistor-to-load capacitance sizing.

Figure 4.9 plots the output voltage responses for different values of the output capacitance when the input low–high transition time is fixed. Curve V_{03} corresponds to a small output capacitor, and the output voltage is almost zero when the input reaches V_{DD}. This case is similar to the static transfer curve. In curve V_{02}, the output voltage is equal to $V_{\text{in}} - V_{tn}$ when the input voltage reaches V_{DD}, whereas in curve V_{01} the output is still above that value when the input gets to its steady state. In all cases, the output voltage initially goes beyond V_{DD} due to overshoot caused by the charge injected from the input through the coupling capacitor. During this period, there is current from the output node through the *p*MOS transistor back to the supply terminal. In none of the three cases does the output voltage start to decay until the input voltage goes beyond V_{tn} at time $t = t_n$. Fast and slow input ramps can distinguish the operating regions of the *n*MOS transistor when the input voltage reaches V_{DD}. In Curve V_{03}, the *n*MOS device is in its linear region when the input

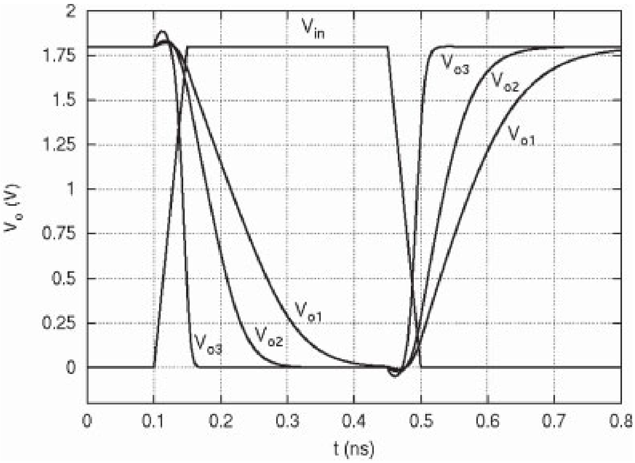


Figure 4.9. Different cases for a dynamic inverter transition.

transition is finished, whereas in curve V_{01} this transistor is still saturated when V_{tp} reaches V_{DD}.

Curve V_{01} is interesting because the *p*MOS transistor never passes a positive current. When overshoot ceases and the output voltage goes below V_{DD}, the input is below V_{tp} and the *p*MOS is off. This condition results in a slight reduction in power consumption.

An exact calculation of the propagation delay of an inverter requires a complex differential equation. We derive a simple formulation, assuming that the device is an ideal current source (i.e., the transistor is always in saturation).

The speed with which an inverter switches logic states depends on V_{tn}, V_{tp}, V_{DD}, W/L, temperature, and the coupling and load capacitances. A simple model shows these parameters that affect inverter rise and fall time. The current source I_0 in Figure 4.10(a) represents the MOS transistor in saturation since that bias state dominates the transition time and C_{L} is the load capacitance. The Shockley MOSFET model is

$$I_0 = I_{\mathrm{D}} = \frac{\mu_n C_{\mathrm{ox}}}{2} \frac{W}{L} (V_{\mathrm{GS}} - V_{tn})^2 \tag{4.4}$$

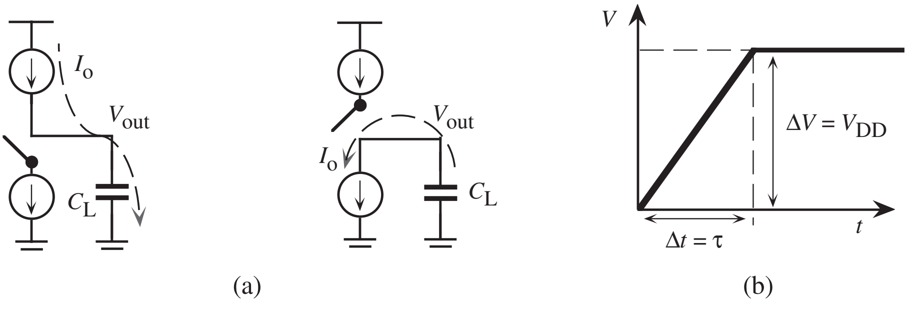


Figure 4.10. (a) Circuit model to estimate rise and fall delays in a CMOS inverter, and (b) voltage response of capacitor to constant current.

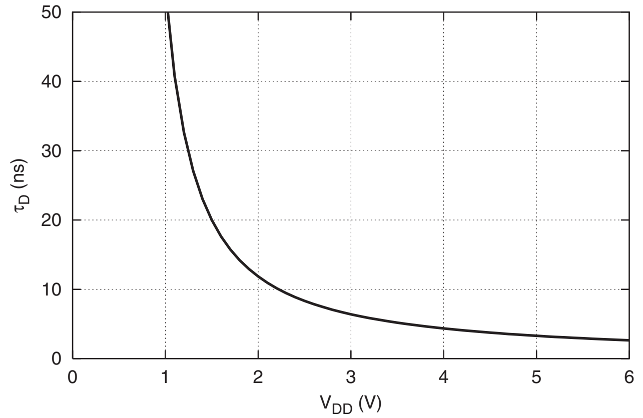


Figure 4.11. Delay versus supply voltage for the model in Equation (4.8).

The capacitor expression for current, voltage, and time from Chapter 1 is

$$i(t) = C_L \frac{dV(t)}{dt} \tag{4.5}$$

If $i(t) = I_0$ (a constant current) and we approximate $dv(t)/dt = \Delta V(t)/\Delta t$, then

$$\Delta V(t) = \frac{I_0}{C_L}\Delta t \tag{4.6}$$

If Δt is the delay time τ_D in Figure 4.10(b) for the signal to rise to $\Delta V(t) = V_{DD}$, then Equation (4.6) is rewritten as

$$\tau_D = \frac{C_L V_{DD}}{I_0} \tag{4.7}$$

Equation (4.7) adequately matches experimental data [2]. Substituting Equation (4.4) into Equation (4.7) with $V_{GS} = V_{DD}$ gives

$$\tau_D = C_L V_{DD} \frac{2L}{W\mu_n C_{ox}} \frac{1}{(V_{DD} - V_{tn})^2} \tag{4.8}$$

Equation (4.8) shows that time delay is geometrically related to the difference in V_{DD} and V_t. Figure 4.11 plots the time delay versus V_{DD} for $C_L = 20$ fF, $\mu_n C_{ox} = 150$ μA/V^2, $V_t = 0.5$ V, and $W/L = 2$. Time delay asymptotically approaches infinity as V_{DD} approaches V_t.

Self-Exercise 4.4

If a *p*MOS transistor has $\mu_n C_{ox} = 56$ μA/V^2, $V_{tp} = -0.6$ V, and $W/L = 6$, what is the expected additional time delay if the gate V_{DD} is reduced from a normal $V_{DD} = 2.5$ V to $V_{DD} = 1.8$ V with $C_L = 25$ fF?

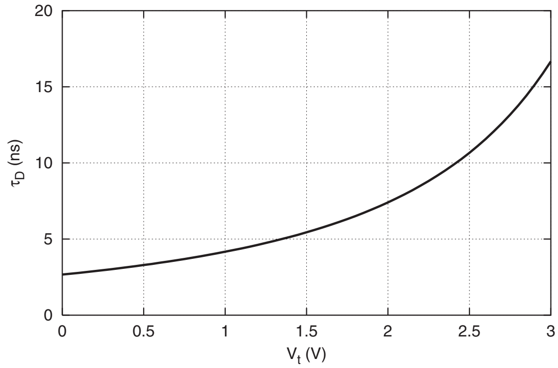


Figure 4.12. Delay versus threshold voltage for the model in Equation (4.8).

The result is similar if V_t varies for a fixed V_{DD}. Figure 4.12 is a plot similar to Figure 4.11, but with time delay plotted against V_t. These figures illustrate why deep submicron technologies strive for low V_t. This is a complicated trade-off for logic speed against the increase in off-state leakage current when V_t is lowered. Chapter 3 discussed these mechanisms.

■ EXAMPLE 4.3

It is given that $C_L = 10$ fF, $\mu_n C_{ox} = 118\ \mu A/V^2$, $W/L = 6$, and $V_{DD} = 2.3$ V. Initially, $V_t = 0.6$ V. If V_t is reduced to 0.2 V, what is the percent decrease in speed and what is the percent increase in I_{OFF} if subthreshold slope of I_D versus V_G is 83 mV/decade?

You can substitute the values into Equation (4.8) and take the ratio or divide Equation (4.8) by itself, substituting $V_t = 0.6$ V and $V_t = 0.2$ V. You get

$$\frac{\tau_D(V_t = 0.6\text{ V})}{\tau_D(V_t = 0.2\text{ V})} = \frac{(2.3 - 0.2)^2}{(2.3 - 0.6)^2} = 1.526$$

The threshold setting of $V_t = 0.6$ V slows the transistor by about 53%

The reduction is V_t affects the off-state leakage current by shifting the log (I_D) versus V_G curve to the right by (0.6 V – 0.2 V) = 400 mV (see Chapter 3). The increase in off-state leakage current is

$$\frac{400\text{ mV}}{83\text{ (mV/decade)}} = 4819\text{ decades} \Rightarrow 65.96 \times 10^3$$

The off-state leakage is increased over four orders of magnitude. A battery-operated circuit, such as a watch or pocket calculator, would choose the higher threshold voltage setting as speed is not a concern and power reduction is. ■

Equation (4.8) is valid for long-channel devices, since the current expression in Equation (4.4) has a quadratic dependence on the gate voltage. An analysis for deep-submicron devices substitutes any of their equations for the drain current into Equation (4.7). In this case, the delay is approximately [1]

$$\tau_{\mathrm{D}} = K\frac{C_{\mathrm{L}}V_{\mathrm{DD}}}{(V_{\mathrm{DD}} - V_{tn})} \tag{4.9}$$

K is a constant that depends on device size and technology parameters. The main difference is the inverse dependence on V_{DD} instead of the inverse square in Equation (4.9). Although deep-submicron delay appears larger than for long-channel transistors, the transistor thin oxide (T_{OX}) is much smaller, reflecting a larger value of K in Equation (4.8).

4.2.4 CMOS Inverter Power Consumption

The energy dissipated by an inverter has static and dynamic components. Dynamic dissipation is due to the charge–discharge of the gate output load capacitance (transient component) and to the short-circuit current from the supply to ground created during the transition. Static dissipation for long-channel transistors is due mainly to reverse bias drain–substrate (–well) *pn* junction leakage current from transistors in the off-state. The deep-submicron off-state current is mostly subthreshold leakage, and it dominates with technology scaling, since the threshold voltage V_{th} is reduced to maintain circuit performance [6]. Gate oxide tunneling current in ultrathin oxides also contributes significantly to IC off-state leakage. The dynamic power calculation requires computation of transient and short-circuit components.

Transient Component. The dynamic power (P_d) to charge and discharge a capacitor C_{L} for a period T of frequency f is

$$P_d = \frac{1}{T}\int_0^T i_{\mathrm{L}}(t)v_0(t)dt \tag{4.10}$$

In one period interval, the output voltage changes from 0 to V_{DD} and vice-versa. Equation (4.5) relates the current and voltage of the output capacitor, so Equation (4.10) is rewritten as

$$P_d = \frac{1}{T}\left[\int_0^{V_{\mathrm{DD}}} C_{\mathrm{L}}v_0 dv_0 + \int_{V_{\mathrm{DD}}}^0 C_{\mathrm{L}}(V_{\mathrm{DD}} - v_0)d(V_{\mathrm{DD}} - v_0)\right] \tag{4.11}$$

giving

$$P_d = \frac{C_{\mathrm{L}}V_{\mathrm{DD}}^2}{T} = C_{\mathrm{L}}V_{\mathrm{DD}}^2 f \tag{4.12}$$

Equation (4.12) shows that transient power can be lowered by reducing the output capacitance, the supply voltage, or the operating frequency. Since the power dependence on the supply voltage is quadratic, lowering V_{DD} is more efficient for reducing power dissipation than the other two parameters.

Short-Circuit Component. When the input is changing and its voltage is between V_{tn} and $V_{\mathrm{DD}} - |V_{tp}|$, then both transistors are simultaneously conducting creating a current path from V_{DD} to ground. This power component (P_{sc}) depends on device size, input transition time, and the output and coupling capacitors that consume about 10%–20% of the overall power. It was recently shown that the ratio of the short-circuit to the dynamic current remains constant for submicron technologies if the ratio $V_{\text{th}}/V_{\text{DD}}$ is constant. The exact computation is complex, so we present an approximation.

Consider a symmetric inverter (i.e., $K'_n = K'_p$, and $V_{tn} = -V_{tp}$) with no output load and an input voltage transition having equal rise and fall times. The time interval when both transistors simultaneously conduct is from t_1 to t_3 in Figure 4.13. During the interval $t_1 - t_2$, the short-circuit current increases from zero to its maximum value $I_{\max}$. Since the nMOS transistor is saturated during this period, its drain current is Equation (4.4)

$$I = K'_n(V_{\text{in}} - V_{tn})^2 \qquad \text{for } 0 < I < I_{\max}$$

where

$$K'_n = \frac{\mu_n C_{\text{ox}}}{2}\,\frac{W}{L} \tag{4.13}$$

Since the inverter was assumed to be symmetric with no load, the maximum current occurs at $V_{\text{in}} = V_{\text{DD}}/2$ and its shape is symmetric along the vertical axis at $t = t_2$. We compute a mean current by integrating from $t = 0$ to $t = T$ and dividing by the period T. There are four equal area current segments to integrate in Figure 4.13 over the whole period T:

$$I_{\text{mean}} = \frac{1}{T}\int_0^T I(t)dt = \frac{4}{T}\int_{t_1}^{t_2} K'_n(V_{\text{in}}(t) - V_t)^2 dt \tag{4.14}$$

If the input voltage is a linear ramp of duration τ,

$$V_{\text{in}} = \frac{V_{\text{DD}}}{\tau}t \tag{4.15}$$

t_1 and t_2 are given by

$$t_1 = \frac{V_t}{V_{\text{DD}}}\tau \qquad \text{and} \qquad t_2 = \frac{\tau}{2} \tag{4.16}$$

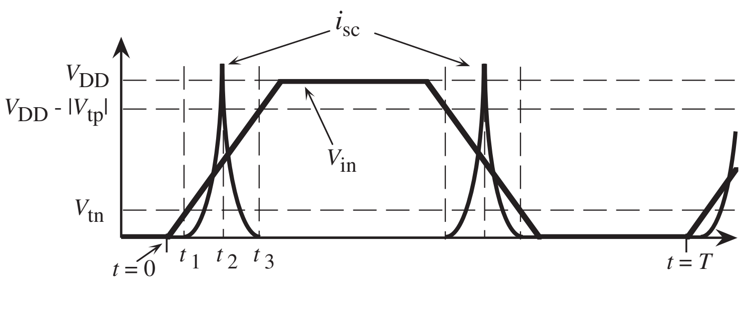


Figure 4.13. A simplified view of the short-circuit current contribution.

Substituting (4.15) and (4.16) into (4.14)

$$I_{\text{mean}} = K_n' \int_{\left(\frac{V_t}{V_{\text{DD}}}\right)\tau}^{\tau/2} \left(\frac{V_{\text{DD}}}{\tau} t - V_{\text{t}} \right)^2 dt \tag{4.17}$$

This integral is of the type $\int x dx$ with $x = (V_{\text{DD}}/\tau)t - V_t$, so the result is

$$I_{\text{mean}} = \frac{1}{6} \frac{K_n'}{V_{\text{DD}}} (V_{\text{DD}} - V_{\text{T}})^3 \frac{\tau}{T} \tag{4.18}$$

Finally, the power contribution is given by

$$P_{\text{sc}} = V_{\text{DD}} I_{\text{mean}}$$

Power Supply Scaling. The ratio of V_t to V_{DD} impacts several inverter properties. Equation (4.4) is repeated:

$$I_{\text{D}} = \frac{\mu_n \varepsilon}{2T_{\text{ox}}} \frac{W}{L} (V_{\text{GS}} - V_{tn})^2 \tag{4.19}$$

$V_{\text{GS}} = V_{\text{DD}}$ for logic circuits, so that Equation (4.19) becomes

$$I_{\text{D}} = \frac{\mu_n \varepsilon}{2T_{\text{ox}}} \frac{W}{L} (V_{\text{DD}} - V_{tn})^2 \tag{4.20}$$

When V_{DD} drops, several things happen:

- The voltage difference in the parentheses (the gate overdrive) is smaller, so the current drive is lower.
- When $V_{\text{DD}} < (V_{tn} - V_{tp}) \approx |2V_t|$ the transition still occurs, but only one transistor is on at a time. There is essentially no transient current spike. Figure 4.14 shows a transfer curve at $V_{\text{DD}} = 1$ V and $V_{\text{DD}} = 0.5$ V for transistors with thresholds on the order of 0.35 V. There is no current spike for the $V_{\text{DD}} = 0.5$ V measurement since $V_{\text{DD}} < 2 \times |V_t|$. V_{in} was swept from 0.5 to 0 V. The power reduction for this condition is large. Low-power, battery-operated products such as electronic watches and medical implants use this technique.
- The fraction of the V_{in} sweep in which both transistors are simultaneously on also drops, reducing the peak I_{DD} current.

There is a trade-off between power savings and delay increase when reducing the supply voltage. Technology scaling includes thinner gate oxides, forcing lower voltages to contain the gate oxide and drain–substrate electric fields to subcritical values. We present some approaches to power supply voltage scaling that focus on reliability, speed, or energy–delay.

Reliability Driven Supply Scaling. The most general voltage scaling trades off long-term reliability, operating speed, and energy. Circuit speed increases with higher V_{DD} but the higher device electric fields increase carrier velocity, creating more hot carriers. Hot carriers contribute to oxide degradation and can limit circuit lifetime. It is possible to de-

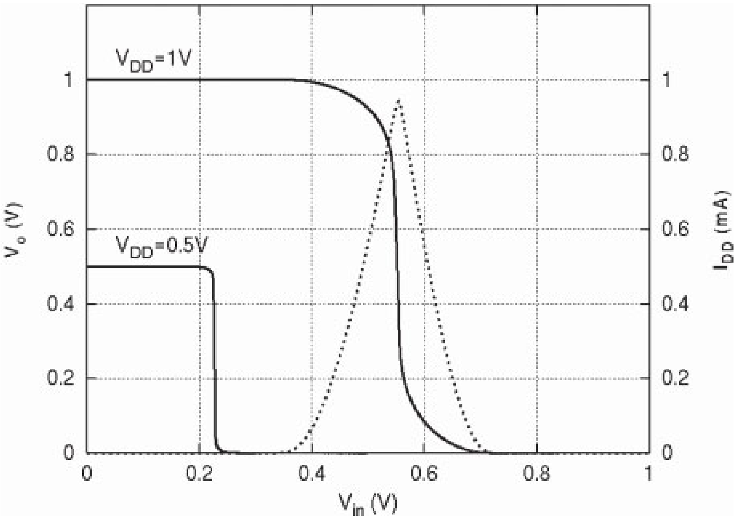


Figure 4.14. Inverter transfer curves at two V_{DD} values. Notice the absence of a short-circuit current spike for $V_{\mathrm{DD}} = 0.5$ V, where V_{in} was swept from 0.5 to 0 V.

velop a model in which circuit delay and hot carrier effects are included and from which an optimum power supply can be determined [2].

Technology Driven Supply Scaling. Transistor current drive in the saturated state for submicron technologies is not quadratic but linear. It is dominated by carrier velocity saturation (v_{max}), as explained in Chapter 3 through Equation (3.40). Since $v_{\mathrm{max}} = \mu\mathscr{E}_{\mathrm{crit}}$, Equation (3.40) is rewritten as

$$I = WC_{\mathrm{ox}}(V_{\mathrm{DD}} - V_t)v_{\mathrm{max}} \tag{4.21}$$

The first-order delay model of Equation (4.7) with this current expression gives a delay almost independent of the supply voltage, provided that carriers move at the velocity saturation v_{max} i.e., at high electric fields. Mathematically, this implies that $V_{\mathrm{DD}} - V_t \approx V_{\mathrm{DD}}$ and V_{DD} vanishes in the delay equation. A "technology" based criterion chooses the power supply voltage based on the desired speed–power performance for a given deep-submicon technology [5]. The relative independence of delay on supply voltage at high electric fields allows voltage reduction for a velocity-saturated device with little penalty in speed performance. This concept of operating above a certain voltage was formalized by Kakamu and Kingawa [5], where the concept of a "critical voltage" was developed.

Energy–Delay Minimum Supply Scaling. Another approach reduces the voltage supply by minimizing the energy–delay product [2]. For a fixed technology, there is a supply voltage that trades off the quadratic dependence of energy and the increased circuit delay.

4.2.5 Sizing and Inverter Buffers

A speed design problem exists when a signal passes through a series of logic gates. A fast charge–discharge of a load capacitor requires large W/L of the driving transistors. However, the large *W*/*L* is defeating since its large gate area increases the load capacitance for its own driving logic gate. Working backward, that would cause all preceding logic gates to have ever larger *W*/*L* ratios. A better solution exists.

A better approach for driving large loads at high speed uses successively larger channel widths in a cascade of inverters to sufficiently increase the current drive of the last stage. A circuit driving a large load is commonly known as a buffer, and a circuit designed with successive inverters is known as a tapered buffer (Figure 4.15). When the area of each stage increases by the same factor, the circuit is called a fixed-taper buffer, and if this ratio is not constant it is called a variable-taper buffer.

The fixed-taper buffer structure was proposed by Linholm in 1975 [8]. He used a simple capacitance model, making the output load of a stage proportional to the size of the input capacitance of the next stage, while the area of each inverter was proportional to the channel width of the transistors. The overall buffer delay was optimized by minimizing the delay of each stage. A better result is obtained if the system delay is considered instead of individual stages.

Let each succeeding stage in the buffer in Figure 4.15 have transistor widths larger than the previous one by a factor α. The first inverter is the smallest, with an input capacitance C_{in}, whereas the *i*th stage has an input capacitance given by

$$C_i = \alpha^{i-1} C_{\text{in}} \qquad i = 1, 2, \ldots, n \tag{4.22}$$

The number of stages n is computed from

$$C_{\text{L}} = \alpha^n C_n \tag{4.23}$$

then

$$\alpha^n = \frac{C_{\text{L}}}{C_{\text{in}}} \tag{4.24}$$

and

$$n = \frac{\ln (C_{\text{L}}/C_{\text{in}})}{\ln \alpha} \tag{4.25}$$

α is computed by optimizing the delay. Assuming that the delay of the first stage driving an identical one is τ_0, the delay of the *i*th stage is

$$t_{di} = \alpha \tau_0 \qquad i = 1, 2, \ldots, n \tag{4.26}$$

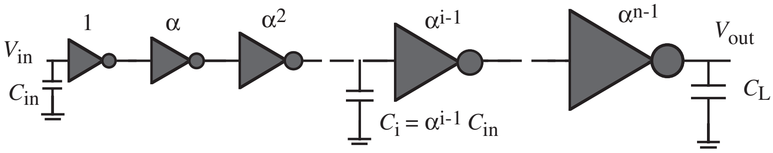


Figure 4.15. Tapered buffer structure.

The global delay of the n stages is

$$t_d = \sum_{i=1}^{n} t_{di} = n\alpha\tau_0 \tag{4.27}$$

giving

$$t_d = \ln\left(\frac{C_\mathrm{L}}{C_\mathrm{in}}\right)\frac{\alpha}{\ln \alpha}\tau_0 \tag{4.28}$$

Differentiating (4.28) with respect to α and equating to zero, the optimum α_opt is

$$\alpha_\mathrm{opt} = e \approx 2.7 \tag{4.29}$$

and the optimum number of stages n_opt is

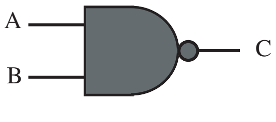
$$n_\mathrm{opt} = \ln\,(C_\mathrm{L}/C_\mathrm{in}) \tag{4.30}$$

Other buffer designs addressed power dissipation, circuit area, and system reliability, many of them leading to results different than shown here. Additionally, more accurate models for capacitance and delay estimation exist. Methods exist that consider interrelated issues of circuit speed, power dissipation, physical area, and system reliability [3]. The point of this section is to impress upon the reader that care must be taken in a design when a logic gate of one size and capacitance drives a logic gate of larger size and capacitance.

4.3 NAND GATES

CMOS technology implements negated functions. This means that the output signals, although controlled by many input lines, are inverted with respect to one or more controlling inputs. Simple examples are the inverter, the NAND gate, and the NOR gate.

A 2NAND gate symbol and truth table are shown in Figures 4.16(a) and (b). There are two properties to note. The first is that any logic 0 to the inputs of a NAND gate causes logic 1 output. The other property is more subtle, but vital to logic design and testing ICs. Certain input levels are called *noncontrolling states.* When A = 1 in rows 3 and 4, then the



(a)

A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

(b)

Figure 4.16. (a) 2NAND gate symbol and (b) truth table.

output C is the negation or complement of B (C = B′). The output C depends only on the value of B if A = 1. If B = 1, C is the complement of A. IC testing requires that signal information from specific nodes be read at an output pin without interference from the other input lines of that gate. The property is essential for passing specific logic values deep in the logic blocks to an observable circuit node such as an output pin.

When a NAND gate input is set to logic 1, then the effects of that signal node are neutralized with respect to signals on the other input lines. For example, if a fault were suspected on the input of B, then a test pattern would drive a signal to B and measure C, but ensure that node A = 1. The noncontrolling logic state for an AND gate is also logic 1.

■ EXAMPLE 4.4

If in Figure 4.17 you want to examine the signal at (a) node B, what should node I_1 be set to? (b) To examine node A, what should I_2 and I_3 be set to? (c) To examine node I_3, what should I_1 and I_2 be set to?

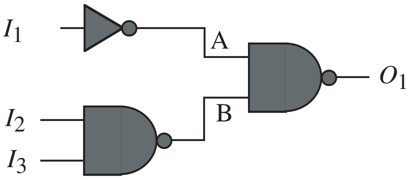


Figure 4.17.

(a) To pass a signal from node B to the output O_1 requires that node A is set at logic 1. Therefore I_1 must be logic 0.
(b) To pass a signal from node A to the output O_1 requires that node B is set at logic 1. Therefore I_2I_3 must be 00, 01, or 10.
(c) To pass a signal from node I_3 to the output O_1 requires that node I_2 be set at logic 1 and I_1 is set at logic 0. ■

Figure 4.18 shows the 2NAND gate transistor schematic. The electronic operation follows the truth table in Figure 4.16(b). A logic 0 on any input line turns off an *n*MOS pull-down transistor and closes the path from the output V_C to ground. A logic 0 ensures that a *p*MOS is turned on. Therefore, for any logic 0 on the inputs, the output is at logic 1, or $V_C = V_{DD}$. If both logic inputs are logic 1 ($V_A = V_B = V_{DD}$), then both *n*MOS transistors turn on, both *p*MOS transistors are off, and $V_C = 0$ V. The noncontrolling logic state for a NAND (and AND gate) input node is logic 1. AND gates can be made by adding an inverter to the output of a NAND gate.

The NAND gate has most of the inverter properties developed in Sections 4.1.2–4.1.5. If we set input B in Figure 4.18 to its noncontrolling state $V_B = V_{DD}$, then a voltage sweep at node A produces static and dynamic transistor curves similar to those measured for the inverter and shown in Figures 4.2 and 4.7.

Inverter speed and power properties apply to the NAND gate with minor exceptions. The goal of matching inverter rise and fall times led to design of $K'_n = K'_p$. This was done

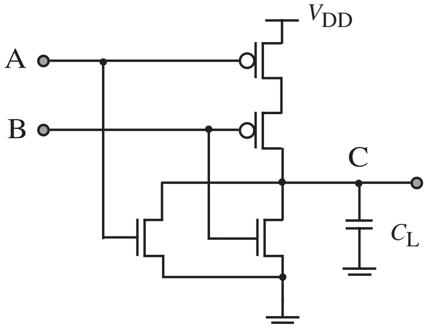
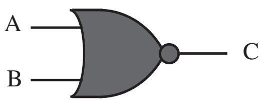


Figure 4.18. Transistor level structure of a static CMOS NAND gate.

by making $(W/L)_p > (W/L)_n$ by a factor of about 2–3.3 depending upon the technology. This ratio compensates for the lower *p*MOS transistor carrier mobility. The NAND gate has more input signal possibilities to deal with if equal rise and fall times are desired. The pull-up current drive strength depends upon the number of *p*MOS transistors that are activated. Two parallel *p*MOS transistors have twice the pull-up strength of a single *p*MOS. Also, when the pull-down path is activated, two series *n*MOS transistors have about half the current drive strength of just a single *n*MOS. Compromises are made with NAND gate *p*MOS transistors and, typically, the $(W/L)_p$ ratios are not as large as the *n*MOS transistor $(W/L)_n$ ratios of inverters.

4.4 NOR GATES

A NOR gate symbol and truth table are shown in Figures 4.19(a) and (b). There are again two properties to note. The first is that any logic 1 to the inputs of a NOR gate causes a logic 0 output. The noncontrolling states are different than the NAND and AND gates. When A = 0 in rows 1 and 2, then the output C is the complement of B (C = B′). A similar property is seen in rows 1 and 3, where the output C = A′ when B = 0. When an input to a NOR gate is set to logic 0, then the effects of that signal node are removed with respect to signals on the other input lines. The noncontrolling logic state for a NOR



A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

(a) (b)

Figure 4.19. (a) 2NOR gate symbol and (b) truth table.

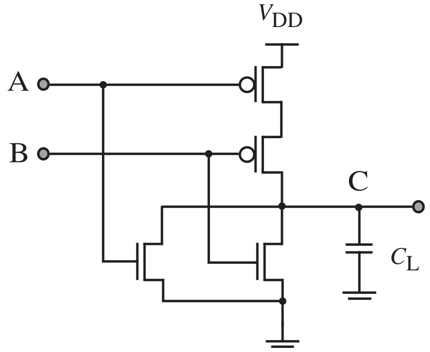


Figure 4.20. 2NOR transistor-level schematic.

gate (and OR gate) is logic 0. OR gates can be made by adding an inverter to the output of a NOR gate.

The NOR gate schematic in Figure 4.20 shows that any high logic input of V_{DD} turns on one of the *n*MOS transistors, forcing the output node C to 0 V. If both inputs are driven high with V_{DD}, then the pull-down strength is large since the *n*MOS transistor mobility is higher than the series *p*MOS transistors and the *n*MOS transistors are in parallel. Two *p*MOS transistors in series are potentially very slow, so the *W*/*L* adjustments for equal rise and fall times favor larger *p*MOS transistors. The NOR gate also has inverter static and dynamic properties. These are seen by setting one of the inputs to its noncontrolling logic state and measuring a transfer curve at the other terminal.

Self-Exercise 4.5

(a) Specify the input signals that allow node I_3 contents to be measured at O_1 (Figure 4.21). (b) Repeat for reading node I_2.

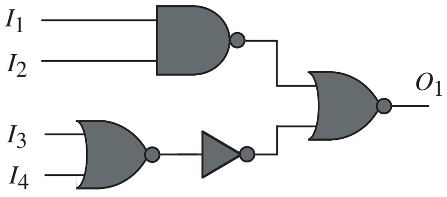


Figure 4.21.

Self-Exercise 4.6

(a) Specify the input signals that allow node I_5 contents to be measured at O_1 (Figure 4.22). (b) Repeat for reading node I_3. (c) Repeat for reading node I_1.

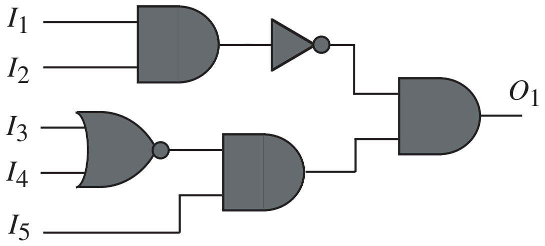


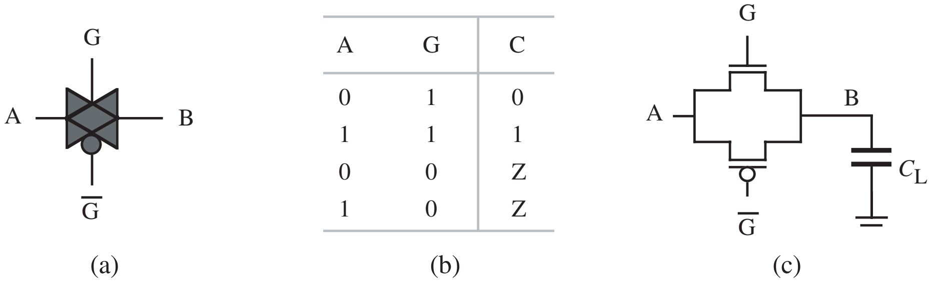
Figure 4.22.

4.5 CMOS TRANSMISSION GATES

A CMOS transmission gate, or T-gate, is a switch with many useful functions. Figure 4.23 shows an early T-gate design with symbol, truth table, and schematic. Signal transmission is controlled by the gating or control signal G in Figure 4.23. When G = 1, both transistors turn on and the signal passes to the output node B. When G = 0, no signal can pass since both transistors are off. The T-gate shown in Figure 4.23 can cause circuit problems in combinational logic since it has a high impedance state (also called floating, hi-Z, or tri-state) when the control signals turn off. Floating nodes allow voltage drift on transistor gates that can upset logic states. T-gates typically appear in CMOS flip-flop designs and to control tri-state levels in IC output buffers. In both applications, the floating node does not cause core logic instability.

Single transistors acting as T-gates are called pass transistors, but have a weakness. Assume that only the n-channel transistor in Figure 4.23(c) drives C_L and the pMOS transistor is removed. When the input A is high, the T-gate opens and charge passes through to C_L. When node B rises to a voltage one threshold drop below V_G, then the n-channel transistor turns off, so node B cannot rise above $V_G - V_{tn}$. An n-channel transistor passes a weak logic high. Similarly, a p-channel transistor will pass a weak logic zero that is one threshold drop above ground.

The cure is to put the n-channel and p-channel transistors in parallel. The nMOS transistor passes low logic levels with no voltage degradation, while the pMOS transistor passes the high logic levels with no V_t degradation. Ideally, a switch should have a constant resistance once turned on for any voltage that is transferred from node A to node B



A	G	C
0	1	0
1	1	1
0	0	Z
1	0	Z

(a) (b) (c)

Figure 4.23. **(a)** Transmission gate symbol, (b) truth table, and (c) transistor-level representation.

(Figure 4.23), but since MOS transistors are nonlinear elements, the on resistance is not constant.

4.6 SUMMARY

This chapter examined detailed electronic properties of the inverter. Much of design and failure analysis uses this information. The inverter properties also align with NAND, NOR, and other logic gates. Static and dynamic transfer curves explain much of the speed and power behavior of integrated circuits. The important design technique of tapered buffers is commonly used in design to match small logic gate drives to larger high-input capacitance load gates. It is essential to understand the operation of NAND, NOR, and transmission gates at the transistor schematic level. The next chapter expands these concepts to show how design of higher functions is achieved.

BIBLIOGRAPHY

1. A. Bellaouar and M. Elmasry, *Low-Power Digital VLSI Design; Circuits and Systems,* Kluwer Academic Publishers, 1995.
2. A. Chandrakasan and R. Brodersen, *Low Power Digital CMOS Design,* Kluwer Academic Publishers, 1995.
3. B. Cherkauer and E. Friedman, “A unified design methodology for CMOS tapered buffers,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 3,* 1, March 1995.
4. C. Hawkins, J. Soden, E. Cole, and E. Snyder, “The use of light emission in failure analysis of CMOS ICs,” in *International Symposium on Test and Failure Analysis (ISTFA),* pp. 55–67, 1990.
5. M. Kakamu and M. Kingawa, “Power supply voltage impact on circuit performance for half and lower submicrometer CMOS LSI,” *IEEE Transactions on Electron Devices, 37,* 8, 1902–1908, Aug. 1990.
6. A. Keshavarzi, K. Roy, and C. Hawkins, “Intrinsic leakage in low power deep submicron CMOS ICs,” in *IEEE International Test Conference (ITC),* pp. 146–155, Oct. 1997.
7. T. Tsang, J. Kash, and D. Vallett, “Time-resolved optical characterization of electrical activity in integrated circuits,” *Proceeding IEEE,* 1440–1459, Nov. 2000.
8. H. C. Lin and L. Lindholm, “An optimized output stage for MOS integrated circuits,” *IEEE Journal of Solid State Circuits, SC-10,* 2, 106–109, April 1975.

EXERCISES

4.1. A logic gate noise margin parameters are: $V_{IH} = 1.6$ V, $V_{IL} = 0.3$ V, $V_{OH} = 1.7$ V, and $V_{OL} = 0.2$ V.

(a) Calculate NM_H.

(b) Calculate NM_L.

(c) The input voltage is down to 1.7 V and a negative 50 mV noise spike appears. What happens to the circuit fidelity?

(d) The input voltage is down to 1.7 V and a negative 150 mV noise spike appears. What happens to the circuit fidelity?

4.2. Given an inverter A whose voltage transfer curve has a logic threshold at $V_{DD} = 3$ V of $V_{TL} = 1.5$ V, and a second inverter B with a logic threshold of $V_{TL} = 1.2$ V for the same V_{DD}:

(a) When the *n*-channel transistor goes into nonsaturation, is V_{in} more or less for B than for A?

(b) When the *p*-channel transistor in B goes into nonsaturation, is V_{in} more or less that that of the *p*-channel transistor in A when it goes into nonsaturation?

(c) What is the design difference between the transistors in inverters A and B?

4.3. Graphically determine the change in logic threshold of the CMOS inverter transfer curve in Figure 4.24 if the curve shifts 0.2 V to the right in the mid-region.

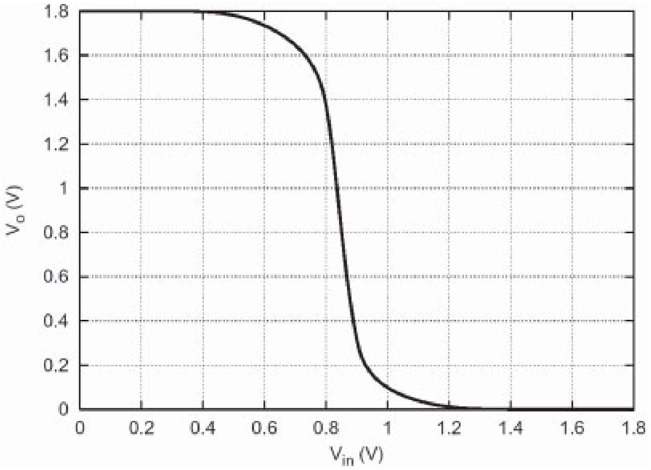


Figure 4.24.

4.4. Figure 4.7 shows how a CMOS inverter transfer curve changes for fast input transitions taking into account the load capacitance.

(a) Do these curves affect logic threshold?

(b) The hot carrier injection phenomenon that was described in Chapter 3 (Section 3.4.5) is aggravated by the electric field of the drain depletion region. Will hot carrier injection damage differ on the rising and falling input signal?

4.5. Calculate the power dissipated by a cardiac pacemaker circuit if $f_{clk} = 32.6$ kHz, $V_{DD} = 1.5$ V, C_L (per gate) = 300 fF, and the number of logic gates = 10 k.

4.6. Figure 4.9 shows that logic gate spikes can go above V_{DD} and below 0 V (GND). Why does this happen? Where does the charge come from, and where does it go?

4.7. Use the transition time delay model of Figure 4.10 if $C_L = 30$ fF, $V_{DD} = 1.5$ V, $K_p = 200\ \mu A/V^2$, $V_{tp} = -0.35$ V, $V_{tn} = 0.35$ V, and $K_n = 125\ \mu A/V^2$. What is the difference between rise and fall time of the transition if defined between 0 V and 1.5 V?

4.8. Repeat previous problem using the short channel model of Equation (4.9).

4.9. Figure 4.14 show CMOS transfer curves as V_{DD} goes above and below $V_{tn} + |V_{tp}|$. If V_{DD} goes below a single threshold voltage (i.e., V_{tn}), then the inverter still exhibits a valid transfer curve, although the curve is not as sharp. From your knowledge of transistors in Chapter 3, how is operation for $V_{\mathrm{DD}} < V_t$ possible?

4.10. An output buffer has an input capacitance of 95 fF and a load capacitance of 100 pF. How many inverters are required in a fixed-taper design to minimize the propagation delay?

4.11. Figure 4.25 shows a single transistor transmission gate. At $t = 0$ the gate voltage moves to $V_{\mathrm{G}} = 2$ V. The load capacitance C_{L} is initially at zero volts. As the node capacitance (the source) charges V_{GS} becomes smaller until it is less than V_{tn}. The transistor cuts off and the node is less than the input drain voltage by $V_{\mathrm{S}} = V_{\mathrm{D}} - V_{tn}$. If the body coefficient is $\gamma = 0.2\mathrm{V}^{1/2}$ and $V_{t0} = 0.4$ V, calculate V_{GS} and V_{S} when the capacitor fully charges before cutoff.

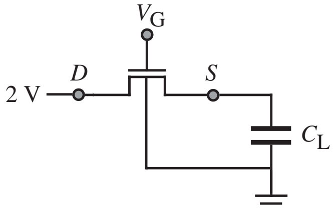


Figure 4.25.

4.12. Given an inverter with $V_{tn} = 0.4$ V and $V_{tp} = -0.4$ V, calculate the peak current during the transition if $W/L = 3$, $K_p = 50\ \mu\mathrm{A/V}^2$, and $K_n = 125\ \mu\mathrm{A/V}^2$.

4.13. Given the logic circuit in Figure 4.26:

(a) What signal values must the input nodes be set to in order to read the contents of node C at output H?

(b) What signal values must the input nodes be set to in order to read the contents of node G at output H?

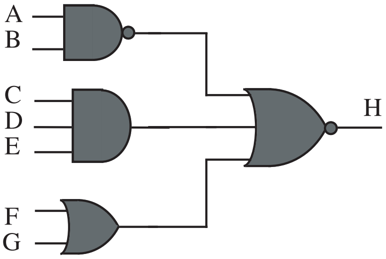


Figure 4.26.

CHAPTER 5

CMOS BASIC CIRCUITS

The previous chapter covered CMOS basic gate construction, emphasizing switching delay and power consumption characteristics. We now look at CMOS logic design styles, including static, dynamic, and pass-transistor logic. Input–output (I/O) circuitry and its protection problems are also discussed.

5.1 COMBINATIONAL LOGIC

Several design options exist for CMOS combinational gates. One reliable, lower-power design style uses complementary static gates, whereas high-performance circuits may use dynamic logic styles more suitable for high speed. Dynamic logic is more sensitive to noise and requires synchronization of signals (with a clock), even for combinational logic. Another logic design style uses pass-transistor or pass-gate elements as basic switches when fewer transistors are needed to implement a function. We want to understand these combinational logic design styles and their trade-offs.

5.1.1 CMOS Static Logic

Static, fully complementary CMOS gate designs using inverter, NAND, and NOR gates can build more complex functions. These CMOS gates have good noise margins and low static power dissipation at the cost of more transistors when compared with other CMOS logic designs. CMOS static complementary gates have two transistor nets (*n*MOS and *p*MOS) whose topologies are related. The *p*MOS transistor net is connected between the power supply and the logic gate output, whereas the *n*MOS transistor topology is connected between the output and ground (Figure 5.1). We saw this organization with the NAND and NOR gates, but we point out this topology to lead to a general technique to convert Boolean algebra statements to CMOS electronic circuits.

ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

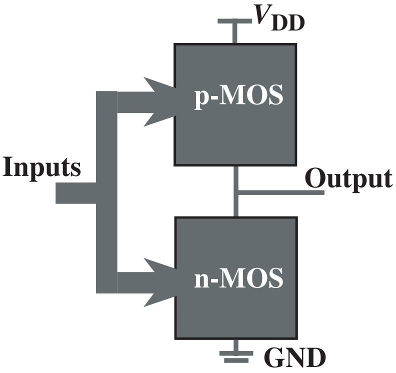


Figure 5.1. Standard configuration of a CMOS complementary gate.

The transistor network is related to the Boolean function with a straightforward design procedure:

1. Derive the *n*MOS transistor topology with the following rules:
   - Product terms in the Boolean function are implemented with series-connected *n*MOS transistors.
   - Sum terms are mapped to *n*MOS transistors connected in parallel.
2. The *p*MOS transistor network has a dual or complementary topology with respect to the *n*MOS net. This means that serial transistors in the *n*MOS net convert to parallel transistors in the *p*MOS net, and parallel connections within the *n*MOS block are translated to serial connections in the *p*MOS block.
3. Add an inverter to the output to complete the function if needed. Some functions are inherently negated, such as NAND and NOR gates, and do not need an inverter at the output state. An inverter added to a NAND or NOR function produces the AND and OR function. The examples below require an inverter to fulfill the function.

This procedure is illustrated with three examples.

■ EXAMPLE 5.1

Design a complementary static CMOS 2NAND gate at the transistor level.

The Boolean function is simply A · B, therefore the *n*MOS net consists of two series-connected transistors, whereas the *p*MOS net will use the complementary topology, i.e., two transistors in parallel. The transistor structure was shown in Figure 4.18. ■

■ EXAMPLE 5.2

Design a complementary static CMOS XOR gate at the transistor level.

The XOR gate Boolean expression F has four literals and is

$$F = x \oplus y = \overline{x}y + x\overline{y}$$

F is the sum of two product terms. The design steps are:

1. Derive the *n*MOS transistor topology with four transistors, one per literal in the Boolean expression. The transistors driven by $\bar{x}$ and y are connected in series, as well as the devices driven by x and $\bar{y}$. These transistor groups are connected in parallel, since they are additive in the Boolean function. The signals and their complements are generated using inverters (not shown). The *n*MOS transistor net is shown in Figure 5.2.

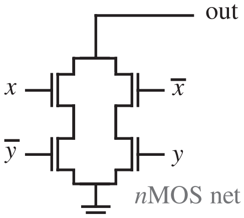


Figure 5.2.

2. Implement the *p*MOS net as a dual topology to the *n*MOS net. The *p*MOS transistors driven by $\bar{x}$ and y are connected in parallel, as are the devices driven by x and $\bar{y}$ (Figure 5.3). These transistor groups are connected in series, since they are parallel connected in the *n*MOS net. The *out* node now implements $\bar{F}$.

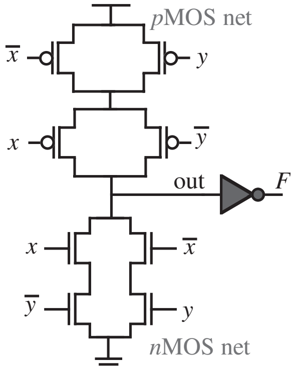


Figure 5.3.

3. Finally add an inverter to obtain the function F, so that F = $\overline{out}$

Steps 1–3 show that any Boolean function, regardless of its complexity, can be implemented with a CMOS complementary structure and an inverter. A more complicated example is developed below. ■

■ EXAMPLE 5.3

Design the *n*MOS transistor net for a Boolean function $F = x + \{\bar{y} \cdot [z + (t \cdot \bar{w})]\}$.

We design this gate with a top-down approach. The *n*MOS transistor network is connected between the output and ground terminals, i.e., the lower box in Figure 5.1. The higher-level function F is a sum of two terms

$$F = x + \{operation\ A\}$$

where *operation A* stands for the logic within the brackets of F. The transistor version of this sum is shown in Figure 5.4.

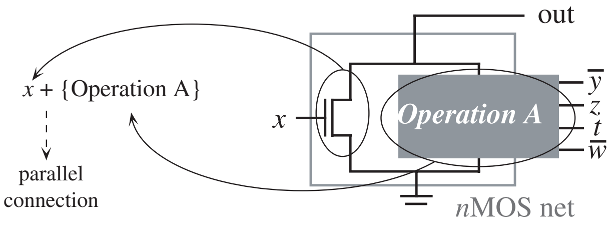


Figure 5.4.

Now we design the transistor topology that implements the block "*operation A,*" whose higher level operation is an AND, i.e.:

$$operation\ A = \overline{y} \cdot \{operation\ B\}$$

Hence, the design topology is a transistor controlled by input $\overline{y}$ in series with a third box that will implement *operation B,* as shown in Figure 5.5.

We then design the topology of box B. This is a transistor controlled by input z, in parallel with two transistors connected in series; one controlled by input t, and the other by input $\overline{w}$. The complete nMOS network is shown in Figure 5.6.

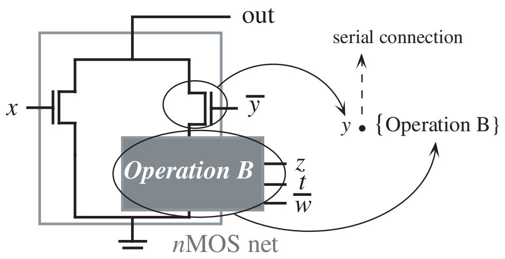


Figure 5.5.

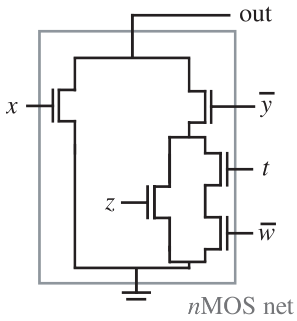


Figure 5.6.

Once the nMOS block is designed, we build the pMOS block with a dual topological structure and then connect an inverter to its output, as shown in Figure 5.7. ■

Self-Exercise 5.1

Design the transistor level schematic of function $F = (x + y)[z + (wt)(\overline{z} + x)]$.

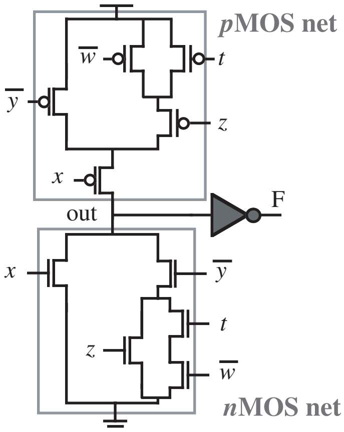


Figure 5.7.

5.1.2 Tri-State Gates

Many logic gates require a tri-state output—high, low, and high-impedance states. The high-impedance state is also called the high-Z state, and is useful when connecting many gate outputs to a single line, such as a data bus or address line. A potential conflict would exist if more than one gate output tried to simultaneously control the bus line. A controllable high-impedance-state circuit solves this problem.

There are two ways to provide high impedance to CMOS gates. One way provides tri-state output to a CMOS gate by connecting a transmission gate at its output (Figure 5.8). The control signal C sets the transmission gate conducting state that passes the non-tri-stated inverter output *out′* to the tri-stated gate output *out*. When the transmission gate is off ($C = 0$), then its gate output is in the high-impedance or floating state. When $C = 1$, the transmission gate is on and the output is driven by the inverter.

A transmission gate connected to the output provides tri-state capability, but also consumes unnecessary power. The design of Figure 5.8 contributes to dynamic power each time that the input and output (*out′*) are switched, even when the gate is disabled in the tri-state mode. Parasitic capacitors are charged and discharged. Since the logic activity at the input does not contribute to the logic result while the output is in tri-state, the power consumption related to this switching is wasted.

This can be avoided by putting a transmission gate "inside" the inverter (Figure 5.9).

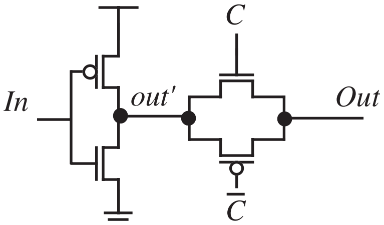


Figure 5.8. Inverter with a transmission gate to provide tri-state output.

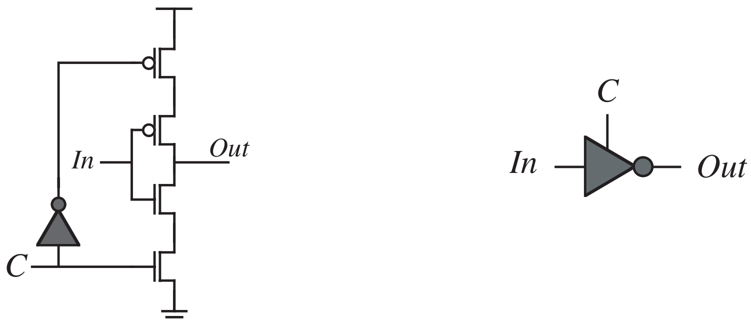


Figure 5.9. Schematic and symbol. The transmission gate "inside" the inverter provides tri-state output.

The *p*MOS and *n*MOS transistors of the transmission gate are in series within the conducting path between the power and ground rails and the inverter transistors. When the gate is in the tri-state mode, the inner transistor source nodes float, and the output is isolated from supply and ground. The activity at the inverter output signal node does not consume power as long as the gate is in the high-Z state ($C = 0$).

A tri-state capability adds delay independent of the configuration, due to the extra resistance and capacitance of the transistors driven by the tri-state control signal.

5.1.3 Pass Transistor Logic

There are many pass transistor (pass gate) logic subfamilies [3], and we will describe a few. Pass transistor logic uses transistors as switches to carry logic signals from node to node, instead of connecting output nodes directly to V_{DD} or ground. If a single transistor is a switch between two nodes, then a voltage degradation equal to V_t for the high or low logic level is obtained, depending on the *n*MOS or *p*MOS transistor type (Chapter 4). CMOS transmission gates avoid these weak logic voltages of single-pass transistors at the cost of an additional transistor per transmission gate.

Advantages are the low number of transistors and the reduction in associated interconnects. The drawbacks are the limited driving capability of these gates and the decreasing signal strength when cascading gates. These gates do not restore levels since their outputs are driven from the inputs, and not from V_{DD} or ground [6].

A typical CMOS design is the gate-level multiplexer (MUX) shown in Figure 5.10 for a 2-to-1 MUX. A MUX selects one from a set of logic inputs to connect with the output. In Figure 5.10, the logic signal c selects either *a* or *b* to activate the output (*out*). Figure 5.10(b) shows a MUX design with transmission gates. The complementary CMOS gates (Figure 5.10(a)) require 14 transistors (four transistors for each NAND and two transistors to complement the control signal), whereas the transmission gate design requires only six devices (more than 50% reduction). Each transmission gate has two transistors plus two more to invert the control signal.

Another pass gate design example is the XOR gate that produces a logic one output when only one of the inputs is logic high. If both inputs are logic one or logic zero, then the output is zero. Figure 5.11 shows an 8-transistor XOR gate having a tri-state buffer and transmission gate with their outputs connected. Both gates are controlled by the same input through a complementary inverter (A-input in this case).

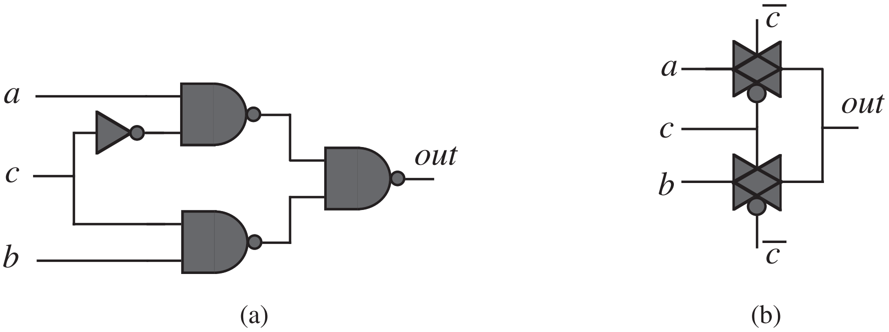


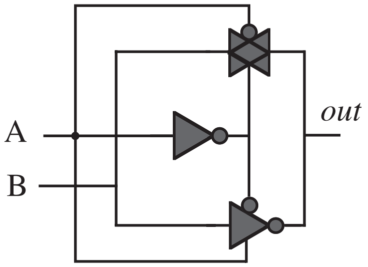
Figure 5.10. (a) Standard 2-to-1 MUX design. (b) Transmission (pass) gate-based version.

The XOR gate of Figure 5.11 is not a standard complementary static CMOS design since there is no *n*MOS transistor network between the output and ground, nor is there a *p*MOS transistor net between the output and the power rail. The XOR standard CMOS design built in Example 5.2 requires fourteen transistors, whereas the design in Figure 5.11 requires only eight.

5.1.4 Dynamic CMOS Logic

Previous sections showed conventional static CMOS circuit design techniques and designs based on tri-state gates and pass transistors. These designs are static, since they do not require a clock signal for combinational circuits. So, if circuit inputs are stopped (elapsed), then the circuits retain their output state (all circuit nodes remain at their valid quiescent logic values) as long as power is maintained. Dynamic CMOS logic families do not have this property, but do have the following advantages:

- They use fewer transistors and, therefore, less area.
- Fewer transistors result in smaller input capacitance, presenting a smaller load to previous gates, and therefore faster switching speed.
- Gates are designed and transistors sized for fast switching characteristics. High-performance circuits use these families.



A	B	out
0	0	0
0	1	1
1	0	1
1	1	0

Figure 5.11. 8-transistor XOR gate and truth table.

- The logic transition voltages are smaller than in static circuits, requiring less time to switch between logic levels.

The disadvantages of dynamic CMOS circuits are

- Each gate needs a clock signal that must be routed through the whole circuit. This requires precise timing control.
- Clock circuitry runs continuously, drawing significant power.
- The circuit loses its state if the clock stops.
- Dynamic circuits are more sensitive to noise.
- Clock and data must be carefully synchronized to avoid erroneous states.

Dynamic CMOS Logic Basic Structure. A dynamic CMOS gate implements the logic with a block of transistors (usually *n*MOS). The output node is connected to ground through an *n*MOS transistor block and a single *n*MOS evaluation transistor. The output node is connected to the power supply through one precharge *p*MOS transistor (Figure 5.12). A global clock drives the precharge and evaluation transistors. The gate has two phases: evaluation and precharge. During precharge, the global clock goes low, turning the *p*MOS transistor on and the evaluation *n*MOS off. The gate output goes high (it is precharged) while the block of *n*MOS transistors float.

In the evaluation phase, the clock is driven high, turning the *p*MOS device off and the evaluation *n*MOS on. The input signals determine if there is a low or high impedance path from the output to ground since the global clock turns on the *n*MOS evaluation transistor. This design eliminates the speed degradation and power wasted by the short-circuit current of the *n*- and *p*-channel transistors during the transition of static complementary designs. If the logic state determined by the inputs is a logic one (V_{DD}) then the rise time is zero. The precharge and evaluation transistors are designed to never conduct simultaneously.

Dynamic circuits with an *n*-input gate use only $n + 2$ transistors instead of the $2n$ devices required for the complementary CMOS static gates. Dynamic CMOS gates have a drawback. If the global clock in Figure 5.12 is set high, then the output node could be in high-Z state with no electrical path to V_{DD} or ground. This exposes the node to noise fluctuations and charge sharing within the logic block, thus degrading its voltage. Also, the output load capacitor will slowly discharge due to transistor off-state leakage currents and lose its logic value. This limits the low-frequency operation of the circuit. The gate inputs

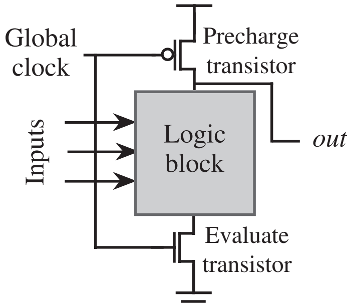


Figure 5.12. Basic structure of a dynamic CMOS gate.

can only change during precharge, since charge redistribution from the output capacitor to internal nodes of the *n*MOS logic block may drop the output voltage when it has a logic high.

Finally, dynamic gate cascading is challenging since differences in delay between logic gates may cause a slow gate to feed an erroneous logic high (not yet evaluated to zero because of the delay) to the next gate. This would cause the output of the second gate to be erroneously zero. Different clocking strategies can avoid this, as shown next.

Domino CMOS Logic. Domino CMOS was proposed in 1982 by Krambeck, et al., [4]. It has the same structure as dynamic logic gates, but adds a static buffering CMOS inverter to its output. In some cases, there is also a weak feedback transistor to latch the internal floating node high when the output is low (Figure 5.13). This logic is the most common form of dynamic gates, achieving a 20%–50% performance increase over static logic [3].

When the *n*MOS logic block discharges the *out′* node during evaluation (Figure 5.13), the inverter output out goes high, turning off the feedback *p*MOS. When *out′* is evaluated high (high impedance in the dynamic gate), then the inverter output goes low, turning on the feedback *p*MOS device and providing a low impedance path to V_{DD}. This prevents the *out′* node from floating, making it less sensitive to node voltage drift, noise, and current leakage.

Domino CMOS allows logic gate cascading since all inputs are set to zero during precharge, avoiding erroneous evaluation from different delays. This logic allows static operation from the feedback latching *p*MOS, but logic evaluation still needs two subcycles: precharge and evaluation. Domino logic uses only noninverting gates, making it an incomplete logic family. To achieve inverted logic, a separate inverting path running in parallel with the noninverted one must be designed.

Multiple output domino logic (MODL) is an extension of domino logic, taking internal nodes of the logic block as signal outputs, thus saving area, power, and performance. Compound domino logic is another design that limits the length of the evaluation logic to prevent charge sharing, and adds other complex gates as buffer elements (NAND, NOR, etc., instead of inverters) to obtain more area compaction. Self-resetting domino logic (SRCMOS) has each gate detect its own operating clock, thus reducing clock overhead and providing high performance. These and other dynamic logic designs are found in [3].

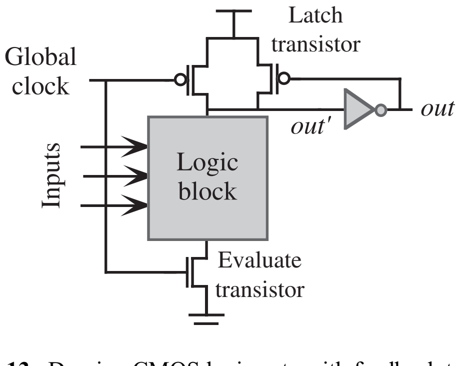


Figure 5.13. Domino CMOS logic gate with feedback transistor

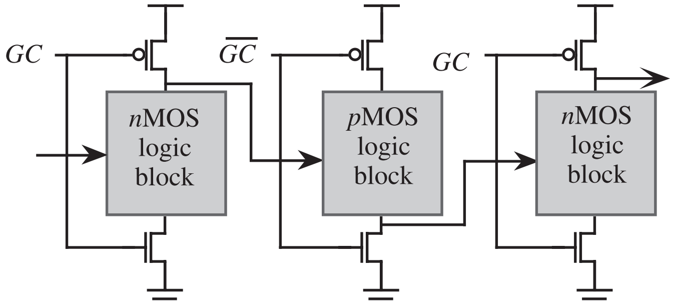


Figure 5.14. NORA CMOS cascaded gates.

NORA CMOS Logic. This design alternative to domino CMOS logic eliminates the output buffer without causing race problems between clock and data that arise when cascading dynamic gates. NORA CMOS (No-Race CMOS) avoids these race problems by cascading alternate *n*MOS and *p*MOS blocks for logic evaluation. The cost is routing two complemented clock signals. The cascaded NORA gate structure is shown in Figure 5.14. When the global clock (*GC*) is low ($\overline{GC}$ high), the *n*MOS logic block output nodes are precharged high, while outputs of gates with *p*MOS logic blocks are precharged low. When the clock changes, gates are in the evaluate state.

Other CMOS Logic Families. Dynamic circuits have a clock distribution problem, since all gates must be functionality synchronized. Self-timed circuits are an alternative to dynamic high-performance circuits, solving the clock distribution by not requiring a global clock. This simplifies clock routing and minimizes clock skew problems related to clock distribution. The global clock is replaced by a specific self-timed communication protocol between circuit blocks in a request–acknowledge scheme. Although more robust than dynamic circuits, self-timed logic requires a higher design effort than other families. These gates implement self-timing (i.e., derivation of a completion signal) by using a differential cascode voltage switch logic (known as DCVS) based on an extension of the domino logic.

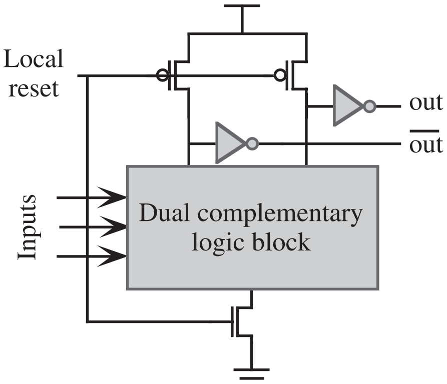


Figure 5.15. Basic DCVS logic gate.

The DCVS logic family (Figure 5.15) uses two complementary logic blocks, each similar to the domino structure. The gate inputs must be in the true and complementary form. Since output true and output negated are available, they can activate a completion signal when the output is evaluated. Since the gate itself signals when the output is available, DCVS can operate at the maximum speed of the technology, providing high-performance asynchronous circuits. The major drawbacks are design complexity and increased size.

5.2 SEQUENTIAL LOGIC

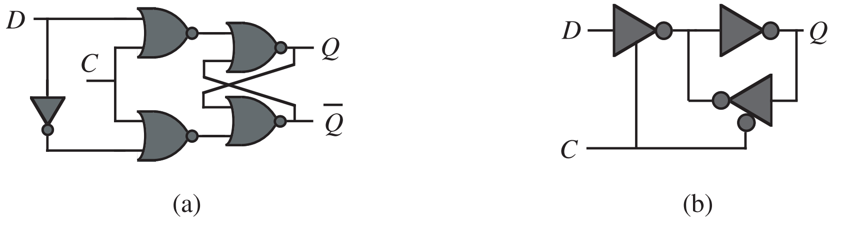
Nonstandard complementary CMOS designs are widely used in sequential logic to achieve compaction, high speed, and data storage. Latches, flip-flops, and registers are basic to many IC circuit designs. We present some of the better known static and dynamic memories.

5.2.1 Register Design

Registers are made with flip-flops that are in turn made with latches. Latches are memory elements whose transparent/memory states depend on the logic value (level) of a control signal. Flip-flops are constructed using latches to obtain a memory element that is transparent during the transition (edge) of the control signal for a better command of the time instant at which data are captured. We describe the basic latch, and then build the higher blocks.

CMOS Latch with Tri-State Inverters. Figure 5.16 shows the gate level and tri-state inverter design of a compact CMOS latch with two tri-state inverters and one regular inverter (Figure 5.16(b)). When *clk* = $C = 1$, the outputs of the first set of 2NOR gates are logic zero. This is the noncontrolling logic state feeding the D signal to the two output 2NOR gates. Therefore, the Q and $\overline{Q}$ signals feeding the inputs of the two output 2NOR gates set a stable logic condition. If $Q = 1$, then the bottom output 2NOR gate is driven to $\overline{Q} = 0$. The $\overline{Q}$ signal feeds a logic zero to the upper 2NOR gate, setting and holding $Q = 1$ (and $\overline{Q} = 0$). The latch holds its logic state indefinitely unless input signals change or the power is lost. When $C = 0$ (noncontrolling logic state to the input 2NOR gates), the Q outputs respond to the data input signal D. This is an example of a circuit that loads data on the low or negative portion of the clock signal.

In Figure 5.16(b), a level-sensitive clock controls the tri-state input of both inverters



(a)

(b)

Figure 5.16. (a) Basic gate-level CMOS latch design. (b) Tri-state inverter-level schematic.

such that when one is in tri-state, the other one is not. When the output of the first tri-state inverter stage is active (C = high), the feedback inverter is in tri-state (off), and the latch output is transparent. When C is low, the output of the first inverter floats, and the feedback tri-state inverter latches the value maintaining a feedback recovery configuration, holding the value. When $C = 0$, the latch is in its memory state. This is an example of a circuit that loads data on the positive portion of the clock.

> ***Self-Exercise 5.2***
>
> Compare the number of transistors in the latch of Figure 5.16(a) with a D latch designed with tri-state inverters [Figure 5.16(b).]

CMOS Latch with Transmission Gates. Another transmission gate latch design further reduces transistor count. The circuit in Figure 5.17 uses two transistors less than that shown in Figure 5.16(b).

CMOS Flip-Flop with Tri-State Inverters. Flip-flops are edge-sensitive memory elements using latches in a "master–slave" (MS) configuration. This edge-sensitive circuit changes logic state not on the level of the clock, but on the leading or falling edge of the clock. This eliminates the transparency properties of the latch since the output signal never sees a direct path to the input. The output is sensitive to change on one of the clock edges, and insensitive to the clock level.

The clock drives the master latch with the slave latch clock signals inverted. The master and slave are coupled through a transmission gate. The master latch configuration captures data at one clock level (high or low), and the slave captures data on the opposite value. The transmission gate between the master and slave latches controls the timing for capture of output data Q.

Figure 5.18 shows a flip-flop design with unequal master and slave cells. The master cell (left portion of the circuit) is the latch design described earlier, and is connected to the slave (right portion of the circuit) through a transmission gate. When the clock is low, the master and slave are isolated, with the master active and the slave in memory. The action of the master tri-state circuit generates a logic value at the master inverter output that equals the input data D. When the clock goes high, the transmission gate connecting the master and slave opens, and data are transferred. Data are read directly to the Q output on the rising edge of the clock. The data could be transferred on the clock falling edge if the coupling transmission gate (and the other clocked signals) reversed their clock signal polarities. The MS design differs from a latch, since the MS output Q sees very little of the

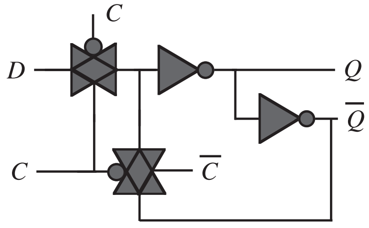


Figure 5.17. Alternate design of a latch cell with transmission gates.

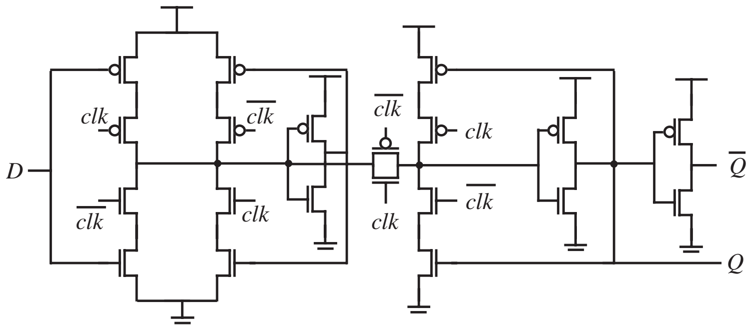


Figure 5.18. CMOS design of a flip-flop combining tri-state inverters and transmission gate design. The slave cell (right side) is only half of the master latch design to further reduce the number of transistors. Data are loaded into the first master latch on the negative clock edge, and data are read by the output Q on the rising clock edge. Then data are stored when the clock returns to logic zero.

input signal D directly. There is a small transient period when all transmission gates are in switching conduction states, and an electrical path may exist throughout the MS flip-flop. However, modern transition times are in the tens of picoseconds, and small clock timing skews make the overlap time very short.

5.2.2 Semiconductor Memories (RAMs)

Memories are a high-volume product in the IC market. The original phrase "random access memory" (RAM) refers to a memory in which all data have equal access procedures. There is no shifting of registers to capture a data bit. Test and reliability engineers also use memories to screen and verify emerging technologies, since they are relatively easy to test and failure analyze for process debugging. Their regularity and high density make them

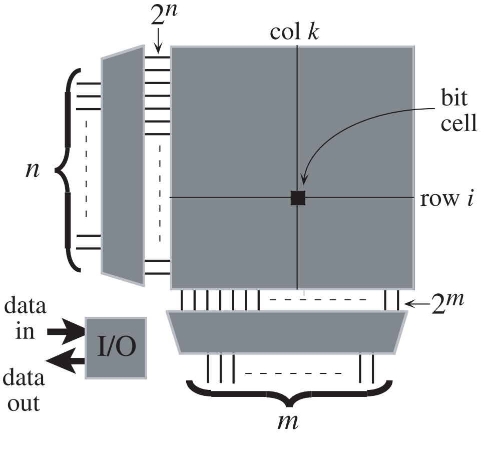


Figure 5.19. General architecture of a semiconductor memory.

good process monitors. Design regularity makes failure analysis easier than in random logic, since it is straightforward to map a logic failure to a physical location. High-density design provides good process monitoring, since transistors are designed for minimum dimensions of the technology, and conducting lines are kept as close as possible. These tight dimensions increase the probability of exposing process deficiencies.

The architecture of a static or dynamic semiconductor memory is shown in Figure 5.19. Memories have three major blocks: the memory array cells, the decoders, and the input–output circuitry. Memories can be bit- or word-oriented, accessing a single bit of the memory or the whole word (8, 16, 32, or 64 bits). In any case, the memory array is organized in rows and columns, with bits located at the intersection between a row and a column.

Each bit (or word) has a unique address that is mapped to physical locations with row and column decoders. The input–output circuitry performs the read or write data operations, i.e., store or retrieve the information in the memory.

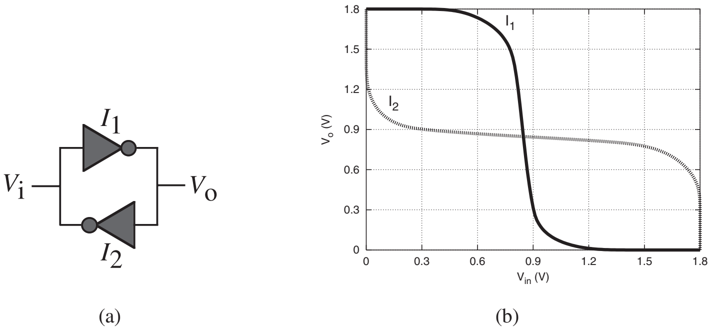
Row and column decoders take an address of $n + m$ bits, and select one word line out of 2^n and one column out of 2^m for bit-oriented memories. In word-oriented memories, the column decoder selects as many columns as the number of bits per word.

Static and dynamic memories have different cell designs. Dynamic memories store information in a capacitor, retaining data for a limited time, after which the information is lost due to leakage. Information can be retained at the expense of additional external circuitry and dedicated working modes to allow memory refreshment. When the memory is being refreshed, it cannot be accessed, and is said to be in a latency period.

Static memories store information in feedback structures (two cross-coupled inverters). They are faster than dynamic memories since static RAMs do not have latency periods, whereas dynamic memory cost per bit is cheaper because fewer transistors per cell are required.

Static Memories (SRAMs). Static semiconductor memories use two inverters in a bistable feedback design (Figure 5.20(a)). Bi-stable operation is illustrated by plotting the output versus input voltages on the same axes for both inverters [Figure 5.20(b)]. The stable quiescent states of the circuit are at the intersections where $V_\mathrm{i} = 0$ and $V_\mathrm{i} = V_\mathrm{DD}$, whereas the intersection voltage at $V_\mathrm{i} = V_0$ is not a stable state (called a metastable state). The system is called bi-stable, since only two states are stable.

The inverter feedback circuit retains its state as long as the power supply is maintained.



(a)

(b)

Figure 5.20. (a) Basic storage mechanism for static memories. (b) Input–output characteristics of the circuit.

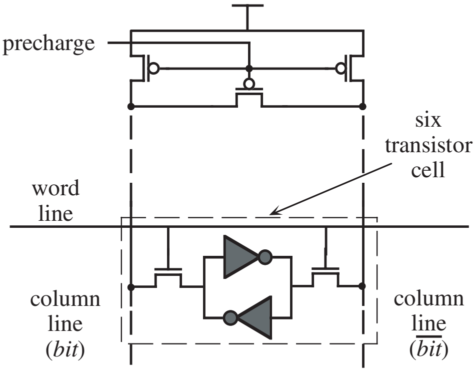


Figure 5.21. Six-transistor CMOS SRAM cell architecture.

Any "soft" voltage perturbation or possible current leakage in one node tending to switch the cell will be compensated for and overridden by the inverter output connected to this node. V_i in Figure 5.20(a) must have a stronger drive than the output of I_2. Memory cells typically set the memory state by driving V_i and V_0 simultaneously with opposite polarity signals.

The six-transistor cell architecture for a CMOS static RAM is given in Figure 5.21. All cell transistors and their interconnections are minimally sized to keep the array as small as possible. The word line controls the access transistors connecting the cell nodes to the column *bit* lines that run in pairs *bit* and $\overline{bit}$. When the word line is high, all cells in that row are connected to their corresponding *bit* and $\overline{bit}$ lines, and can, therefore, be accessed to read or write.

Memory read–write access time is reduced by precharging the *bit* and $\overline{bit}$ lines, i.e., forcing lines to the same voltage before any operation takes place. The precharge signal at the top of Figure 5.21 turns on all three *p*-channel transistors, forcing V_DD on both *bit* lines. When a write operation is performed, the *bit* and $\overline{bit}$ line drivers rapidly unbalance these lines, so that the correct value is stored in the memory. The precharge avoids the significant time for charging the highly capacitive *bit* lines when signals go from low to high. *n*MOS transistors pull down faster than equal sized *p*MOS pull-up transistors.

Memory cell inverters are minimally sized, but must drive long *bit* lines through a pass transistor during read operations. This potential delay can be improved using small analog circuits, called *sense amplifiers,* that are placed at each *bit* column output. Figure 5.22

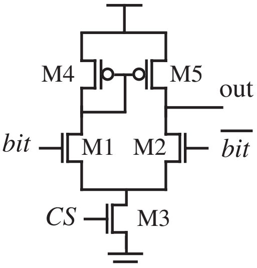


Figure 5.22. A differential sense amplifier used in SRAM memories.

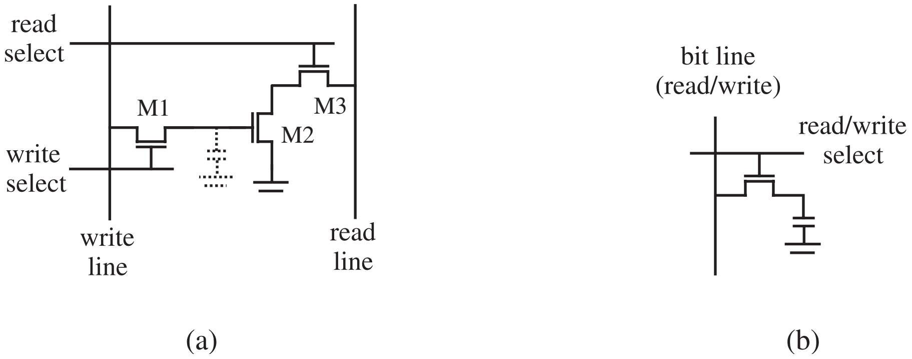


Figure 5.23. DRAM cells (a) Three-transistor cell. (b) One-transistor cell.

shows a typical differential sense amplifier used in CMOS SRAM designs. When the control signal CS is low, M_3 is off, and the sense amplifier output is floating. This corresponds to write operations. When CS is high, the circuit is activated. The sense amplifier reads *bit* and $\overline{bit}$ line voltages after precharge, and quickly transfers the cell value to the input–output circuitry, even before internal *bit* and $\overline{bit}$ lines reach steady voltages. If *bit* and CS are high, then M_1 drives current through M_4. The voltage drop across M_4 reduces the drain voltage at M_1. M_2 is off, and the *out* signal is pulled to V_{DD} through M_5. When *bit* is low and CS high, then M_2 turns on and *out* goes low. Sense amplifiers are only used during the read phase, and are disabled in other operations.

Dynamic Memories (DRAMs). Dynamic memory retains data as charge stored on a capacitor. This allows smaller memory cells, but since charge is not maintained by a feedback structure, stored values are lost with time and require refresh periods.

Two dynamic cell configurations are shown in Figure 5.23. Both cells use the parasitic gate capacitance of a MOS transistor to store the charge. The three-transistor cell (Figure 5.23(a)) has separate read and write select lines, giving a faster operation, but occupying more space. When the write select line is high, M_1 acts as a pass transistor, transferring the write line logic state to M_2 and putting M_2 in the off or conducting state. The drain $M_1 - M_2$ node capacitance holds that state. The read signal turns on M_3 and the data bit on the M_2 drain is passed through M_3 to the read line. This configuration allows for a nondestructive read operation, meaning that the cell does not lose its contents after a read is performed.

The single-transistor cell (Figure 5.23(b)) is popular since it has the smallest memory cell area. The charge stored in the cell storage capacitor is lost during the read operation because of charge sharing with the *bit* line parasitic capacitor, thus requiring a refresh operation during the same access cycle. The refresh operation uses circuitry that restores the original value in the cell once it is read.

5.3 INPUT–OUTPUT (I/O) CIRCUITRY

Input–output circuitry must link logic signals inside the IC to the outside world. The major I/O design problems are sufficient signal strength to drive large loads on printed circuit boards (PCBs) and IC internal circuitry protection from outside electrical assaults. Output current drive is typically achieved by using large output buffers that can have *W*/*L* ratios in the range of 1000–4000. I/O design is challenging and very technology dependent.

5.3.1 Input Circuitry: Protecting ICs from the Outside Environment

CMOS circuits need protection from electrical assaults of the outside environment, especially for circuit inputs since they are connected to transistor gates. Input devices are often exposed to electrical overstress (EOS) and electrostatic discharge (ESD) phenomena that are responsible for gate oxide ruptures and interconnect damage [1]. A person walking on a carpeted floor can accumulate over 20 kV of static charge. Contact between a charged human body and an IC pin can cause a several nanosecond discharge, leading to Ampere current peaks and pin voltages up to 4,000 V or greater!

ESD is the rapid transient discharge from picoseconds to nanoseconds of static charges when two dissimilar bodies come in contact. The transistor thin silicon dioxide (SiO_2) film of less than 25 Å is easily damaged. The operating electric fields that gate oxides typically use are between 2–5 MV/cm, whereas breakdown occurs between 10–18 MV/cm. It takes only a small extra gate voltage to push the oxides into rupture. Some IC fabrication steps induce ESD on internal transistors of the circuit, so that the phenomena are not just related to those transistors physically driven by pin connected inputs. ESD protection structures are designed within the IC, and can protect the circuit if designed well.

EOS delivers a high voltage for a longer time than ESD. EOS times between microseconds to seconds cause more visible damage to the IC than ESD. ESD and EOS have different properties and root causes, but both destroy ICs. ESD typically occurs when a circuit contacts a charged machine or human, and EOS comes from aberrant longer pulses from power supplies, testers, lightning, or general misuse, such as mounting a package backward. EOS protection strategies often seek to eliminate the problem at the system level.

Different strategies are adopted to protect input structures against ESD. Elements are connected between the input PAD, the transistor gates, and the power rails to provide safe discharge paths when ESD occurs. These elements are inactive as long as the voltage levels of the node are within the normal operating conditions of the device.

When ESD assaults the IC, the protection circuit must drive the excess charge to the power or ground rails, steering that damaging energy away from the transistor gate oxide or metal interconnects. These protection devices are diodes and/or transistors working out of their normal operating ranges at high voltages and currents. Protection devices must sink large currents in nanosecond response times, suppressing heating effects and high electric fields. The protection circuits must survive the static energy assault to continue their protective function.

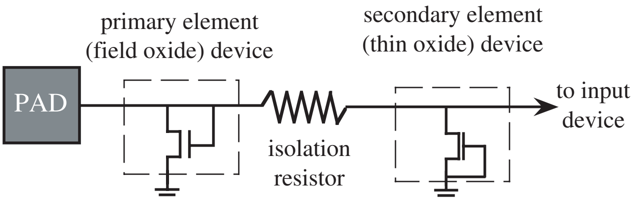


Figure 5.24. Example of input protection scheme against ESD.

ESD protection circuit design greatly depends on the technology, and is very layout sensitive. A common protection circuit with two protecting elements and a resistor is shown in Figure 5.24. The primary element takes most of the current during the ESD event, whereas the secondary element gives rapid initial protection to the logic gate input until the primary device turns on. The resistor provides a voltage drop to isolate both elements, allowing high voltage operation of the primary element while the voltage at the gate input can be maintained at a lower value. There are several ways to design the primary and secondary elements, so only the basics are described here (for more information refer to [1]).

The primary input protection circuitry in MOS technologies may use a field oxide transistor with a triggering voltage of about 30–40 V (this greatly depends on the technology). The secondary protection device is a grounded gate *n*MOS transistor reaching its trigger breakdown voltage (called snapback) rapidly before the primary protection circuit turns on. The current through this secondary device causes a voltage drop across the resistor that increases the PAD voltage to a value at which the field oxide transistor triggers, and takes most of the current.

5.3.2 Input Circuitry: Providing “Clean” Input Levels

Input circuitry must provide “clean” or noise-free logic levels to the internal circuitry. An external noisy or ringing input transition may induce multiple switching at the gate output. A solution uses a circuit with a static hysteresis transfer characteristic. The input voltage for which the output responds to a low-to-high or high-to-low transition depends on the output voltage. This circuit is known as an Schmitt trigger. Figure 5.25 shows both the Schmitt trigger transistor-level design and the input–output transfer characteristic. The feedback transistors whose gates are connected to the output provide the hysteresis [2].

The Schmitt trigger circuit in Figure 5.25 has only six transistors, but it has complexity whose explanation will bring together many CMOS concepts [2]. Start by setting $V_{\text{in}} = 0$ V and tracking the three *n*MOS transistors as they change state. When $V_{\text{in}} = 0$ V, M_1 and M_2 are off and M_3 is a pass transistor driven fully by the high logic voltage at V_{out}. M_3 will pass V_{DD} to the source of M_2 (drain of M_1) with a weak voltage of $V_{\text{DD}} - V_{tn3}$. As V_{in} rises to V_{tn1}, M_1 turns on. M_2 is still off since $V_{\text{GS2}} = V_{tn1} - (V_{\text{DD}} - V_{tn3})$. As V_{in} rises above $V_{\text{in}} = V_{tn1}$, M_1 conducts through M_3 and V_{S2} begins to fall. M_2 will turn on when $V_{\text{in}} - V_{\text{S2}} = V_{tn2}$. M_2 source and bulk are at different voltages so M_2 will have an elevated threshold voltage, the same as M_3. M_2 conduction now allows a more rapid drop in V_{S2} and V_{out} with

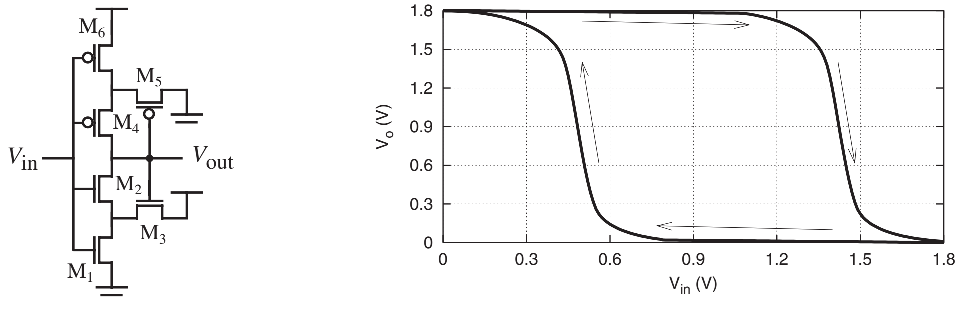


Figure 5.25. **(a)** Schmitt trigger CMOS design and (b) transfer characteristic.

the onset of transition higher than V_t as in a normal inverter. A similar analysis exists, starting with $V_{\text{in}} = V_{\text{DD}}$, that watches the transistor actions as V_{in} drops. The *p*-channel transistors respond to a different level when switching the output voltage to a logic high.

Analytically, the design is examined by equating the saturated state drain current expressions for M_1 and M_3:

$$K_1'(V_{\text{GS1}} - V_t)^2 = K_3'(V_{\text{GS3}} - V_t)^2 \tag{5.1}$$

We define the switching point of the low- to high-input transition V_{SPH} as the input voltage at which M_2 starts to conduct ($V_{\text{GS2}} = V_{t2}$), i.e.,

$$V_{\text{in}} - V_{\text{S2}} = V_{t2} \tag{2.2}$$

Setting the body effect threshold voltages of M_2 and M_3 equal, from Equation (5.2) the switch point is $V_{\text{SPH}} = V_{\text{S3}} - V_{tn3}$, giving

$$\frac{K_1'}{K_3'} = \frac{(V_{\text{DD}} - V_{\text{SPH}})^2}{(V_{\text{SPH}} - V_{tn1})^2} = \frac{L_3}{W_3}\frac{W_1}{L_1} \tag{5.3}$$

The conduction control of the *p*MOS transistors follows a similar analysis, giving

$$\frac{K_6'}{K_5'} = \frac{(V_{\text{SPL}})^2}{(V_{\text{DD}} - V_{\text{SPL}} - V_{tp6})^2} = \frac{W_6}{L_6}\frac{L_5}{W_5} \tag{5.4}$$

The transistor widths and lengths can be designed to achieve a given V_{SPH} and V_{SPL}.

5.3.3 Output Circuitry, Driving Large Loads

IC output circuitry must have strong signal strength to drive other circuits at the PCB level. Large capacitive loads driven at high speed require a large current in a small time. Driving large capacitances is not only an I/O design problem, it also appears within the device when driving long lines or bus lines. This drive is achieved with large transistors that have large input capacitance values because of their size. Specific circuits can drive large loads such as the tapered buffers described in Section 4.2.

Latchup in CMOS Technologies.* Large currents can be triggered by a bipolar mechanism called *latchup.* CMOS technology uses *n*-type and *p*-type transistors on the same substrate. Many processes start with a uniformly doped substrate, and construct wells of opposite doping to fabricate both MOS transistor types. This structure has an inherent *pnpn* parasitic bipolar transistor structure shown in Figure 5.26 that is off in normal operation and does not contribute to the circuit behavior. Latchup occurs when a parasitic *pnpn* structure underlying the CMOS structure is turned on, driving large currents and damaging the whole circuit.

The underlying parasitic bipolar transistors are connected with positive feedback, so that once the structure is triggered, the current increases until the device is destroyed. If proper rules are not followed during design or the circuit is operated improperly, then the parasitic bipolar structure may be triggered on, causing severe circuit damage.

*This subsection requires the reader to have a knowledge of bipolar (BJT) transistor princicples.

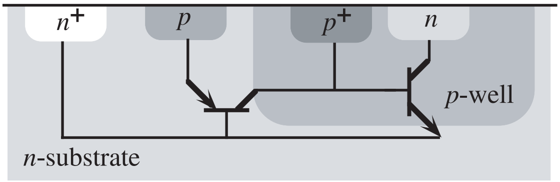


Figure 5.26. Cross section of a CMOS circuit fabricated with a single well and parasitic bipolar devices associated with such a technology.

The CMOS structure with diffused wells in Figure 5.26 shows the parasitic bipolar transistor structure underlying the circuit. The parasitic bipolar devices are connected such that the collector terminal of one device is connected to the base of another in a closed positive-feedback loop.

If an excess of carriers reach the base of some of the parasitic bipolar transistors, the current is amplified at its collector terminal, driving the base of the other bipolar device. This positive feedback connection can increase the current without limit. Figure 5.27 shows the current–voltage characteristic of the parasitic bipolar structure within a CMOS single-well process. Once the structure is triggered (the voltage goes beyond V_{trig}), lowering the voltage does not decrease the current because of the positive feedback. The only way to cut the current through the device is to completely switch off the power supply of the circuit.

Latchup is prevented by proper design that avoids activating a parasitic structure, since this cannot be eliminated. One latchup mechanism uses hot electrons from saturated MOS devices, causing holes to be injected into the substrate. If those holes are not properly collected at substrate and bulk contacts, they may diffuse and cause a voltage drop within the substrate (or well) that is enough to turn on a parasitic bipolar device. High substrate currents are another latchup source. Design strategies to avoid latchup are beyond the scope of this book and can be found in [1]. The modern trend toward SOI technologies and power supplies lowered to around $V_{\mathrm{DD}} = 1$ V lessen the threat of latchup

5.3.4 Input–Output Circuitry: Providing Bidirectional Pins

Microprocessors, microcontrollers, programmable logic (FPGA), and memories use bidirectional (i.e., input–output) pins. Depending on the circuit design, certain pins are logic inputs for some operations and logic outputs for others. Bidirectional pins reduce the overall circuit pin count. These pins must have proper protections for the gates that will process the inputs, and also provide enough driving capability when acting as outputs.

Figure 5.28 shows a commonly used design to control bidirectional I/Os. When the control signal OE is low, the logic from inside the circuit (data out) is driven onto the output PAD through the strong output transistors. When the control signal input/output (OE) is high, both output devices are off, and the PAD acts as an input to the circuit.

5.4 SUMMARY

This chapter raises the level of transistor integration, showing how primitive CMOS complementary combinational logic gate designs are built from Boolean algebra equations. More compact circuits that have different power dissipation and speed properties illustrate

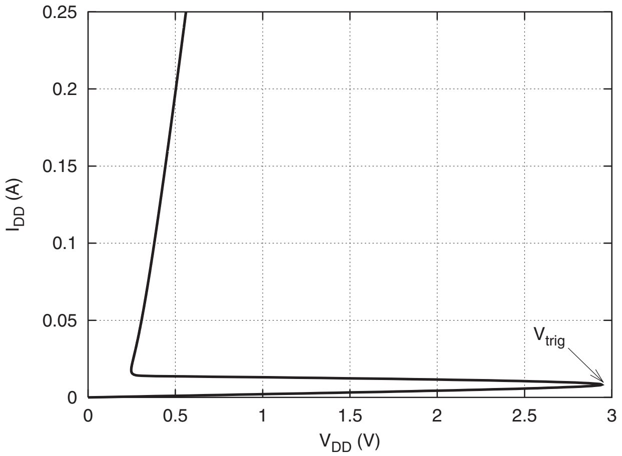


Figure 5.27. Current voltage characteristics of a parasitic bipolar structure underlying a CMOS single-well process.

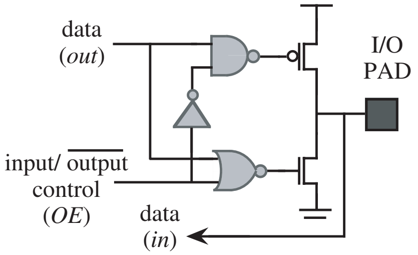


Figure 5.28. A bidirectional I/O circuit.

the popular tri-state gate, pass transistors, and dynamic logic gates. All versions appear in modern CMOS IC design. Sequential or memory-storing circuits partner with combinational logic to build complete ICs. Latches are the first building block, but have transparency properties eliminated by combining latches and transmission gates into flip-flops. Finally, the latchup failure mechanism and important input/output circuits were described.

REFERENCES

1. A. Amerasekera and C. Duvvury, *ESD in Silicon Integrated Circuits,* Wiley, 1995.
2. R. J. Baker, H. W. Li, and D. E. Boyce, *CMOS Circuit Design, Layout, and Simulation,* IEEE Press, 1997.
3. K. Bernstein, K. Carrig, C. Durham, P. Hansen, D. Hogenmiller, E. Nowak, and N. Rohrer, *High Speed CMOS Design Styles,* Kluwer Academic Publishers, 1998.
4. R. H. Krambeck, et al., "High speed compact circuits with CMOS," *IEEE Journal of Solid State Circuits, 17,* 3, June 1982, 614–619.
5. J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits,* Prentice-Hall, 2003.
6. N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design, A Systems Perspective,* 2nd ed., Addison-Wesley, 1993.

EXERCISES

5.1. Given the Boolean function $F = z[\overline{x}yz + x\overline{z}]$, draw the static CMOS transistor schematic.

5.2. Write the Boolean expression F for A, B, and C in the circuit in Figure 5.29.

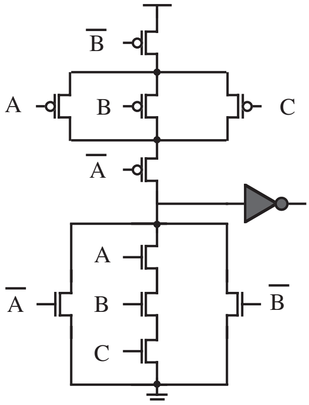


Figure 5.29.

5.3. Draw the static CMOS transistor schematic that performs the Boolean function $F = (g + f) \cdot (m + n)$.

5.4. Draw the CMOS transistor schematic that fulfills the function $F = \overline{[(A \cdot B) + C] \cdot D}$ for both a static and a domino CMOS logic gate.

5.5. Given the schematics of Figure 5.30:

(a) If it corresponds to the CMOS pull-up network of a static circuit, what is the resultant Boolean expression F?

(b) If the p- and the n-channel transistors are sized for equal drive current, discuss whether the pull-up will be faster than the pull-down network, or they will they be the same.

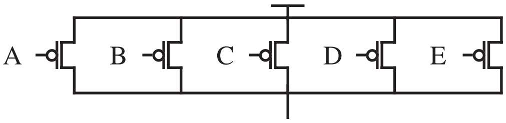


Figure 5.30.

5.6. What Boolean function will the circuit in Figure 5.31 perform?

5.7. Determine the logic function of the circuit in Figure 5.32.

5.8. Given the circuit of Figure 5.33:

(a) Determine the role of A, B, and C nodes (input or output).

(b) Determine its Boolean function.

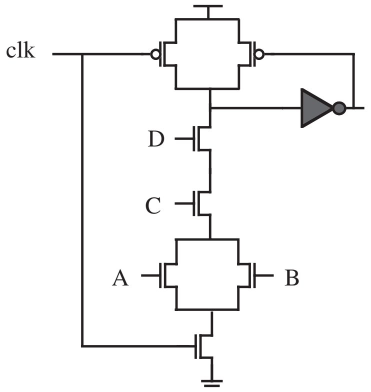


Figure 5.31.

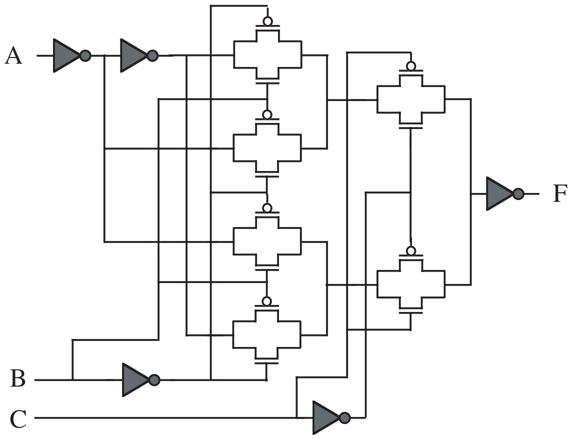


Figure 5.32.

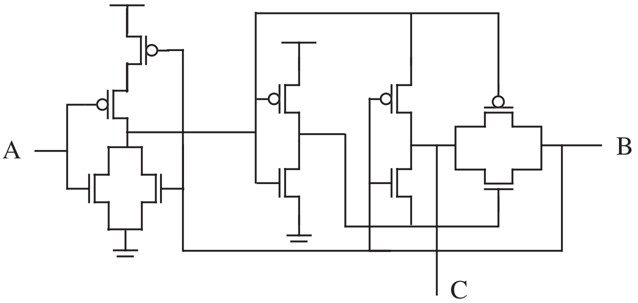


Figure 5.33.

5.9. The circuit of Figure 5.34 has the same function as a basic block used in sequential circuits. Identify the circuit type and the conventional names given to the inputs and outputs. Hint: analyze the equivalent circuit for $y = 0$, and then for $y = 1$.

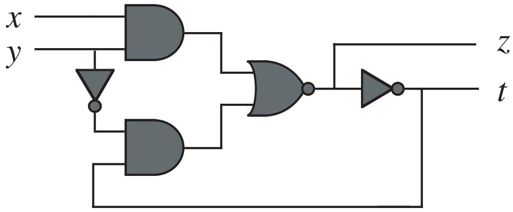


Figure 5.34.

5.10. Figure 5.20(b) shows the transfer properties of a simple static memory circuit. Suppose the input V_i is a short pulse with amplitude 0.6 V_{DD}. If V_0 drives another latch, what is the effect on (a) overall timing, (b) noise sensitivity (margin).

5.11. Identify the function and the input/output conventional node names for the circuit in Figure 5.35.

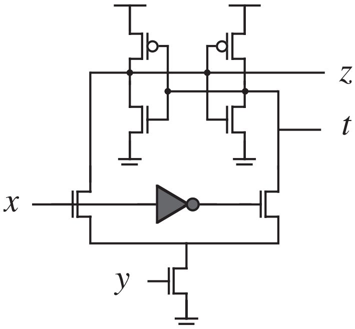


Figure 5.35.

5.12. Combine two circuits of Exercise 5.11 to get a flip-flop.

5.13. (a) What is the difference between EOS and ESD?
(b) If an input protection circuit protects the inner core logic from an ESD assault, but it is damaged, has the protection circuit done its job?

5.14. The DRAM circuits in Figure 5.23 store the bit (voltage) information on a capacitor. Use knowledge from Chapter 2 to determine the affect on refresh frequency if the temperature rises.

5.15. Observe the Schmitt trigger circuit in Figure 5.25(a). Explain how the transfer curves in Figure 5.25(b) behave as the input signal drops from V_{DD} to 0 V. Describe the transistor action.

5.16. If latchup occurs in a CMOS circuit and draws a large current, how do you stop it?

PART II

FAILURE MODES, DEFECTS, AND TESTING OF CMOS ICs

CHAPTER 6

FAILURE MECHANISMS IN CMOS IC MATERIALS

6.1 INTRODUCTION

A single modern IC may have more than a billion transistors, miles of narrow metal interconnect ions, and billions of metal vias or contacts. As circuit entities, metal structures dominate the semiconductor transistors, and their description is as challenging and necessary as that of semiconductors. Transistor oxides have shrunk, and dimensions are now on the order of 5–7 SiO_2 (silicon dioxide) molecules thick. Metal and oxide materials have failure modes that have always been with us, but are now even more significant in the deep-submicron technologies. This chapter addresses these IC failure modes that are caused by failure of materials.

This chapter has three sections:

1. Materials Science of IC Metals
2. Metal Failure Modes
3. Oxide Properties and Failure Modes

We do not seek to overwhelm the reader with mathematical detail, but rather to use visual models and learn the conditions that cause IC materials to fail in time due to interconnect bridges opens or damaged oxides. Circuit failure modes challenge the quality and reliability of large deep-submicron ICs. High-performance ICs now push safety margins closer to expected product lifetimes than before to achieve high clock frequencies. Added concerns are that previously dormant metal and oxide failure modes may now appear at test or during product life.

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

We refer to *intrinsic* material as defect-free, and *extrinsic* material as containing defects. Reliability failure modes are typically identified with failures of intrinsic material, whereas extrinsic material is linked to burn-in yield or production test failures. There are gray areas in this division, since some latent failure modes are related to (extrinsic) defects present during fabrication, but are too difficult to detect at test. Also, violation of fabrication procedures or design rules are known to cause postfabrication failures. Defective thin oxide that leads to rupture of the oxide, or metal stress voiding as a serious metal open failure mechanism are examples that are described later. Parts can fail at the next level of assembly or in the field, but detection of defects causing these serious failures may be almost impossible earlier in the product cycle. Test engineering deals with extrinsic materials, whereas reliability and failure analysis engineers deal with both material types.

Materials science studies the defects of a solid and their relation to the solid's physical properties [2, 3, 4]. The materials foundation of modern ICs were built on this old science. Several material classes exist, but we will only study metals and dielectrics. For many years, Al (aluminum) dominated signal and power interconnections, and SiO_2 was the thin and thick oxide material of the IC. Cu (copper) is now the popular interconnect material and new dielectrics are being developed to replace SiO_2. It is a nontrivial challenge to introduce new materials into the known and controlled CMOS IC manufacturing process.

6.2 MATERIALS SCIENCE OF IC METALS

Interconnect metals are thin films made of small, single crystals called grains. Metal grains are crystals similar to silicon crystals, but their grain surfaces are irregular and not smooth like silicon crystals. Figure 6.1 is a photograph of a polycrystalline Al line with its grain boundaries marked artificially for clarity. The irregular *grain boundaries* are important interface regions that influence the metal resistance against forces that can move atoms and lead to open or bridge metal lines. Metal failures involve extrusions or voids, and so require movement of Al atoms along easy paths, such as grain boundaries. Grain boundaries are about 1–2 atoms wide and are relatively open spaces, allowing easy travel for moving atoms. If a line has large grains and thus fewer grain boundaries, atoms have less opportunity for displacement. The metal in Figure 6.1 has many paths for a mobile metal atom to follow, increasing the likelihood of net Al atom dislocation in the presence of forces in the metal. This chapter looks at these metal forces that derive from electron current, temperature gradients, atomic concentration gradients, and mechanical stress forces.

Figure 6.2 shows two of the 14 possible crystal structures (Bravais lattices). The dots represent the center of the atoms since atomic volume usually extends to neighboring atoms. Figure 6.2(a) shows the corner atoms of a unit crystal with an additional atom

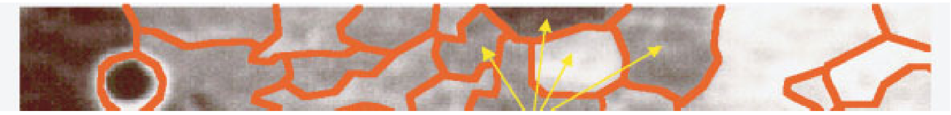


Figure 6.1. Al grains in IC interconnect. (Reproduced by permission of Bill Miller, Sandia National Labs.)

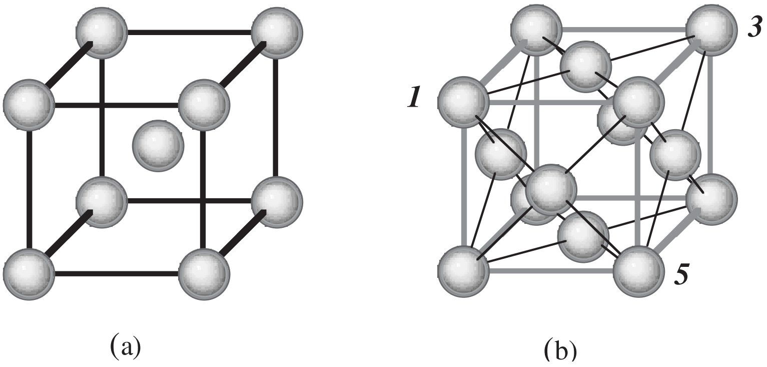


Figure 6.2. (a) Body-centered cubic cell (W). (b) Face-centered cubic cell (Al, Cu).

placed in the middle of the cube. This structure is called a *body-centered cubic* (bcc) cell, and is found in W (tungsten). The dots on the corners are drawn larger than the one in the middle for viewing clarity. Figure 6.2(b) shows a *face-centered cubic* (fcc) cell with an atom placed in the center of each face of the unit crystal. This is the structure of Al and Cu.

The unit crystals in Figure 6.2 can interpret the variable strength of a metal when forces are applied in any direction. Some directions are very resistant to applied force, and other directions are weak. Figure 6.2b numbers corner atoms 1, 3, and 5. If the unit crystal is viewed perpendicular to the plane of these three atoms, and the atoms are enlarged to represent their full diameters, then we see six atoms (Figure 6.3). The corner and face-centered atoms touch, and that surface is called the close-packed plane. The atoms bond strongly in this plane, and lateral forces that try to pull the atoms apart find it difficult to do so. This plane is called the 111 plane, following crystal convention. A force perpendicular to the close-packed plane dislodges atoms more easily because nearest neighbor atoms in this direction are further apart, and bonding strength is less. A fortunate result of laying down Al interconnections on the substrate of an IC is that the 111 texture is energetically favorable. Deposited atoms fill up available space next to the substrate just as oranges do when dumped on the bottom of a grocery fruit bin. This forms a crystal in the 111 plane that is the densest plane, so that more atoms can get closer to the substrate. This places the close-packed plane parallel to the majority of forces that act on the horizontal plane (especially electron flow), providing a natural resistance to dislocation of atoms.

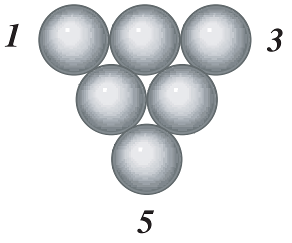


Figure 6.3. Close packed plane of fcc metal.

Metal crystals are imperfect. They have defects that can be intentional, such as alloyed metals, or unintentional from contamination by other atoms or thermal dislocations in a pure metal. Figure 6.4 illustrates three important metal crystal defects. Small atoms, such as B (boron) or H (hydrogen), can squeeze between larger metal atoms and form *interstitials.* Their small size allows easy diffusion within the host material. Some atoms can replace a host metal atom and are called *substitutionals.* Copper forms substitutionals in Al, and is intentionally alloyed with Al to strengthen the metal. A third defect appears naturally due to thermal vibration in otherwise perfect crystals, and is called a *vacancy.*

The vacancies in Figure 6.4 are locations in the metal lattice where an atom is missing. This is a natural condition of the high-frequency, thermal-induced vibrations of atoms. At any instant, there is a probability that an atom has vibrated out of its host position. There are no vacancies at 0 K, but they increase exponentially as temperature rises. Vacancy existence is a condition for movement of metal atoms. When metal atoms move within the lattice, such as in Figure 6.4, they must have a place to go. If an atom jumps into a vacancy, then it creates a vacancy at the location it left. Motion of vacancies is similar to the hole motion concept in semiconductors, except that vacancies don't have an electrical charge.

Metals have line, area, and volume defects besides the atomic defects in Figure 6.4. Volume defects are large metal voids or precipitates. A common line defect called an *edge dislocation* is drawn in Figure 6.5. The regular array of metal atoms is missing a few atoms in one of the crystal planes. Stress forces appear in the region of the circle. The top three atoms in the circle are in compression, and the bottom two atoms are in tension. The metal wants to relieve this energy. The compressive and tensile forces encourage a movement of atoms along the slip plane. The significance of these defects is that the metal is less resistant to holding its structure in the face of external forces (described later). Its strength is less than that of a perfect metal. When stresses are relieved to lower energy states, such as in heat annealing of metal, then conductivity increases. Electrons move easier in a crystal without alloy elements or other defects, although alloy elements are intentionally added to provide mechanical strength.

Grain boundaries are area defects. Figure 6.6 is a photograph of a narrower metal line than the one in Figure 6.1. The arrows locate points where three grain boundaries merge to form a *triple point.* If Al atoms move in a single grain boundary that merges with two other grain boundaries, then at the merger, more atoms can leave that point than enter (or vice versa). This is called a *flux divergence site.* For example, if Al is moving from right to left in Figure 6.6, then the first triple point will show more atoms entering the point than

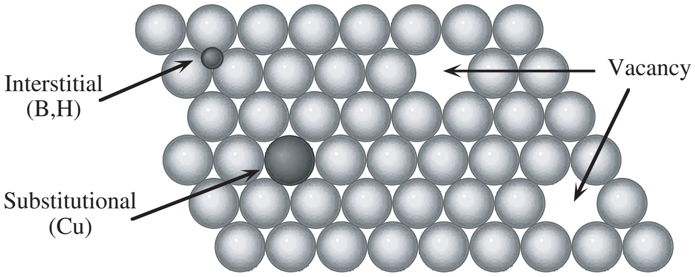


Figure 6.4. Imperfections in a metal crystal. Cu is not drawn to scale; it is about two times larger than Al.

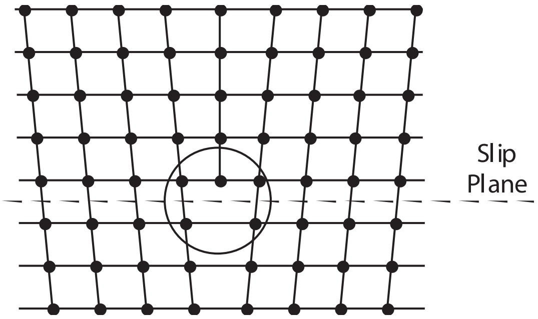


Figure 6.5. Edge dislocation and slip plane.

leaving, causing compressive buildup of atoms. At the second triple point, more atoms will leave than enter, causing tensile forces. The opposite polarity stress sets up a stress gradient.

The vertical grain boundary shown toward the left in Figure 6.6 is typical for metal lines of <0.5 μm, in which grains are large compared to the metal width. The very small lines in deep-submicron ICs have mostly these vertical or *bamboo* structures. You will notice that whereas an Al atom may move laterally in Figure 6.6 along longitudinal grain boundaries, it virtually stops when it encounters an orthogonal grain boundary. Metal lines with bamboo structures are generally stronger than those with triple points

The last topic in this capsule view of materials science concerns the physical gradients that can drive metal atoms. The link between physical gradients and particle (atoms) motion is subtle. Consider four potential gradients that drive metal atoms:

1. *dC*/*dx* Concentration gradient of atoms—diffusion
2. *dT*/*dx* Temperature gradient in solids—thermotransport
3. *dV*/*dx* Voltage gradient in solid—electromigration
4. *d*σ/*dx* Stress gradient in solid—stress voiding

The concentration gradient causes net atom motion because all atoms are in thermally induced motion. The direction of motion is random, so the region of denser atoms will have more atoms diffusing toward the region of lower density than vice versa. A net dislocation of atoms occurs. Similarly, a temperature gradient has a region of higher-energy (chemical potential) atoms, and the more energetic ones can move to the lower chemical potential energy, similar to the motion induced by a concentration gradient. The voltage and stress gradient forces are discussed in detail since they relate to the electromigration and

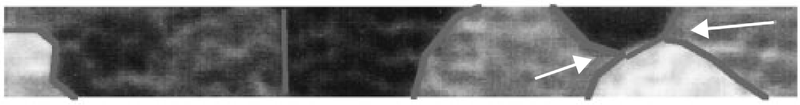


Figure 6.6. Al metal and (marked) grain boundaries. (Reproduced by permission of Bill Miller, Sandia National Labs.)

stress voiding metal failures discussed later. Fick's diffusion laws and the concept of chemical potentials underlie all four forces.

Diffusion can be measured for the many atomic elements under diverse conditions such as solid, liquid, or gas phases, or for different structural conditions. Diffusivity denotes the measurement of diffusion, and is defined by Einstein's relation: $D = \mu \cdot kT$ (cm^2/s), where μ = particle mobility and $k = 1.38 \times 10^{-23}$ Joules/K (Boltzmann's constant). D is exponentially related to Kelvin temperature (T), an activation energy (E_a), and a material-dependent constant (D_0) by

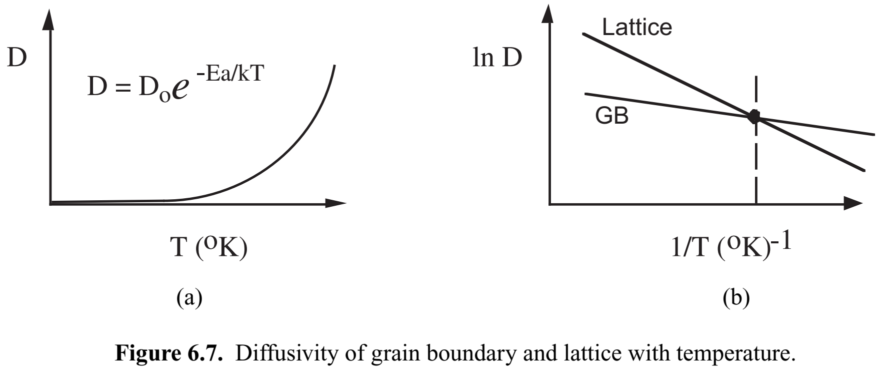
$$D = D_0 e^{-E_a/kT} \tag{6.1}$$

Figure 6.7(a) shows the exponential diffusivity (D) relation to temperature, and Figure 6.7(b) plots the commonly presented straight line of ln(D) versus T^{-1}.

Generic curves are shown for diffusivity (Figure 6.7b) when the metal atoms are in a GB (grain boundary) or diffusing within the crystal lattice. GBs have a diameter that is about twice the size of an atomic diameter, and at low temperatures (right-hand side of x-axis), the grain boundaries have a higher diffusivity than the tightly packed crystal lattice region. As temperature rises, the diffusivity of atoms in the lattice increases and surpasses that of grain boundaries. Metallurgists estimate that the crossover temperature is between 0.4 T_{mp} to 0.6 T_{mp} (T_{mp} = metal melting point). Al melts at T_{mp} = 660°C = 933 K, so that the estimated temperatures at which diffusivity crossover occurs are between 100°C to 287°C.

Modern high-performance ICs have average package temperatures above 100°C and IC hot spots of even higher temperatures. The junction temperature of an IC is defined as the temperature of the silicon substrate, and it is a crucial parameter of reliability-prediction procedures and burn-in testing. The measured junction temperature of a 1 GHz 64-bit RISC microprocessor implemented in 0.18 μm CMOS technology was reported as 135°C at V_{DD} = 1.9 V [1]. This microprocessor had 15.2 million transistors packed in the 210 mm^2 chip area. High temperature allows higher diffusivity of metal atoms, which can lead to shorter failure times.

Review. Metal studies began over 100 years ago, and their application to IC thin metal films (interconnections) has a solid knowledge base. Metals are polycrystalline, with



(a)

(b)

Figure 6.7. Diffusivity of grain boundary and lattice with temperature.

grain boundaries that separate the grains and influence the quality of the metal. Metal defects and high temperature make it easier for atoms to move within a metal, and reduction of these effects requires strong effort by industry. Concentration, voltage, temperature, and mechanical-stress gradients will move metal atoms within an interconnect. The movement of atoms leads to open or bridging circuit defects that are the next topics.

6.3 METAL FAILURE MODES

6.3.1 Electromigration

Electromigration is the net movement of metal under the influence of electron flow and temperature. A metal line will fail if sufficient current density and high temperature are applied. Metals can form a void or may form an extrusion that projects from one of the surfaces of the metal. This failure mechanism is called *electromigration* (EM). The abundant knowledge about aluminum (Al) failures is presented first, followed by a description of copper (Cu) failure mechanisms.

Figure 6.8 is an unusual photograph of two electromigration failure sites in a wide, unpassivated Al line. Here, Al atoms presumably exited the voided region and moved to the extruded bulging region on the left. Electromigration almost stopped the IC industry in the 1960s until methods were found to control electromigration. Electromigration studies began over 35 years ago, and much is known about this failure mechanism.

Electrons are believed to transfer a small but sufficient momentum to thermally active metal atoms forcing those atoms, out of their lattice sites, and moving them under diffusion in the same direction as the electrons. Figure 6.9 illustrates an Al metal line with electrons moving from right to left and colliding with Al atoms. If the thermal energy of Al is at a level such that a small nudge from many electrons dislodges it, then it will move if there is a vacancy to move into. That critical energy is called a saddle point. A small tensile stress is

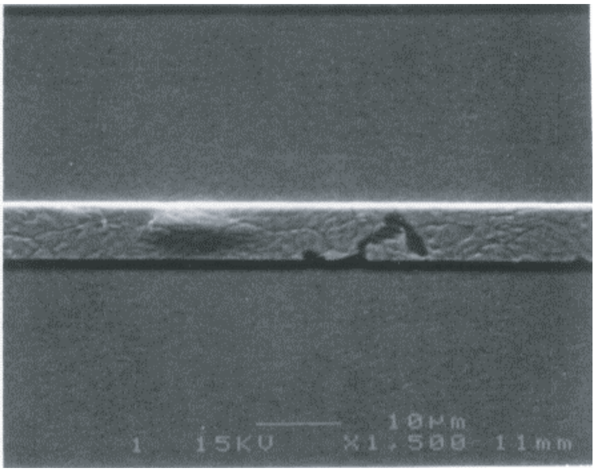


Figure 6.8. Electromigration SEM photo in a wide, unpassivated metal line. (Reproduced by permission of Joe Clement, Sandia National Labs.) Electron current and Al atom motion is from right to left.

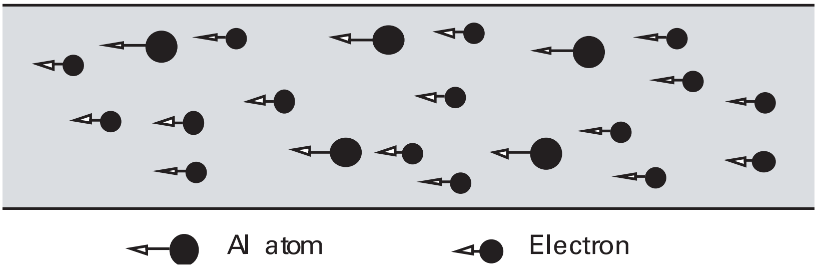


Figure 6.9. Representation of moving electrons and metal atoms during electromigration.

created where an atom is knocked from its lattice site, while downstream, the displaced metal atom creates compressive forces and possible extrusions. The region between the compressive and tensile stresses is under a stress gradient. Figure 6.8 shows these concepts, where the electron current (j) direction displaces Al atoms in the region of the void, and those atoms pile up downstream, forming an extrusion through passivation.

The variables that affect electromigration are clues used to reduce the threat, and we will present a model of electromigration using these variables. A flux J is the number of particles (atoms in this case) crossing a unit area per unit time. Table 6.1 lists the equations and subequations that lead to an expression for the atomic flux J_{em} of an electromigrating metal atom [9]. Equations are given in the middle column and substituted variables are given in the right column.

The final flux expression for electromigrating atoms is

$$J_{\text{em}} = \frac{D}{kT}(Z^* \cdot q)\rho \cdot j_{\text{e}} \cdot N\left(\frac{\text{atoms}}{\text{cm}^2 \cdot \text{s}}\right) \tag{6.2}$$

Equation (6.2) is not typically used to calculate numbers, but to identify variables that influence electromigration. The flux is governed by diffusion and electromigrating atoms, and is said to be under a biased or directed diffusion. The temperature sensitivity of electromigration diffusion in Equation (6.2) is seen in Equation (6.1) showing the exponential temperature dependence.

Table 6.1. Electromigration Flux J_{em} Derivation

Magnitude	Equation	Symbols
Atomic flux J_m	$J_{\text{em}} = vN$	v = mean velocity of metal atoms N = concentration of moving metal atoms
Effective charge Q	$Q = Z^*q$	Z^* = effective charge factor q = electron charge
Atom mobility	$\mu = QD/kT$	D = diffusivity kT = Boltzmann's thermal energy
Atom mean velocity	$v = \mu\mathscr{E}$	μ = atom mobility $\mathscr{E}$ = electric field
Electric field	$\mathscr{E} = \rho j_{\text{e}}$	ρ = metal resistivity j_{e} = electron current density

The RHS of Equation (6.2) has an electron current flux term j_e, whereas the LHS is a flux of atoms. The Z^* term performs a cross coupling in the equation from electron current density to metal atom flux. One qualitative interpretation of Z^* is that it is the ratio of impinging electrons to a single metal atom that is moved. The driving electric field for electrons is the product of resistivity and electron current density across a 1 cm length, and the units are V/cm. Equation (6.2) reinforces that at a given current density and temperature, metal atoms will move along the metal stripe in the same direction as the electrons.

Electromigration Failure Time. A different analysis looks at the electromigration variables that affect the time for a metal line to fail, t_F. James Black made major contributions to understanding electromigration failure [5]. He empirically derived Black's law, which states that the median time to failure (t_F) of a group of Al interconnects is

$$t_F = \frac{A_0}{j_e^2}\, e^{E_a/kT} \tag{6.3}$$

A_0 is a technology-dependent constant, T is temperature in Kelvin, j_e is electron current density (A/cm^2), and E_a is the activation energy (eV) for electromigration failure. Metal grain structures vary considerably, so that t_F is a statistical quantity not a predictor of single line failure time. Black's law shows the electromigration failure dependence on j_e^{-2}. When a design is shrunk, current density usually increases, so that designers must work to reduce j_e. An older DC design rule for CMOS technologies kept $j_e < 0.2 - 1$ mA/μm^2. This was a crude estimate with a built-in safety factor. Electromigration design rules for metal dimensions are now taken from circuit simulators that compute average current density and expected temperature in a local interconnection. The waveshape of the current pulses and the dimensions of the metal line needed to assure metal safety are calculated from these variables. Although j_e reduction is important, the exponential temperature term has a more acute effect on t_F.

The Black's law exponent of j_e is given as $n = 2$, as the original equation stated. Subsequent knowledge showed that n can vary from $1 < n < 2$ depending on the metal structure. We will use $n = 2$ for simplicity.

Black's law predicts failure time, and that is useful in test structure studies under accelerated temperature and current density. t_F is typically measured as the median time to failure of many test structures. However, Equation (6.3) is more often used to calculate E_a for the metal when t_F is measured. E_a is the accepted parameter to assess the electromigration quality of a metal technology. If t_F is measured over a temperature range, then E_a can be calculated from the slope of a plot of ln (t_F) versus (1/kT). If E_a is in the low range of $E_a \approx$ 0.4 to 0.6 eV, then the quality of the metal is poor, usually indicating an abundance of grain boundary paths. If $E_a \approx$ 0.8 eV to 1.0 eV, then metal quality is good and indicative of bamboo structures. The calculated upper limit on E_a is that of a pure Al crystal, where $E_a \approx$ 1.48 eV. A metal interconnect line made of a pure, single Al crystal would be the best structure possible, but this has not yet been practical.

■ EXAMPLE 6.1

Use Black's law [Equation (6.3)] to estimate the reduction in useful product life if a metal line is initially run at 55°C at a maximum line current density of 0.6

MA/cm^2, and then run at 110°C and 2 MA/cm^2. Use E_a = 0.7 eV and Boltzmann's constant of k = 86.17 μeV/K.

Black's law can be written for T = 55°C = 328 K and j_e = 0.6 MA/cm^2. Then divide by Black's law for T = 110 °C = 383 K and j_e = 2 MA/cm^2. A_0 cancels, and we get the electromigration acceleration factor A_{EM}

$$A_{EM} = \frac{t_{F1}(328\text{ K})}{t_{F2}(383\text{ K})} = \frac{j_{e2}^2}{j_{e1}^2} = e^{E_a/k[(1/328)-(1/383)]} = \frac{(2.0\text{ MA})^2}{(0.6\text{ MA})^2}\, e^{0.7/k[(1/328)-(1/383)]} = 389.4$$

■

Typical practice rounds A_{EM} off to a single digit, or A_{EM} = 400. The result is that the hotter part has an estimated useful life shortened by a factor of 400. If nominal life is projected for 15 years for the cool part, then the hot part has a predicted life of about 2 weeks. This example is realistic since modern high-performance ICs have package temperatures of 100 °C, and the dies are even hotter. Designers push the j_e limits, since this increases clock frequency performance. This is one example where reliability margin and improved performance can be traded off.

Self-Exercise 6.1

An IC part has an operating temperature 10°C above its specification due to mounting in a package with poor thermal impedance. If metal E_a = 0.8 eV and normal use temperature is 85°C, (a) what percentage must the metal current density be reduced to maintain the same expected time to fail (t_F), and (b) if the part is run 10°C hotter, what is the reduction in lifetime.

How Metals Fail in Electromigration. Metals need an imperfection or defect in the structure to begin the failure process. Unfortunately, metals have unavoidable vacancies and irregular grain boundary patterns that can initiate electromigration. The ultimate failure may be an open circuit, or the metal may exert pressure at a site and break the passivation layer, and possibly form a defective bridge. Figure 6.10(a) shows an electromigration failure caused by extrusion of the Al through the passivation material, forming a bridge to another interconnection. Figure 6.10(b) shows a commonly seen open circuit notching characteristic of narrow metal lines (< 0.5 μm).

Failures commonly occur at a flux divergence site, and Figure 6.11 sketches two examples. Figure 6.11(a) shows a region of increased granularity that provides many more paths for Al to move in. The left side of the granular region undergoes tension as atoms leave this location and are easily transported in the granular region. The right border of the granular region undergoes compression as atoms find fewer travel paths and are stuffed into a volume smaller than their natural spacing. The tension and compression regions set up a stress gradient that actually acts to retard the left-to-right motion of Al atoms under electromigration. That is an important failure relief concept that will be developed later.

Figure 6.11(b) shows temperature differences along a line that can lead to flux divergence. The hot region has a higher concentration of vacancies, and Al atoms will diffuse more readily here than in the cold region. The J_{m1} flux tends to cause voiding and tension in and near the hot region, and compression as the moving atoms leave the hot region and approach the cold region. Since the mechanism for motion is diffusion, the cold region

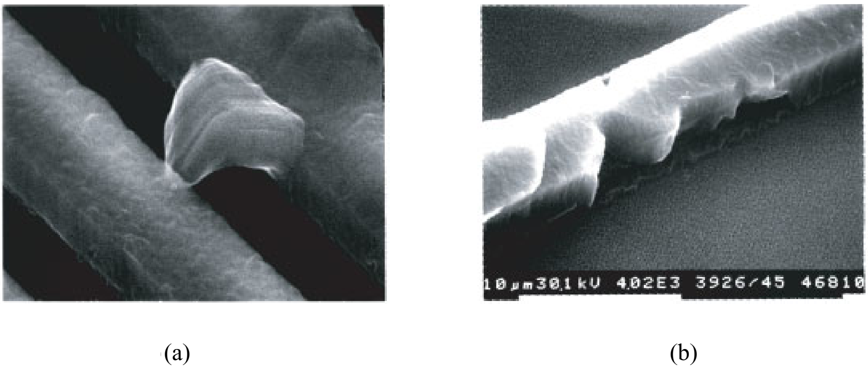


Figure 6.10. (a) Extrusion and bridging defect from EM. (b) Narrow metal line electromigration. (Reproduced by permission of Rod Augur of Philips Semiconductors; reprinted with permission from "Diffusion at the Al/Al oxide Interface during Electromigration in Wide Lines," *J. Appl. Physics, 79,* 6, 15 March 1996, pp. 3,003–3,010. Copyright 1996, American Institute of Physics.)

relatively stops the moving atoms (J_{m2}). The compression can be large enough to rupture the passivation material, allowing extrusion of the metal into the ruptured region [Figure 6.10(a)].

Another serious flux divergence site appeared when W (tungsten) was adapted as a via material in multilayer Al metal ICs. Tungsten has a high melting point, stronger atomic bonds, and high resistance to electromigration. Figure 6.12(a) sketches a metal interconnect with W vias and a TiN (titanium nitride) liner over the top and bottom of the Al line. Electrons entering the structure at the upper right direct Al atoms toward the W via. There is no net motion of atoms in the W, so the Al atoms pile up, causing compression at the top. Since electrons move unimpeded through the W via, they can dislodge Al atoms on the other side of the via, leading to voiding at the bottom (Figure 6.13). The left-side via experiences the opposite effect—compression on the bottom and tension (voiding) at the top.

Figure 6.12(b) shows a metal structure with Al vias. Here, Al atoms will pass through all structures. Unless there are local flux divergence sites along the path, there will not be voiding or compression. W replaced Al as the via material several years ago when reduced scaling made it difficult to retain Al. Cu (copper) interconnections with Cu vias unfortu-

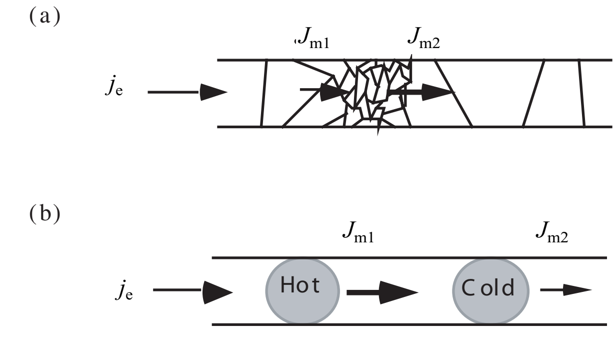


Figure 6.11. Flux divergence sites due to (a) granularity differences, (b) temperature gradient.

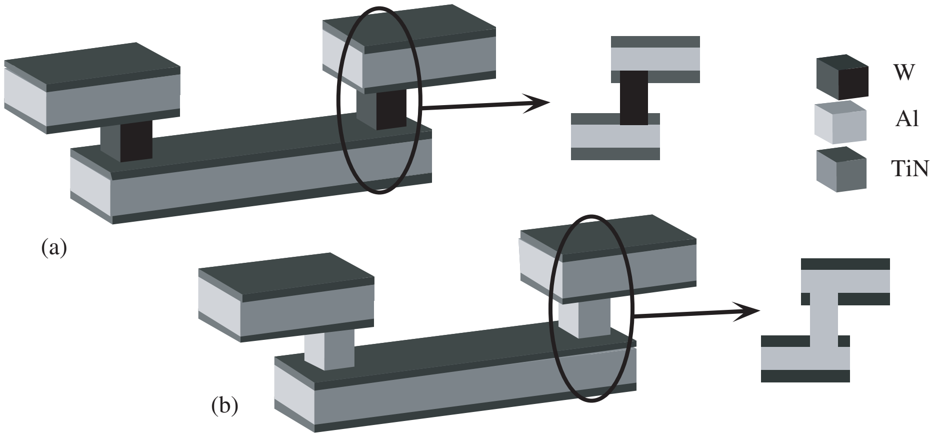


Figure 6.12. Metal–via structures (a) Al–W–Al. (b) Al–Al.

nately have a thin barrier metal interrupting the Cu flow at the bottom of the via. This flux divergence site is a common location for Cu electromigration voiding.

The TiN metal adjacent to the Al is a safeguard against breaks that might occur in the Al (Figure 6.13). If an open circuit forms in the Al, then current will pass through the TiN shunt path. This open-circuit protection does not prevent extrusions that occur horizontally and form bridges. However, barrier metals on the Al are a good retardant to electromigration voids, and one reason why electromigration is not more prevalent. Barrier metals such as TiN safely shunt current around an Al void location, but resistivity of barrier metals can be 30–40 times higher than Al. Voltage drops are typically small, but surprisingly large amounts of heat may be generated. Self-Exercise 6.2 illustrates this.

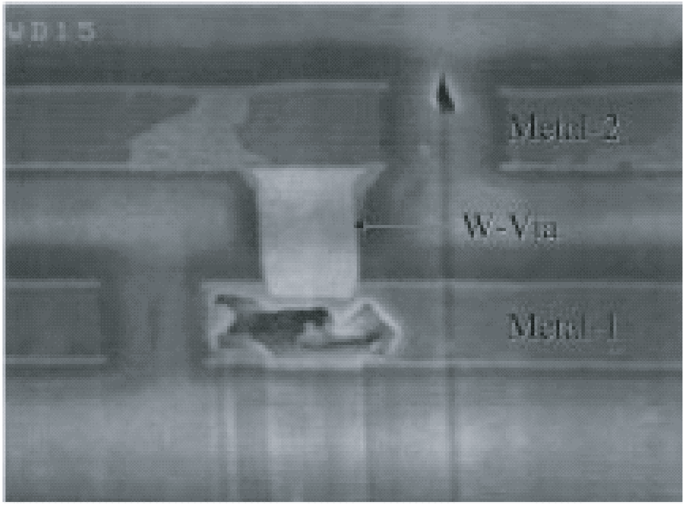


Figure 6.13. Void at the bottom of the blocking tungsten plug in metal-1 aluminum–copper alloy due to accelerated electromigration stressing [7].

Self-Exercise 6.2

The Al void in Figure 6.14 is 2.5 μm long and 1 μm wide, and the resistance of the TiN conduction path is 3 Ω. The current shunts through the 2.5 μm of TiN and then back to the Al. Calculate the heat flux (Watts/cm^2) at the TiN surface, assuming that all heat goes out the bottom surface when: (a) current is 100 μA and (b) current is 10 mA. (c) What are voltage drops across the shunted void? (The problem assumes only one side of the Al is shunted by TiN.)

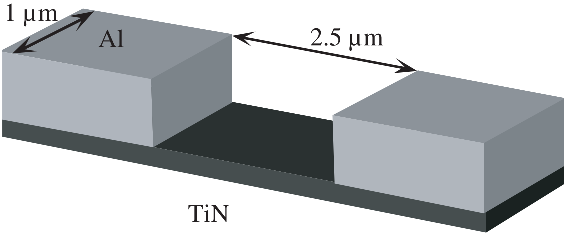


Figure 6.14. Al metal line with a void.

Current Polarity and Pulse Frequency. Three current waveform types exist in metal: pure DC, unipolar pulse currents with an average DC value, and bidirectional (bipolar) currents. DC currents exert the most electromigration stress, and are the typical condition for process characterization and long-term reliability studies. With the exception of off-state leakage, pure DC currents do not appear in CMOS circuits that are fully static and fully complementary. Fully static means that the circuit can operate at zero clock frequency, and fully complementary means that there are equal numbers of *p*MOS and *n*MOS transistors per gate. DC currents may appear if pull-up or pull-down resistors are used, or in any design that allows a continuous path from V_{DD} to V_{SS}. A significant reliability implication is that CMOS DC currents are also caused by most bridge defects and certain open-circuit defects. If these defect-induced currents exceed electromigration design rules in a given interconnect, electromigration may unexpectedly occur.

Unipolar pulses have an average DC current that occurs in drain or source terminals. Positive current always enters the source of a *p*MOS transistor (dotted line in Figure 6.15) and exits the drain, whereas current always enters the drain of an *n*MOS transistor and exits the source. The transistor gate voltage turns source–drain current on and off. The average (DC) current is used as the stress parameter for electromigration.

Bipolar current (solid line in Figure 6.15) occurs in the interconnections from a gate output terminal to the next logic gate input. Current enters the load gate interconnection, charging the capacitance during pull-up. The current reverses direction during pull-down. Interconnects carrying bipolar current have a low susceptibility to electromigration because damage caused on the forward current phase tends to heal on the reverse current phase. Metal atoms move in one direction during pull-up, and then reverse their direction during pull-down.

Blech Effect and Electromigration. The force between atoms at a given temperature is similar to the force between balls attached by a mechanical spring. If you push against a crystal of Al atoms, they compress and exert a back force. That back force acts on the com-

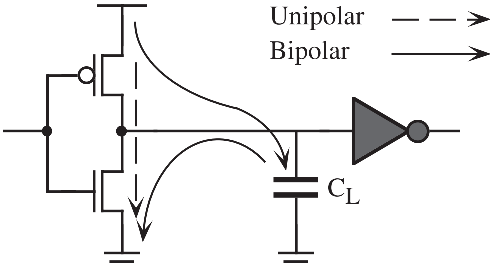


Figure 6.15. Unipolar and bipolar currents.

pressed atoms, tending to eject them from the stressed space. If tension is put on a crystal, then the opposite effect occurs as the region will tend to pull atoms back into the tensile space. Ilan Blech found an interesting and beneficial relation between electromigration and stress gradients [6]. He discovered that for a fixed metal line length, there was a current density below which electromigration would not occur. Conversely, for a given current density, there was a line length below which electromigration failure also would not occur. This so-called Blech effect is important in containing electromigration in deep-submicron ICs. We will derive the Blech effect flux equations as we did earlier for electromigration in Table 6.1.

Stress analysis uses mechanical concepts. We will derive the Blech effect starting with the definition of stress (σ):

$$\sigma = \frac{F}{A} \tag{6.4}$$

where F is the force across a material with area (A). A material or atom with high stress (σ), but no stress gradient ($d\sigma/dx = 0$) has no driving force to move it. However, when a stress gradient is present, then the force on an atom with atomic volume Ω is

$$F_\sigma = \Omega\frac{d\sigma}{dx} \tag{6.5}$$

This expression is subtle, but quite important for understanding stress voiding in metals. The equation predicts that applying equal high stress throughout a material has no displacement influence on particles, molecules, or atoms, but a difference in stress across a material will tend to move an atom. How is this so?

Figure 6.16 shows a unit cube representing a small solid. Initially, the block has no net force on its sides, but when a force F is applied to the right-hand face, the x-dimension compresses by Δx. If the left-hand side is constrained, we can derive the relation between the force, cube dimensions, pressure (p), and induced stress σ.

The energy statement (w) including pressure (p) and volume (v) is

$$dw = F \cdot \Delta x = \Delta p \cdot \Delta v \tag{6.6}$$

or

$$F = \Delta v\frac{\Delta p}{\Delta x} \tag{6.7}$$

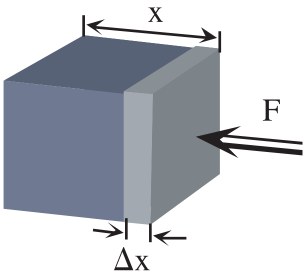


Figure 6.16. Element volume of particle or atom under stress forces.

Since stress $\sigma = p$, and the smallest volume is an atom of volume Ω, then

$$\Delta V = \Omega \text{ (smallest unit volume, approximately an atomic volume)} \tag{6.8}$$

and in the limit

$$F = \Omega \frac{d\sigma}{dx} \tag{6.9}$$

If a stress gradient exists in the metal, then Equation (6.9) provides the force on atoms to move them with the stress gradient. Equation (6.9) allows us to derive the flux equation using an electromigration-induced stress gradient, and then an equation for the Blech observation. The initial flux equations are

$$J_\sigma = vC = (\mu F)C = \frac{D}{kT} FN \tag{6.10}$$

Substituting Equation (6.9) into Equation (6.10) gives the flux term for metal under a stress gradient J_σ:

$$J_\sigma = \frac{D}{kT} \Omega \frac{d\sigma}{dx} N \tag{6.11}$$

The diffusion mechanism is evident in the stress effect. Electromigration simultaneously creates a tensile force and a compressive force as an atom is displaced. Blech recognized that the electromigration force and an oppositely directed stress gradient force may achieve a flux balance, thus stopping the net flow of atoms. Those two electromigration (J_{em}) and stress driven (J_σ) fluxes are repeated:

$$\begin{aligned} J_{\text{em}} &= \frac{D}{kT}(Z^* \cdot q)\rho \cdot j_{\text{e}} \cdot N \\ J_\sigma &= \frac{D}{kT}\Omega \frac{d\sigma}{dx} N \end{aligned} \tag{6.12}$$

At equilibrium, the absolute values of the fluxes are equal, $J_\sigma = J_{\text{em}}$.

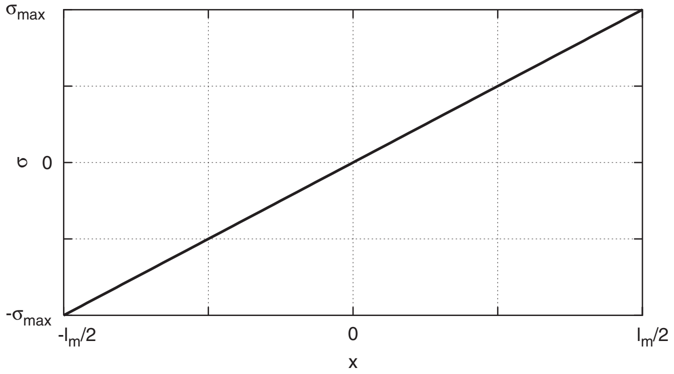


Figure 6.17. Stress versus distance curve for a thin film of metal.

Figure 6.17 shows a stress versus distance curve for a passivated metal. The stress gradient is assumed to be due to metal atom migration from the LHS of the curve to the RHS. $\sigma_{\max}$ is the maximum stress that the passivation can take before it cracks.

We can rearrange Equations (6.12) as

$$j_e dx = \frac{\Omega d\sigma}{Z^* q\rho} \tag{6.13}$$

This balance equation requires one last concept. The balance ends when the passivation ruptures at a distance l_m under the high pressure of a maximum stress $\sigma_{\max}$. If dx is integrated from $x = -l_m/2$ to $+l_m/2$ and stress from $-\sigma_{\max}$ to $+\sigma_{\max}$ we get

$$\int_{-l_m/2}^{l_m/2} j_e dx = \int_{-\sigma_{\max}}^{\sigma_{\max}} \frac{\Omega}{Z^* q\rho} d\sigma \tag{6.14}$$

or

$$l_{\max} j_e = \frac{2\Omega\sigma_{\max}}{Z^* q\rho} \tag{6.15}$$

The labor needed for this derivation is worth the result. The RHS of Equation (6.15) is a constant for a given technology. If we raise j_e, then $l_{\max}$ must drop and vice versa. Equation (6.15) explains Blech's law. Experimental measurements find the $l_{\max} \cdot j_e$ product is about 1,000–3,000 A/cm for unpassivated metal. An example and exercise will illustrate this.

■ EXAMPLE 6.2

An Al interconnect has a length of 100 μm. If the ($l_{\max} \cdot j_e$) product is 3,000 A/cm, what current density limit should be assigned to prevent electromigration?

$$l_{\max} \cdot j_e = 3{,}000 \text{ A/cm} = (100\ \mu\text{m})(j_e)$$

So if $j_e < 3$ mA/μm^2, electromigration will not occur in this unpassivated line. ■

Self-Exercise 6.3

If the ($l_{\max} \cdot j_e$) product is 3,000 A/cm, what maximum length should the IC interconnect lengths be if designers keep effective current densities in all lines at less than 1.2 mA/μm^2.

Electromigration and High Frequency. Lines carrying pulsed currents show more resistance to electromigration as the frequency increases. Figure 6.18 shows the increase in t_f for a 50% duty cycle as current pulse frequency increases from DC [16]. During the off-portion of the cycle, a back stress is exerted on the metal from atoms moved during the on-cycle. As pulse frequency increases, average line temperature rises, increasing back-diffusion efficiency and healing during the off-state. This is an unusual case of metal reliability increasing with temperature.

6.3.2 Metal Stress Voiding

A bad "discovery" was made in the early 1980s when certain metal lines pulled apart, forming open circuits even if the IC was not powered. This failure mechanism, called *stress voiding,* (or *stress-induced voiding*) was linked to the TCE (thermal coefficient of expansion) differences of metal and the passivation materials surrounding it. When deposited metal is taken to 400°C or higher for a passivation step, the metal expands and tightly bonds to the passivation material. When cooled to room temperature, enormous tensile stresses arise from the differences in the TCE of the metal and passivation material. The passivation material basically does not move, so the metal bonded to the passivation material undergoes extreme tensile stress. These stresses are parallel to the metal line, and can pull lines apart if stress gradients appear. The time required to do this varies with the quality of the metal. It can happen during the fabrication process itself, or can take weeks to years for voiding to appear. Figure 6.19 shows photographs of two stress void failures.

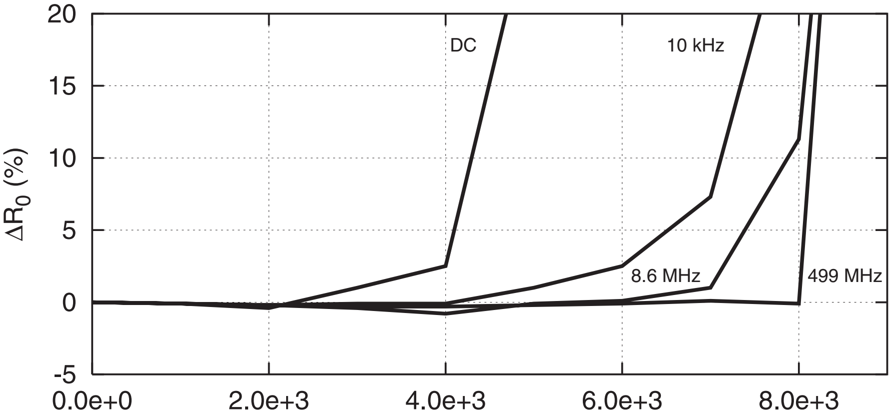


Figure 6.18. Time to failure versus current pulse frequency [16].

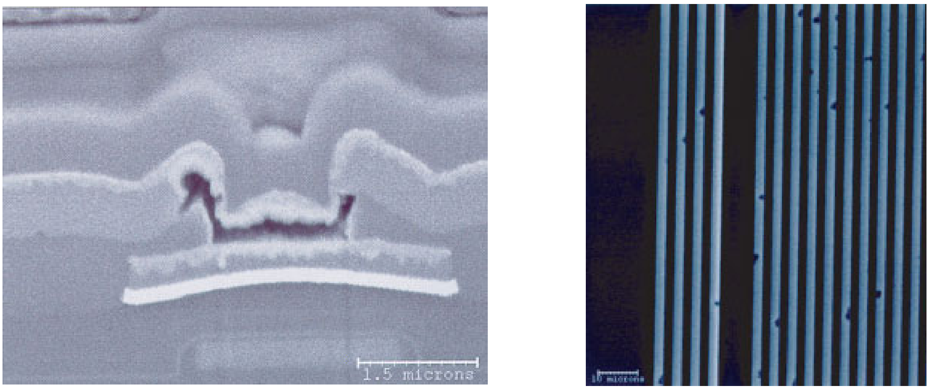


Figure 6.19. Stress void photos. (Reproduced by permission of Bill Filter, Sandia National Labs.)

Stress void analysis uses concepts from materials science, physics, and mechanical engineering. It is a mechanical failure mechanism that involves no electron current, but often it is the most prevalent metal failure mode in modern ICs. We will examine unpassivated and passivated metal line responses to the large mechanical forces that they undergo in normal IC environments. Passivation layers provide electrical isolation and protection to the metals, but the unavoidable problem lies in the sharp differences of thermal expansion coefficients. Al has a TCE = $\alpha = 23.6 \times 10^{-6}$ parts per °C, and silicon dioxide has an $\alpha = 0.5 \times 10^{-6}$ parts per °C. This means that a 1 meter Al line will expand by 24 μm for each degree of elevation in temperature, and SiO_2 by 0.5 μm per degree.

When SiO_2 is deposited and tightly bonded to Al at 400°C, there are no thermally generated stresses between the materials. However, when the two bonded materials cool to room temperature, enormous lateral stresses are generated, since the SiO_2 moves little while the Al strains to contract. A related problem is when Al reacts with a metal such as Ti. If both materials are passivated before the reaction, then the formation of $TiAl_3$ occurs with an approximate volume reduction of 5%. This is another high-stress-generating mechanism. We will calculate the stresses on Al for an unpassivated and a passivated metal line, developing simple equations that predict the stresses and strains in modern IC metals. We will then use numerical examples to show these enormous values, and conclude with methods to reduce the probability of stress void occurrence.

We begin with a simple example of metal stress forces, using a metal rod suspended in air and pulled at its ends. Figure 6.20(a) shows a metal rod with an applied force F, an area A, and a stress $\sigma = F/A$. When F is applied, the rod has a uniform stress along its axial length, and stretches a small amount, ΔL. The amount of stretching is called strain ε, and its measurement is normalized with respect to its original unstressed length L, or $\varepsilon = \Delta L/L$ [Figure 6.20(b)]. When metal is pulled in air, lengthening corresponds to a decrease in the diameter or circumference of the metal. The length increase is compensated for by a diameter reduction in the lateral walls, so that the volume remains constant. The surface atoms have no restraining force to prevent them from moving inward. Accurate strain measurements take this area change into account.

Figure 6.20(b) shows a generic stress–strain curve. Material in the linear region can be stretched to a point that upon release of the force returns the material to its original length.

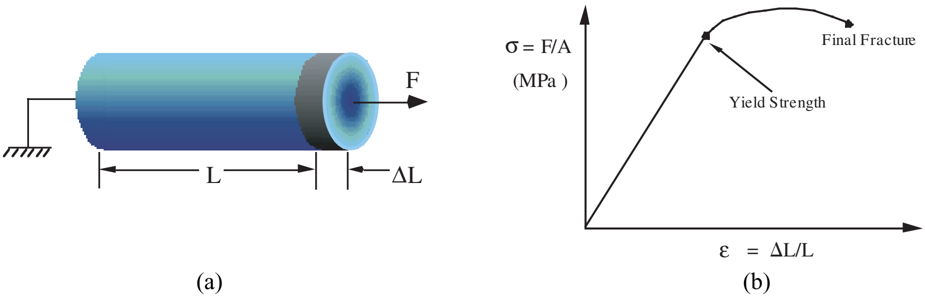


Figure 6.20. (a) Measurement setup for stress and strain. (b) Stress (σ) versus strain (ε) curve.

Atoms in this region are acting in accordance with the atomic spring model. The slope of this straight line is called Young's modulus, Y_m. Y_m characterizes a material's resistance to an applied force, and also allows calculations of strain given the stress (or vice versa). Stress is related to strain by a simple but important relation:

$$\sigma = Y_m \varepsilon \tag{6.16}$$

The stress point at which the material enters the nonlinear or plastic deformation region is called the yield strength of the material. Metal stretched beyond the yield strength will not return to its original length. It has undergone plastic deformation, after which atoms no longer act as springs, but slide past each other. Finally, the material ruptures. The pressure unit is Pascals (Pa) (Newtons per meter^2), where 1 MPa $\approx$ 146 lb/in^2. Al alloys have a yield strength of about 95 MPa (about 14,000 psi) and a Young´s modulus of about 71.5 GPa.

Temperature and length are typically linear, and related by the thermal coefficient of expansion (TCE) where

$$\alpha = \frac{\Delta L/L}{\Delta T} = \frac{\varepsilon}{\Delta T} \tag{6.17}$$

Stress, strain, and temperature are combined from these equations to give

$$\sigma = Y_m \varepsilon = Y_m \alpha \Delta T \tag{6.18}$$

Equation (6.18) is an important link between stress, temperature, and the material constants Y_m and α. An example will show how to estimate the stress forces on an unpassivated Al line.

■ EXAMPLE 6.3

Given an unpassivated Al line surrounded by air [except at the ends, where forces are applied (Figure 6.21)]. Let the length at 430°C be $L + \Delta L$, and the length at 30°C be L. When the metal shrinks as temperature drops from T = 430°C to 30°C, what stress is required to hold the ends at the T = 430°C dimension ($L + \Delta L$). Assume $\alpha_{\mathrm{Al}} = 23.6 \times 10^{-6}$/°C and $Y_m = 71.5$ GPa.

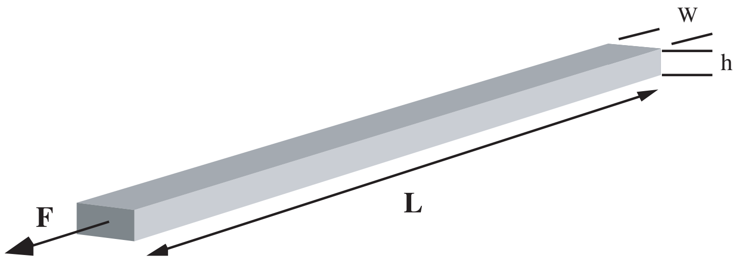


Figure 6.21. Unpassivated Al line.

The fractional change in the Al line is equal to the strain ε, where

$$\varepsilon = \frac{\Delta L}{L} = \alpha \Delta T = 23.6\ 10^{-6} \times (430 - 30) = 0.00944$$

Since $Y_m = 71.5$ GPa and

$$Y_m = \frac{\sigma}{\varepsilon}$$

then

$$\sigma = \varepsilon\ Y_m = 0.00944 \times 71.5 \text{ GPa} = 675 \text{ MPa}$$

■

This 675 MPa stress greatly exceeds the yield and fracture strength of Al, therefore the TCE forces would tear the Al line apart. Most Al metal interconnections do not pull apart, so what protects the thin metal lines? The answer lies in the unique properties that passivation brings to these material systems. Surprisingly, even higher stress forces are generated in passivated metals than for the case of Al in air.

Self-Exercise 6.4

Assume a 100 μm long unpassivated Al metal line (Figure 6.22) with width of 1 μm, height of 0.4 μm, and yield strength of 95 MPa.

(a) How many pounds of force F are required to cause the metal to enter the plastic deformation region? How many kg?

(b) How far can you pull the 100 μm line before it goes into plastic deformation?

We next look at the more relevant situation in which the metal is bound and constrained by a dielectric. We are indebted to Bill Filter of Sandia National Labs for his stress void lectures at the University of New Mexico, giving his insights and examples of stress void calculations. Unpassivated metal atoms on the surface of the line do not have a bond to another atom on the surface plane, since there is only air at the surface. As the metal line is pulled at the ends, all atoms feel a tension, but the surface atoms, having no

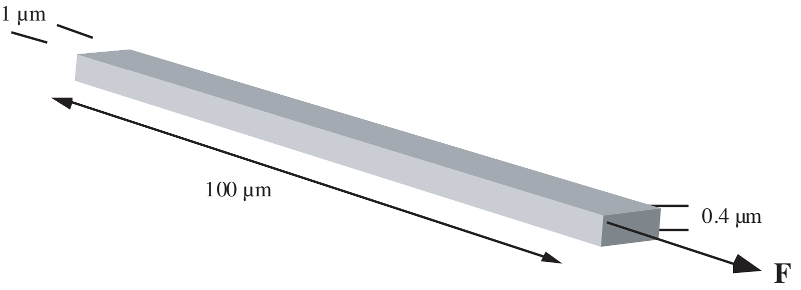


Figure 6.22. Al line.

restraint at the surface plane, tend to move inward, reducing the section of the metal line. The inward displacement provides some stress relief, but the stress remains high.

When a metal line is passivated, its surface atoms bond strongly to the fairly rigid passivation material and, basically, these metal atoms cannot move, even under very large stresses. The stresses are larger when a metal is passivated, and unless there are defects present in the metal, it maintains its shape. A key issue is that the surface atoms that could move inward under tensile stress now remain in position, and a stress exists at the surface of the metal that is orthogonal to the stress applied at the ends of the line. The stress equation is modified from that of Figure 6.20 to reflect this increase, using a parameter called Poisson's ratio. Equation (6.19) shows this modification for passivated metal stress [18]:

$$\sigma = \frac{Y_m \varepsilon}{1 - 2v} \tag{6.19}$$

where v is called Poisson's ratio; $v = 0.35$ for aluminum. When two materials are bonded, the thermally induced strain must consider the differences in their TCE, so that using Equations (6.16)–(6.18),

$$\varepsilon = \frac{\Delta L_2 - \Delta L_1}{L_2} = \frac{L\alpha_2 \Delta T - L\alpha 1_1 \Delta T}{L} = \Delta\alpha\ \Delta T \tag{6.20}$$

where $L_1 = L_2 = L$ for constrained metal. When we substitute Equation (6.19) into Equation (6.20) we get

$$\sigma = \frac{\Delta\alpha \Delta T Y_m}{1 - 2v} \tag{6.21}$$

A simplified example shows the even larger stress calculation for a passivated metal.

■ EXAMPLE 6.5

Given a passivated Al line [Figure 6.23)], if $\alpha_{\mathrm{Al}} = 23.5 \times 10^{-6}$/°C, $\alpha_{\mathrm{SiO}_2} = 0.5 \times 10^{-6}$/°C, $v = 0.35$, and $Y_m = 71.5$ GPa, calculate the stress on the Al at its ends after the material drops from 430°C to 30°C. In other words, what is the required

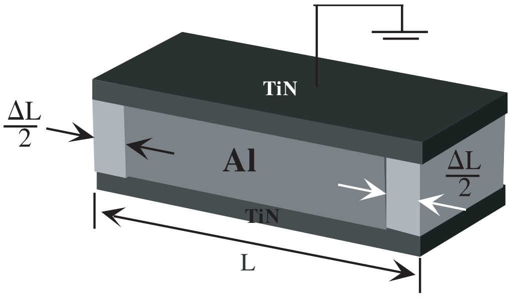


Figure 6.23. Passivated Al.

stress on the ends needed to maintain the longitudinal dimension when the metal and passivation material cool to 30 °C?

The fractional change in the Al line is equal to the strain ε, where

$$\varepsilon = \Delta\alpha\ \Delta T = (23.6 \times 10^{-6} - 0.5 \times 10^{-6})(430 - 30) = 0.00924$$

Since

$$\sigma = \frac{\varepsilon Y_m}{1 - 2v}$$

then

$$\sigma = \frac{(0.00924)(71.5\ \text{Gpa})}{1 - 2 \times 0.35} = 2{,}202\ \text{MPa}$$

■

The stresses in this calculations are enormous, far exceeding the Al yield and fracture strengths. Again, why doesn't Al instantly pull apart? The answer lies in the ability of the passivation molecules to bond to the metal atoms. They can withstand stresses of GPa's, whereas Al–Al bonds permanently deform at less than a hundred MPa. Al is a soft metal.

How does this impact an integrated circuit? Stresses of hundreds of MPa have been measured in passivated Al lines. The everyday ICs that we use, such as wristwatch ICs, and those in personal computers, pocket calculators, etc., are subject to these large stresses throughout their product life. What is the concern? It is the threat of stress voiding, and that is final subject of this section.

How do stress voids appear in this high-stress environment? Voids in the metal do not just appear spontaneously. Voids form only in a stress gradient. Equation (6.22) is the flux equation for materials under a stress gradient:

$$J_\sigma = \frac{D}{kT}\Omega\frac{d\sigma}{dx}N \tag{6.22}$$

Net atomic displacement at room or operational temperature can only happen if $d\sigma/dx$ is not zero. If the *whole* material has an equal stress of 2.2 MPa, then net atomic displacement does not happen. Metals and their passivation material can lie in these large stresses, and nothing much happens unless a gradient occurs. That is what causes stress voids.

A void requires an imperfection in the metal, such as a small nucleation at the metal–passivation material boundary. The surface of the metal atoms of a void nucleus have zero stress, so that a large stress gradient forms in the metal. The metal atoms now want to move and relieve the stress gradient. The last condition for destructive stress void growth is a diffusion path and sufficient temperature for these metal atoms to move. A neighboring grain boundary unfortunately satisfies this condition. In summary, there are three conditions for a stress void to occur:

1. A large stress must be present in the metal.
2. A defect must be present to convert the stress to a stress gradient.
3. A diffusion path and sufficient temperature must be present that allow the void to grow.

The flux J_σ in Equation (6.22) is proportional to the stress gradient, but atomic travel is still controlled by the diffusivity of the metal. These two factors combine to give an interesting temperature dependence to stress void sensitivity. Figure 6.24 shows a family of acceleration factor curves relative to room temperature as a function of processing temperature. The top curve has a deposition temperature of 435°C and the bottom curve of 300°C. Each curve has a peak with zero acceleration factors at high temperatures and near zero at low temperatures. At high temperatures, the TCE-induced stress differences are small, so little net metal atomic motion occurs, even though diffusivity is large. At low ambient temperatures, the TCE-induced stress is very large, but diffusivity is small, and, again, little net motion of the metal occurs. A peak occurs at which diffusivity and stress gradients

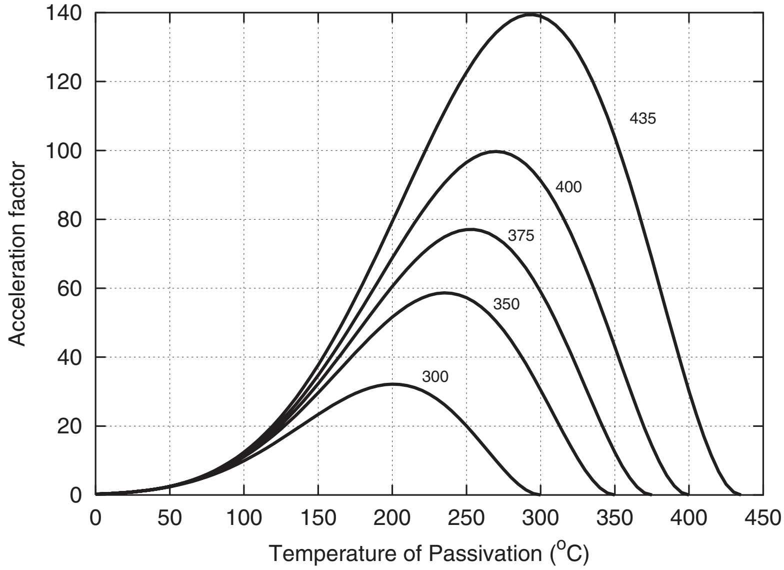


Figure 6.24. Stress void sensitivity versus test or ambient temperature. (Reproduced by permission of Bill Filter, Sandia National Labs.)

combine for maximum effect. The peak sensitivity shifts at lower process temperature, and the magnitude of the peak sensitivity is lower. A warning from Figure 6.24 is that modern high-performance ICs have die temperatures above 100°C.

Stress Voiding and Electromigration Comparisons. Electromigration and stress voiding have distinctions that are important when seeking the root cause of a failure (Table 6.2). Electromigration requires circuit operation that provides the necessary current density and elevated temperature, and it worsens with temperature. Stress voiding needs only a moderate temperature; it has a temperature peak in sensitivity, and it requires no power. When small stress voids nucleate during the cool-down or quenching process, then lines become more sensitive to electromigration during circuit operation. The small stress voids finitely reduce the metal area, thus increasing j_e.

6.3.3 Copper Interconnect Reliability

Cu appeared in 1998 as a substitute for Al in some high-performance ICs. Cu resistivity is about 1.9 $\mu\Omega \cdot$ cm compared to Al–Cu alloy resistivity of about 3 $\mu\Omega \cdot$ cm. This is about a 37% reduction in interconnect resistivity and RC time constant. Cu has interesting properties besides resistivity. Its melting point is 1358 K, considerably higher than Al at 933 K. This means that Cu has stronger bonds than Al, and should be more resistant to atomic motion by electromigration or stress voiding. Lee et al. reported a Cu Young's modulus of 110 GPa compared to 70 GPa for Al [10]. They also reported a Blech threshold of 3700 A/cm for passivated Cu lines.

It was originally thought that Cu would never electromigrate or have stress void failures, but that is not the case. Cu electromigration activation energies can show $E_a \approx 0.8$ eV, about the same as bamboo Al. Why the similarity if Cu–Cu bonds are stronger? Metal

Table 6.2. Comparison of Electromigration and Stress Voiding Failure Mechanisms

Property	Electromigration Failure	Stress Voiding Failure
Powered On	Necessary	Not necessary
Defect Class	Bridges and opens	Only opens
Ambient Temperature	Gets exponentially worse as temperature rises	Has a temperature peak sensitivity to failure
Passivation Strength	Blech Length gets longer as passivation hardness increases, making metal more durable	TCE-induced stress gets worse as hardness increases; thus, more vulnerable to failure
Initial Event	Displaced atoms cause tension and compression regions; stress gradient formed from displaced atoms	Stress gradient appears due to void nucleation before metal atoms are displaced
Atomic Displacement Mechanism	Diffusion	Diffusion
Final Failure Mode	Displaced atoms cause voiding (opens) and compression (bridging)	Displaced atoms cause voiding

migration in Al bamboo structures primarily takes place at the Al_2O_3–passivation interface. Cu electromigration also takes place at the Cu–passivation interface, and Cu granularity is higher than Al. Al bonds much tighter to the passivation than does Cu, so that Cu migration at its interface is larger. Improvements in Cu processing have raised its measured activation energies.

The Cu advantage in RC time constant can be taken from the R or C. For example, if a Cu and Al metal line have the same geometry, then the Cu line resistance is about 37% lower than the Al with the C unchanged. However, if the height of the Cu metal line is reduced, then its resistance increases, but its sidewall capacitance drops. The load C_L and crosstalk capacitance are smaller for Cu, and the IC power dissipation drops, as we saw in Chapter 4:

$$P = C_L V_{DD}^2 f_{clk} \tag{6.23}$$

Cu has a unique reliability risk not found in Al and that is the high diffusivity of Cu in SiO_2 and Si. Cu lines must be bound with a thin ($\approx$ 150 Å) barrier metal liner such as TaN. Cu can ruin transistor *pn* junctions if it is not contained. TaN liners bind three surfaces of the Cu line: the bottom and the two sides. A partial containment solution uses W (tungsten) at the first metal layer that connects drain, source, and gate contacts. This further separates Cu from the transistor *pn* junctions. Although W resistivity is higher, the signal lines are kept short so that IC performance is not compromised.

A Cu process uses the dual-damascene process, which is quite different from the Al metal sputtering process. The metal regions are first etched as trenches in the dielectric. Then a thin barrier metal layer is deposited on the bottom and two sidewalls of the trench. Next, a thin Cu seed layer is deposited in the trenches using PVD or CVP techniques followed by a Cu electroplating that fills the trenches and upper dielectric surface. The surface is then polished flat using the chemical mechanical planarizing (CMP) technique. This removes the excess Cu on the top surface, leaving Cu interconnects in the trenches.

The quality of the Cu interconnects depend critically on the properties of the seed layer. A rough Cu seed layer promotes small grain sizes, and this degrades the ability of Cu to resist EM fluxes. The dual-damascene process is more complex, with its required barrier protection metal liner as an essential part of Cu interconnect integrity. Cu is also a strong contaminant of the ICs in a fabrication facility if it spreads to the equipment in the lab. Originally, Cu was thought to be a perfect metal, without electromigration or stress void reliability risk. However, studies show otherwise [8, 10, 13, 15].

Song et al. noted that Cu is a noble metal that is less reactive than Al [15]. This may seem to be an advantage, but Cu forms weaker bonds with the surrounding dielectric than Al. Therefore, we see higher EM fluxes occurring at the Cu-to-dielectric interface, and even see lateral (intrametal layer) breakdown leakage paths.

The high via aspect ratios make liner dimension and liner continuity integrity challenging for the billions of vias that may populate an IC. The total metal length of a modern IC is on the order of several kilometers. The initial via etch through the dielectric must be taken just to the level of the bottom metal and no more. A shallow etch will leave a thin layer of dielectric in the serial path of the via. A 90 nm technology can use minimum vias on the order of 90 nm in diameter, subject to via–metal interconnect line design rules. The via fabrication challenges translate to lower IC yield, more test escapes, and increased reliability concerns.

Many via electromigration failures occur at the bottom of the vias, where the liner intersects and interrupts the Cu via path [8]. The subsequent Cu flux divergence identifies an electromigration weak spot. This site is similar to the electromigration voiding found under a tungsten via in Al systems. Also, an initial defect-induced voiding in the Cu could electromigrate to a complete open in the via. This forces the via current to pass through the parallel path of the liner. Although this offers protection, failure analysis showed that about 20% total via voiding could lead to excessive heat in the high-resistive liner, and even lead to a thermal opening of the liner itself.

The evolution of Cu interconnects to low-k dielectrics will impact Cu reliability. Lee et al. studied Cu with the low-k dielectric SiLK™ and found that t_{50} values were 3 to 5 times lower for Cu–SiLK™ than for Cu–oxide materials [10].

Doong et al. reported design rules for stress-induced voiding (SIV) in a 130 nm Cu–damascene technology [19]. The interconnect variables were width of the metal lead, via location in a wide metal lead, and width of the other connecting metal. They found that SIV was more severe on a via fed by a wide metal line than on one fed by a narrow line. Design rules are a key ingredient to preventing stress voiding in an IC.

6.4 OXIDE FAILURE MODES

Transistor gate oxides made of SiO_2 (silicon dioxide) are the beating hearts of a MOS transistor. Gate control of channel charge depends on the dimensions and quality of this oxide. Although SiO_2 appears in different parts of an IC, this section specifically uses the word oxide to refer to the thin dielectric material that separates the transistor gate from the channel substrate. Financial penalties for poor quality oxides are longer time to market and customer dissatisfaction. Oxide thickness in the 1970s was about 750 Å, and now oxide dimensions are below 20 Å. Gate oxide electric fields at the turn of this century were higher than burn-in field strengths in the early 1990s. Test, field failure, and burn-in are just three examples of why we must understand the chemical and electronic nature of oxides. We will look at the chemical structures of the thin oxide and then two oxide failure mechanisms: wearout and hot carrier injection. We will close with a description of a relatively recent *p*MOS transistor oxide reliability concern called negative bias temperature instability (NBTI).

Figure 6.25 is a remarkable TEM (transmission electron microscope) photograph of a MOS capacitor structure, showing the atoms of the single-crystal Si material, the noncrystalline or amorphous SiO_2 thin oxide molecules, and the polysilicon gate material above [24]. Imperfections cannot be avoided when the amorphous SiO_2 surface abuts the Si crystal. The interface is the site of numerous dangling bonds in which Si atoms or SiO_2 molecules have unshared bonds, leading to charges that are readily filled if an electron, hole, or hydrogen atom (H^+) is near. Process steps use or generate hydrogen and water that can bond with the unfulfilled states at the interface. The transistor threshold V_t is altered by these charge exchanges, with an important impact on speed.

Figure 6.26 shows the molecular orientation of SiO_2 molecules. A Si atom (open circle) appears to bond to four O atoms (shaded circles), but since each O atom also bonds to another Si atom, the chemical ratio is one Si to two O atoms, or SiO_2. McPherson and Mogul described the oxide structure in which each SiO_2 tetrahedron molecule forms rigid 109° angle bonds between Si and O [41]. Significantly the bonding between tetrahedrons is not rigid, but bond angles form from about 120° to 180°, with a mean of about 150°.

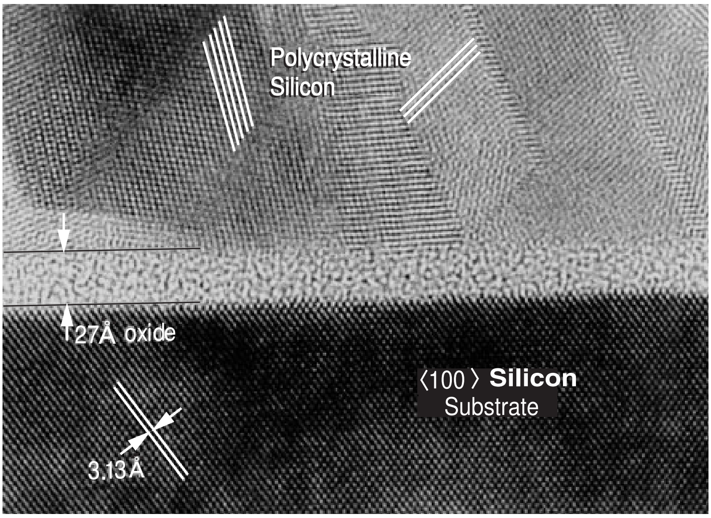


Figure 6.25. MOS capacitor cross section. (Reproduced by permission of Doug Buchannan, IBM Corporation.)

The bond angle weakens as the angle deviates from the mean. The variable bond strength is one source of the statistical behavior of oxide wearout and breakdown. Another weak bond occurs when an O atom is absent, allowing two Si atoms to bond to each other (Figure 6.26). These weaker (strained) bonds are more susceptible to rupture, leaving sites for holes, electrons, or atoms such as hydrogen to attach to.

The dangling bonds, the variable bond strength of SiO_2–SiO_2 molecular angles, and the absence of O atoms in the normal pairing leads to defects in the oxide called *traps*. A trap is an oxide defect, and the electronic charge on that trap is called a *state.* Traps can

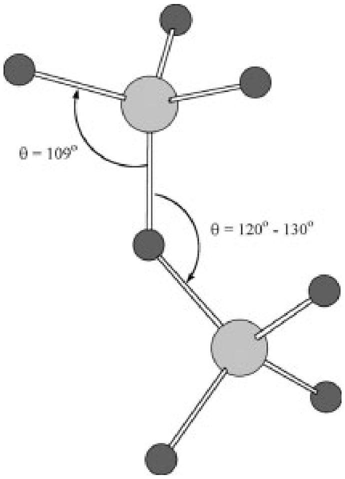


Figure 6.26. Si and O bond geometry in SiO_2 [41].

exist after the processing steps, or can be created when bonds are broken by energetic particles such as electrons, holes, or radiation.

Traps lying at the Si–SiO_2 border are called *interface traps.* Interface traps can rapidly exchange charge with channel carriers, since they are in close proximity to the channel. A trap 25 Å into the oxide will exchange channel charge in about one second. The oxide depth of the trap from the interface determines the exchange rate with the channel. Each trap that is 2.5 Å deeper in the oxide increases charge tunneling time by about one decade [30]. Border traps are those that lie deeper than interface traps, but less than 50 Å deep. Fixed oxide traps lie deeper than 50 Å and, basically, do not exchange charge with the channel. Most oxide dimensions are now less than 50 Å, so that these deep traps are less relevant to modern failure mechanisms. Charge exchange between the channel carriers and the oxide traps has a negative influence on transistor performance. The next section builds on these physical descriptions, describing oxide wearout and subsequent rupture.

6.4.1 Oxide Wearout

Good oxides wear out and rupture if a charge is continuously injected. This has nothing to do with defects from the fabrication, and the actual failure mechanism has eluded good people doing expensive experiments for many years. Each time a logic circuit has a voltage put across its gate oxide, a small amount of charge is injected into the oxide. The question is how long will it take for a normally operating transistor to wear out and rupture. That time must exceed expected product lifetime, since miscalculation could have severe consequences if premature oxide wearout caused ICs to fail during product life. Oxide wearout time decreases as oxide stress increases, and a concern is that the voltages and electric fields of thin oxides will cause premature oxide wearout. Oxide field strength is the force that accelerates electrons across the oxide. The example below illustrates the rising oxide field occurring for deep-submicron transistors in their use condition. The values are typical for the succession of technologies since the late 1980s.

■ **EXAMPLE 6.6.**

Calculate the oxide field strengths in V/cm for the following technologies: (1) 5 V and T_{ox} = 300 Å, (2) 5 V and T_{ox} = 200 Å, (3) 3.3 V and T_{ox} = 100 Å, (4) 2.8 V and T_{ox} = 60 Å, (5) 2.5 V and T_{ox} = 40 Å, and (6) 1.2 V and T_{ox} = 20 Å.

1 Å = 10^8 cm, so:

1. $\mathscr{E}_{ox} = 5\text{ V}/300 \times 10^{-8}\text{ cm}= 1.7\text{ MV/cm}$
2. $\mathscr{E}_{ox} = 5\text{ V}/200 \times 10^{-8}\text{ cm} = 2.5\text{ MV/cm}$
3. $\mathscr{E}_{ox} = 3.3\text{ V}/100 \times 10^{-8}\text{ cm} = 3.3\text{ MV/cm}$
4. $\mathscr{E}_{ox} = 2.8\text{ V}/60 \times 10^{-8}\text{ cm} = 4.7\text{ MV/cm}$
5. $\mathscr{E}_{ox} = 2.2\text{ V}/40 \times 10^{-8}\text{ cm} = 5.5\text{ MV/cm}$
6. $\mathscr{E}_{ox} = 1.2\text{ V}/20 \times 10^{-8}\text{ cm} = 6\text{ MV/cm}$

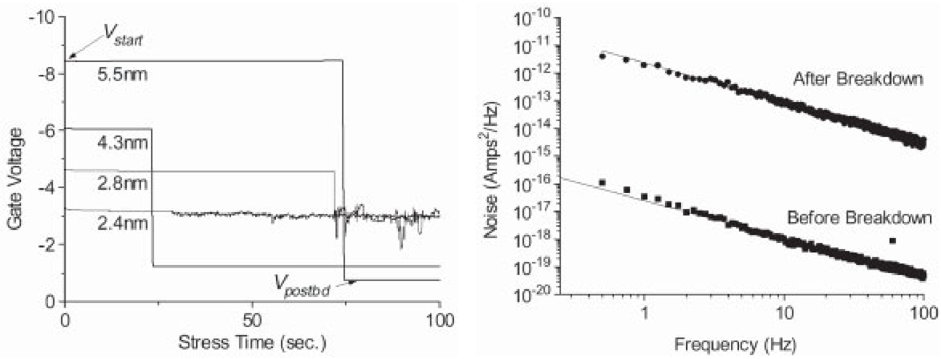
■

Significant tunneling of electrons through the gate oxide can occur when the oxide thickness becomes less than about 40 Å. As T_{ox} goes to 20 Å and 15 Å, the tunneling current is worse. These increased gate currents are reliability and power concerns in modern ICs.

Most of the research on oxide reliability has used MOS capacitor structures. Some early work on transistor gate oxide shorts showed that the gate capacitance could store sufficient energy (½ CV^2) so that when a breakdown rupture occurred, this energy was released into the small, weakened oxide site, causing severe local damage. The silicon on either side of the oxide became temporally molten and joined; i.e., the polysilicon gate material physically bonded to the silicon substrate. An *n*-doped polysilicon gate joined with the *p*-well (*n*MOSFET) or *n*-well (*p*MOSFET) to form parasitic diodes or resistors. As transistors were scaled to modern technologies, power supply voltages dropped from 5–10 V to 1.0–1.2 V. Gate dimensions scaled from channel lengths of 1–5 μm to 90–130 nm. The gate area scaled by $(0.7)^2$ for each technology node so that gate capacitance scaled on the order of 2^7 as we went from 1.0 μm to 130 nm technologies. This dropped the gate capacitance by a factor of over 100 and V_{DD}^2 by about 20–25. The stored gate capacitance then became sufficiently small so that the violent thermal ruptures were replaced with the more gradual and subtle breakdowns that are described next.

What causes oxide wearout? The answer lies in which technology we work with. The older-technology oxides greater than 40 Å thick have a breakdown model quite different than the oxides we now build (below 30 Å). Oxides less than 30 Å thick are known as the ultrathins. Ultrathin oxide breakdown shows a distinct soft breakdown. Soft breakdown results in an irreversible damage to the oxide. Its most significant effect is an increase in noise of the gate voltage or current. Figure 6.27(a) shows this breakdown for oxide thicknesses from 2.4 nm to 5.5 nm [52]. The oxides were stressed with a constant current and the 5.5 nm oxide shows a precipitous drop in gate voltage at 75 s when a stressing gate current is applied. The 2.4 nm gate oxide did not change gate voltage with the oxide damage event, but shows an increase in noise.

The noise plotted in Figure 6.27(b) shows a four orders of magnitude increase after soft breakdown. Noise increase is the only certain evidence of the irreversible damage to ultrathin oxides. The noise associated with soft breakdown is thought to be trap-assisted conduction through a small conducting path in the oxide. The electrons hop noisily from trap to trap. In contrast, hard breakdowns in the thick oxides of older technologies showed severe gate voltage or current changes (Figure 6.27(a)). A hard breakdown was defined as a thermal event that merged the material above and below the oxide. The physical touch-



(a) (b)

Figure 6.27. (a) Oxide breakdown with stress time. (b) 1/*f* noise before and after breakdown [52].

ing of two differently doped materials can create diodes, or resistors if the doping is of opposite polarity.

The normal functioning of a transistor with an ultrathin oxide is not as effected as those with thicker oxides following rupture [49]. The ultrathin wearout and breakdown model shows that rupture is primarily related to the gate voltage V_G and the amount of charge driven through the oxide (fluence). Evidence for the voltage model is shown in Figure 6.28, which plots the log of time to breakdown (T_{BD}) versus V_G. The interpretation is that electrons tunnel through the gate oxide, accelerating in the oxide field. The oxide electric field is constant across the oxide, but the internal oxide voltage drops as the electron reaches the anode of the structure. The relation of oxide rupture to gate voltage implies that the electron travels through the oxide without interaction, achieving a maximum kinetic energy before striking the anode, where it causes bond breakage. The likely weak bonds are H–Si and H–O. One subsequent damage mechanism is believed to be release of a hydrogen ion that reenters the oxide, causing trap damage. This is the anode hydrogen release model (AHR). The other damage mechanism is thought to be creation of a hole that migrates back into the oxide. This is the anode hole injection model (AHI). H^+ and a hole feel the attractive pull of the oxide electric field, causing trap damage as they enter and interact with the oxide molecules. There is evidence that both AHI and AHR contribute to wearout and breakdown [37].

Oxides do not breakdown after a single hole or electron are reinjected into the oxide. Oxides have a wearout and a breakdown phase. The wearout is believed to be the continuous addition of damage sites (traps) distributed throughout the oxide. When a statistical distribution of these traps is critically aligned in a vertical path supporting an increase in conduction, then a thermally damaging current goes through the oxide. This model is known as the *percolation model* of wearout and breakdown [28]. Figure 6.29 sketches such a statistical distribution of traps. The path in the middle of the figure indicates a trap distribution that is sufficiently close to form a breakdown percolation path.

Ultrathin breakdown has been characterized into three stages:

1. Slow defect generation within the oxide (wearout) until a defect path links the gate terminal to the substrate (percolation model)

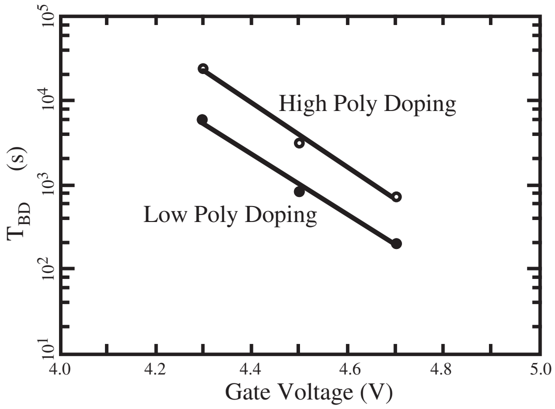


Figure 6.28. TBD as function of V_G [37].

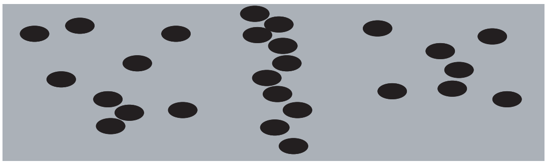


Figure 6.29. Percolation model of wearout and breakdown [28].

2. A soft breakdown (SBD) at low voltages that permanently increases gate current (< 100 nA at 1.2 V in 150 Å oxide) and gate noise
3. The appearance of a "hard breakdown" (HBD), showing continuous exponential increase in I_G.

There is evidence that SBD and HBD may be independent events [40, 44].

An ultrathin-oxide, voltage-dependent time-to-breakdown model (T_{bd}) has been proposed [43]. This breakdown model [Equation (6.24)] includes the gate oxide thickness (T_{ox}) and the gate voltage (V_G):

$$T_{bd} = T_0 \cdot e^{\gamma\left(\alpha \cdot T_{ox} + \frac{E_a}{kT_j} - V_G\right)} \tag{6.24}$$

where γ is the acceleration factor, E_a is the activation energy, α is the oxide thickness acceleration factor, T_0 is a constant for a given technology, and T_j is the average junction temperature. Time-to-breakdown physical parameter values were extracted from experiments as follows: $(\gamma \cdot \alpha) = 2.0$ 1/Å, $\gamma = 12.5$ 1/V, and $(\gamma \cdot E_a) = 575$ meV [43]. The voltage model and its supporting data suggest that ultrathin oxide rupture will be a greater concern with the increased electron tunneling (fluence) of thin oxides. V_{GB} decreases with each shrinking technology, but it is still high enough to support electron tunneling and subsequent reentry of high-energy particles into the oxides. However, recent data suggest that the soft ruptures of ultrathin oxides may not pose as serious a reliability threat to actual transistors.

There is general agreement on the ultrathin voltage-driven wearout model and the percolation theory of breakdown, but there is need for ultrathin technology data relating wearout and breakdown to logic circuit failure, not just to oxide capacitors. The ultrathin oxide experiments indicate that reliabilities may not be as risky as for breakdown in older technologies, but we must take care with these conclusions from wearout studies since they are predominantly done on capacitor oxides and to a lesser extent on transistors having drain, channel, and source regions. The studies reported on the effect of transistor oxide rupture on circuit functionality are now reviewed.

The evolution of ultrathin oxide studies from MOS capacitors to MOSFET transistors shows distinct characteristics. A rupture of the older technology gate oxide shorts may or may not cause logic failure in the IC [32, 33, 49]. However, the effect of ultrathin oxide soft breakdown on transistor V_t and g_m was reported as negligible [52]. Figure 6.30 shows the small time-varying changes in V_t and g_m during the pre-soft-breakdown stress and afterward—V_t dropped by 1.3% and g_m increased by 3.1%.

Crupi et al. [27] stressed 24 Å thin oxide transistors, and found breakdown in the overwhelming majority of *n*MOSFET devices. I_{Doff} was significantly increased, and the $|V_G/I_G|$ ratio showed hard breakdown values in the range from about 1 kΩ to 100 kΩ. High I_{Doff}

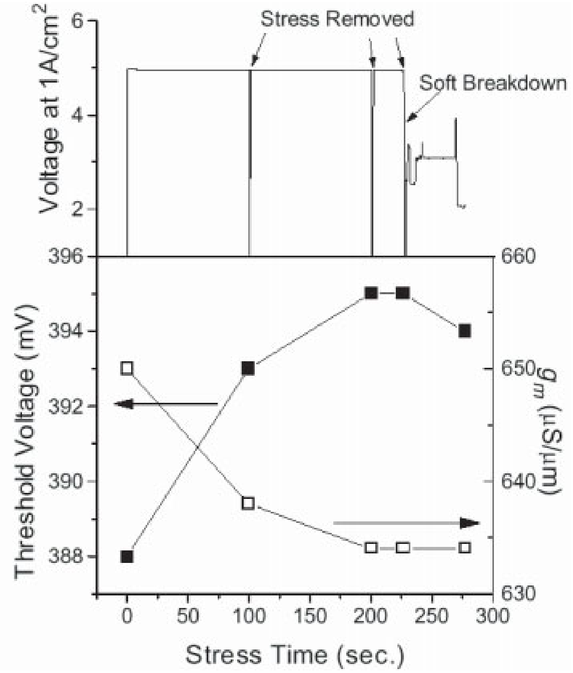


Figure 6.30. Time-varying change in V_t and g_m of an ultrathin oxide during current stress [52].

drain currents from about 1 μA to 1 mA occurred only for breakdown in the in the gate-to-drain region. Soft breakdown with much lower I_{Doff} occurred dominantly in the gate-to-source and gate-to-channel regions of the transistor. Hard breakdowns were not found for any of the three regions in the *p*-channel MOSFETs. The implication is that only the gate-to-drain breakdowns of *n*MOSFETs are serious reliability threats. Although the experiment clearly shows a sensitivity of the *n*-MOSFET, it should be noted that the oxides were protected from harder breakdown by a 1 kΩ series resistor in the gate electrical path. A normal logic IC may show more variation in breakdown hardness from soft to hard categories. Also, the implications for a logic circuit with damaged transistors, such as a NAND gate, were not shown.

Rodriguez et al. measured the effect of ultrathin gate oxide breakdowns on inverter properties [48]. Inverter transfer curves showed weakened logic voltages and, finally, functional failure for inverters that underwent a stronger stress. The weak logic voltage compromises noise margins, and could also cause I_{DDQ} elevation if the weak voltage output is sufficient to turn on downstream load gates. HBD can significantly load a previous logic gate stage to the point of logic failure or severe weakening of noise margin.

Dumin et al. showed an interesting result that the stress on an *n*MOS transistor oxide is greater if $V_G = 0$ V and $V_D = V_{DD}$ [29]. An inverter in the high-output-logic state would show this stress. This contrasts with traditional thinking, which assumed that the gate voltage was set at V_{DD} and source and drain terminals were at ground potential.

E versus E^{-1} Models for Oxide Breakdown. The research community debated for several years two oxide breakdown models that pertain to oxides of $T_{ox} > 40$ Å [51]. These are the E-model and the E^{-1} (or 1/E) model. The E- and E^{-1} models are increasingly less relevant with ultrathin technology use, but these models consumed large research resources, and they taught us a great deal about oxide properties. The E-model is thermodynamic, and one interpretation is that the initial event is field emission of an electron in the oxide. When field strength is high enough, an electron can be pulled from an atom in the material. This field emission causes a trap, and when a sufficient number of traps are vertically lined up, the oxide field strength exceeds that needed to rupture it. The E^{-1} model assumes an initial preferential tunneling of charge into a spot that has local thinning with respect to neighboring regions. A trap occurs at the thinner spot, resulting in a higher oxide field strength and leading to more tunneling and damage. The damage increases the oxide field in the region of traps until rupture occurs. The difficulty in distinguishing between the two models is that test data must be taken at abnormally high field strengths to accelerate the failures in a reasonable time. Wearout and breakdown data would take months or years to collect at normal use field strengths. The data on high oxide field stress (> 8 MV/cm) overlay almost exactly for the E- and E^{-1} models, and that was the problem.

Figure 6.31 compares time to breakdown, t_{BD}, found when plotting the data with the E-model or E^{-1} model. Although t_{BD} predictions are virtually identical in the high field region, a wide discrepancy is seen for projections of data to user conditions between 2–5 MV/cm. Experiments done at lower field strengths and elevated temperatures show that high-temperature breakdowns occurred with the same mechanism as those at lower temperatures. Suehle and colleagues then used this observation to fit t_{BD} data to the E-model over a broad range, including the user region [50].

6.4.2 Hot Carrier Injection (HCI)

This second major oxide failure mechanism occurs when the transistor electric field at the drain-to-channel depletion region is too high. This leads to the hot carrier injection (HCI)

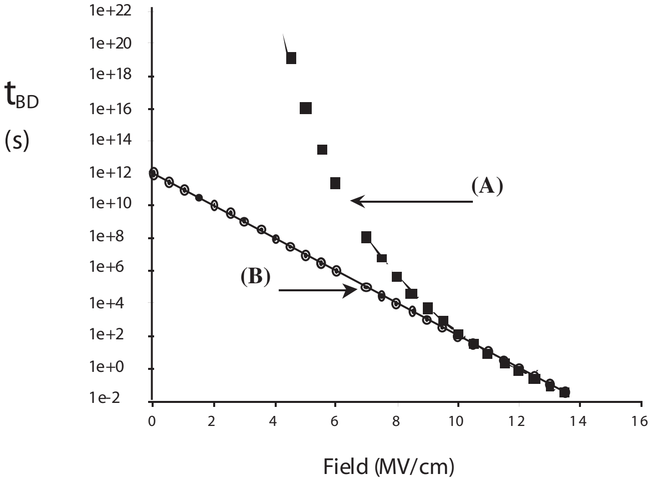


Figure 6.31. Time to breakdown t_{BD} for E^{-1} model plot (A) and E model plot (B) using extrapolation of data from high-oxide fields (8–10 MV/cm) [51].

effects that can alter circuit timing and high-frequency performance. HCI is a systematic failure resulting in a decline in the maximum operating frequency (F_{max}) of the IC. It seldom leads to catastrophic failure. The typical parameters affected are: I_{Dsat}, transistor transconductance (g_{m}), threshold voltage (V_t), weak inversion subthreshold slope (S), and increased gate-induced drain leakage (GIDL).

HCI can happen if the power supply voltage is higher than intended for the design, the effective channel lengths are too short, there is a poor oxide interface or poorly designed drain–substrate junctions, or overvoltage accidentally occurs on the power rail. Figure 6.32 sketches an *n*MOS transistor cross section showing the drain depletion field. The horizontal electric field in the channel $\mathscr{E}_{\mathrm{ch}}$ gives kinetic energy to the free electrons moving from the inverted portion of the channel to the drain. When the kinetic energy is high enough, electrons strike Si atoms around the drain–substrate interface causing impact ionization. Electron–hole pairs are produced in the drain region and scattered. Some carriers go into the substrate, causing an increase in substrate current I_{SUB}, and a small fraction have enough energy to cross the oxide barrier and cause damage. It is estimated that an electron needs at least 3.1 eV to cross the barrier and a hole needs 4.6 eV. Even more energy is needed to break bonds leading to trap formation. Typically, damage is creation of acceptor-type interface traps near the drain by electrons with energies of 3.7 eV or higher. A possible mechanism is that a hot electron breaks a hydrogen–silicon bond at the Si–SiO_2 interface. If the silicon and hydrogen recombine, then no interface trap is created. If the hydrogen diffuses away, then an interface trap is created [42].

The energy follows a Boltzmann distribution in which particle thermal energy is $E_t = kT/q$ where k is Boltzmann's constant, T is degrees Kelvin, and q is the electron charge. An electron of 3.1 eV then has an equivalent mean temperature of $T = E_t/k = 3.1$ eV/(86.17 μeV/K) = 36,000 K. This is the basis for the expression "hot electrons." Ambient temperature has an interesting relation to HCI since carrier mobility increases as temperature decreases. Carriers with higher mobility more efficiently create hot holes and electrons, so that HCI increases as temperature is lowered. This property is sometimes used when using HCI reliability test structures to rapidly show damage.

Once a hot carrier enters the oxide, the vertical oxide field $\mathscr{E}_{\mathrm{ox}}$ determines how deeply the charge will go. If the drain voltage is positive with respect to the gate voltage, then holes entering the oxide near the drain are accelerated deeper into the oxide, and electrons in the same region will be retarded from leaving the oxide interface. $\mathscr{E}_{\mathrm{ch}}$ restricts the damage to oxide over the drain–substrate depletion region, with only a small amount of dam-

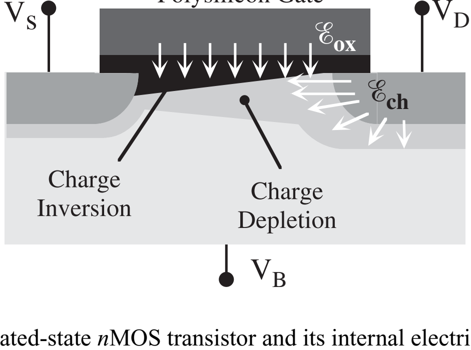


Figure 6.32. Saturated-state *n*MOS transistor and its internal electric fields $\mathscr{E}_{\mathrm{ch}}$ and $\mathscr{E}_{\mathrm{ox}}$.

age just outside the depletion region. In practice, the I_{Dsat} parameter is typically used to measure HCI degradation. I_{Dsat} is the transistor parameter that most closely approximates the impact on circuit speed, since it impacts the charge and discharge of load capacitors. Also, the MOSFET current model equations in Chapter 3 showed that I_{Dsat} is a function of V_t. The increased trap density and subsequent charging of the traps alters transistor threshold voltage V_t. Typically, *n*MOS transistors show increased V_{tn} causing the transistor to slow, and decreased I_{off} and g_{m}. *p*MOS transistors typically show the opposite effect: $|V_{tp}|$ decreases, I_{off} and g_{m} increase, and the transistors switch faster.

Figure 6.33 shows time degradation for a *n*MOS transistor under a hot-carrier stress. The important circuit speed parameter is the I_{Dsat} parameter that shows only slight degradation in time. This is the parameter that largely controls the oscillation frequency of a circuit such as a ring oscillator or a microprocessor. V_{tn} also changes slightly in time. Other parameters change more readily such as the g_{m}, I_{Dsat} reverse, and I_{Dsat} forward. The forward and reverse designations refer to normal bias of the drain and source (forward), and reversing the normal bias of the drain and source (reverse). The point is that whereas some transistor parameters change markedly, I_{Dsat} is the overall speed determining parameter, and it changes slowly.

Figure 6.34 shows $I_{\mathrm{D}} = I_{\mathrm{off}}$ (off-state leakage current) versus gate voltage for a *p*MOS transistor subjected to a drain-to-source overvoltage. This nominal 2.8 V transistor had $V_{\mathrm{DS}} = -4.5$ V during the time of the measurements. After 1 minute of stress, I_{off} increased over two orders of magnitude. The damage is quick and easily measured, but, surprisingly, such damage does not affect circuit performance to the same degree. Chatterjee et al. reported that a stressed ring oscillator failure, defined as a 5% reduction in oscillation frequency, occurred for a time 100 times longer than the 10% damage criteria for measuring

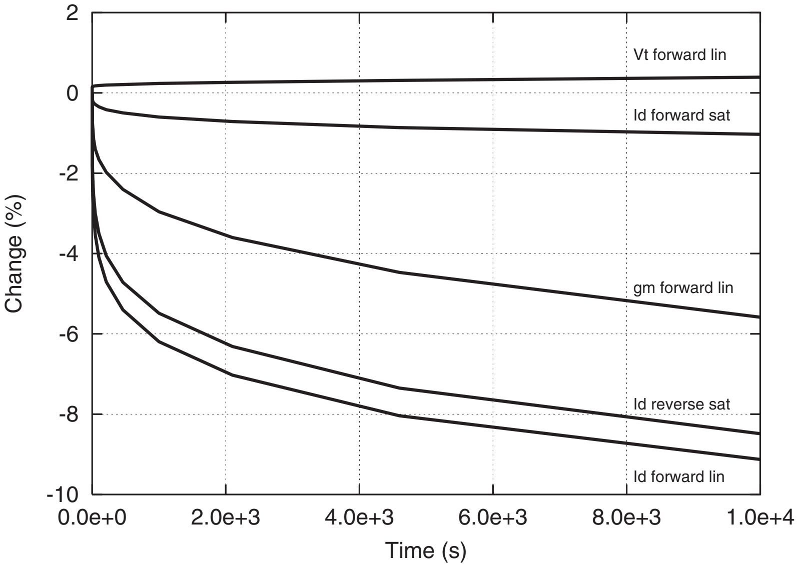


Figure 6.33. Plot of HCI degradation for transistor parameters. (Reproduced by permission of Duane Bowman, Sandia National Labs.)

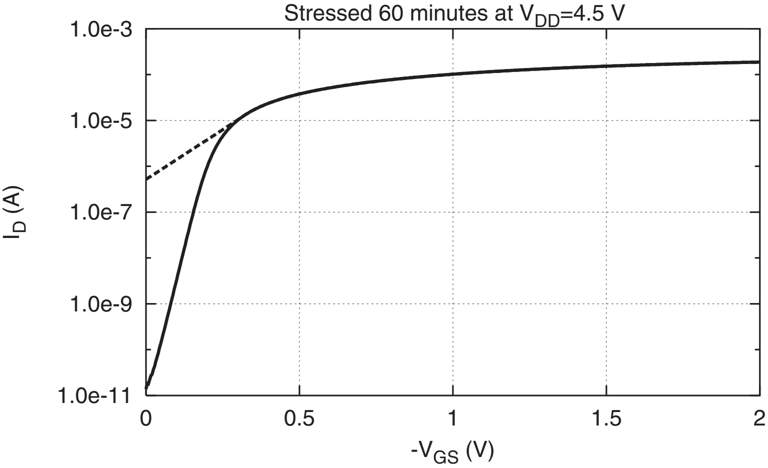


Figure 6.34. Stress time and *p*MOS transistor damage due to hot-carrier injection in a 0.35 μm technology.

I_{LIN}, the drive current of the transistor in an ohmic bias state [25]. Reasons for this paradox are developed next.

Why does the obvious damage to a transistor by HCI not evoke the same measure of IC performance reduction? One reason is that the dominant speed parameter I_{Dsat} is minimally affected by HCI stress. The oxide damage occurs dominantly in the oxide over the drain-depletion–substrate-depletion region when the transistor is in the saturated bias state, and that only occurs during the logic transition. If HCI damage is present in a *n*MOS oxide, then V_{tn} is increased only in the drain-depletion region. Charge inversion and V_{tn} are irrelevant in the depletion region during the logic transition. In the transition from V_{DD} to GND, the transistors are in the depletion state for about 75% of the excursion. The damage is in an unusual “don’t care” location. If a transistor drain and source are electrically exchanged, then much more damage is observed, since the threshold voltage is critically altered near the source region affecting normal carrier inversion. Another effect offsetting IC performance is that a *p*MOS transistor undergoes a drop in V_{tp} and operates faster than normal.

The typical HCI effect is reduction in F_{max} (maximum measured operating frequency). A production concern is that ICs with statistically short effective channel lengths will have better F_{max} performance, but higher drain depletion fields. The question is whether F_{max} will degrade by HCI during customer use. Unless HCI is severe, then expected reductions in operating frequency are on the order of 1–3%. Guardbanding at test by raising the F_{max} limit by 5% is one protection. However, all manufacturers must make these determinations from their own parts characterization.

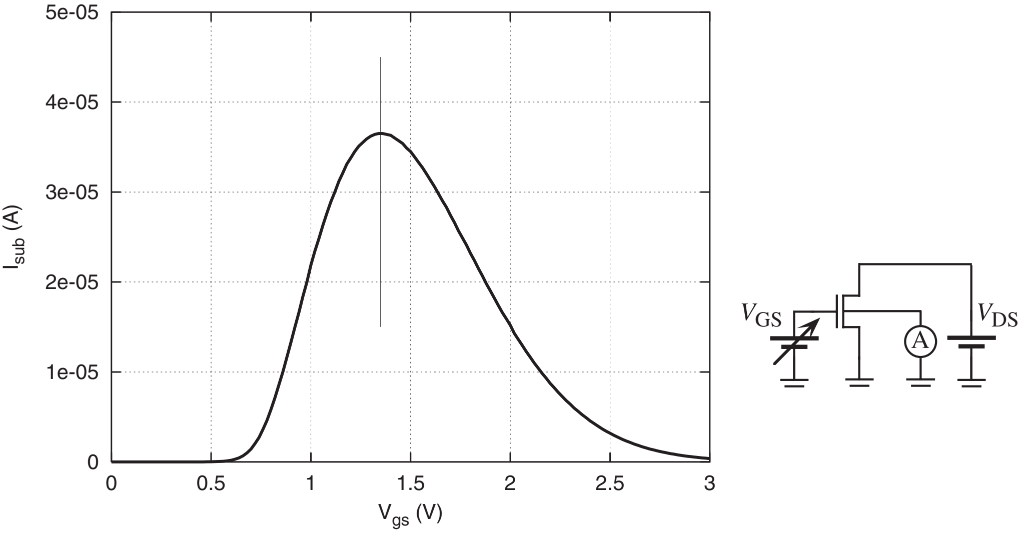
Hot-Carrier Injection and Bias State. Figure 6.35(a) shows the substrate current I_{SUB} versus V_{GS} when V_{DS} is high, holding the transistor mostly in the saturated state. The schematic for this measurement is shown in Figure 6.35(b). Hot-carrier generation in the drain-depletion region causes impact ionization with holes and electrons entering the substrate as well as the oxide. I_{SUB} is larger and more easily measured than I_{G}, and is often used as a proportionality indicator of hot-carrier generation.

The plot for an n-channel transistor in Figure 6.35(a) peaks near $V_{\mathrm{GS}} \approx 0.5\ V_{\mathrm{DS}}$. When V_{G} is below threshold, few carriers exist in the channel, and I_{SUB} is near zero. When V_{GS} approaches and becomes larger than V_{DS}, the transistor enters the nonsaturation state, and the depletion field at the drain disappears. Again no hot carriers are generated. When V_{GS} goes just above V_t then the device is in saturation and free carriers exist in the channel, some of which cause hot-carrier generation. Holes and electrons that enter the gate experience an oxide electric field whose positive field is at the drain. The gate voltage is lower than the drain voltage so that holes are preferentially attracted to the gate.

The left-hand portion of the curve in Figure 6.35(a) is the region in which holes enter the oxide and become gate current. When V_{GS} goes well beyond the peak, the gate voltage rises, and electrons are drawn to the gate. The peak in the I_{SUB} curve is a condition in which both holes and electrons are entering the oxide, but neither with maximum field attraction such as at the ends of the curve. The holes and electrons in the middle portion of the curve tend to cause more interface damage here, in contrast to traps deeper in the oxide.

These bias curves are useful for engineers who design reliability monitor structures for measuring HCI. Wafer-level reliability (WLR) monitor structures are designed for rapid measurement of damage. The HCI is maximized by biasing transistors at these worst-case conditions (i.e., $V_{\mathrm{GS}} \approx V_{\mathrm{DD}}/2$) and even at lower temperatures. The curves also show that HCI occurs during the logic state transitions, and not during the quiescent states. HCI requires the saturated bias state of the transistor, and that only occurs during the logic gate transition. The terminal polarities of an inverter are more dynamic than the curves in Figure 6.35 since V_{DS} is not constant, but drops as V_{GS} increases. $\mathscr{E}_{\mathrm{ch}}$ and $\mathscr{E}_{\mathrm{ox}}$ vary in a complex manner over a wide range during a logic transition.

Prediction of Integrated Circuit HCI Reliability. One method of estimating how long an IC will perform in a normal hot-carrier injection environment surprisingly uses



(a) (b)

Figure 6.35. (a) n-channel transistor hot-carrier generation curves as function of V_{GS} and I_{SUB}. (b) Schematic for measurement.

data collected from individual transistors. These data are then combined with estimates of the IC duty cycle and other operating parameters. We thank Steve Mittl of IBM, who lectured on these subjects at the University of New Mexico, for the following discussion.

A theoretical model forms the basis of HCI reliability prediction for most companies [34]. The end result is a calculation of transistor lifetime τ as

$$\tau = C'\frac{W}{I_D}\left(\frac{I_{SUB}}{I_D}\right)^{-\phi_{it}/\phi_i} \tag{6.25}$$

where C' is a constant, W is transistor width, I_D is drain current, I_{SUB} is substrate current, ϕ_{it} is the interface state activation energy ($\approx$ 3.7 eV), and ϕ_i is the impact ionization activation energy ($\approx$ 1.2 eV). The ratio of ϕ_{it}/ϕ_i is a constant of about 3. τ is the defined lifetime degradation due to HCI (i.e., a 10% drop in V_t or g_m, etc.). If we rearrange the terms, we get

$$\frac{\tau I_D}{W} \propto \left(\frac{I_{SUB}}{I_D}\right)^{-\phi_{it}/\phi_i} \tag{6.26}$$

We can measure τ, I_D, and I_{SUB} for each transistor in the stress experiment, and we know W and $\phi_{it}/\phi_I \approx 3$. If we plot these variables on log-log scale, we get a straight line such as shown in Figure 6.36. The data on the lower right are for transistors that failed in a few seconds under a HCI stress condition.

The slope in Figure 6.36 is constant. The line will be offset for transistors of different quality. The next step is to estimate chip HCI lifetime. The method predicts chip HCI lifetime by estimating I_D and I_{SUB} at use conditions, and then derives τ from those use conditions in Figure 6.36. We then use the conversion equation from DC stress into chip power-on hours (POH)

$$\text{DC stress time} = \text{duty factor} \times \text{switch factor} \times \text{chip POH} \tag{6.27}$$

where duty factor is (device DC equivalent stress per cycle)/(cycle time). Switch factor is a fraction that may consider that stress only occurs during low-to-high input transitions, and

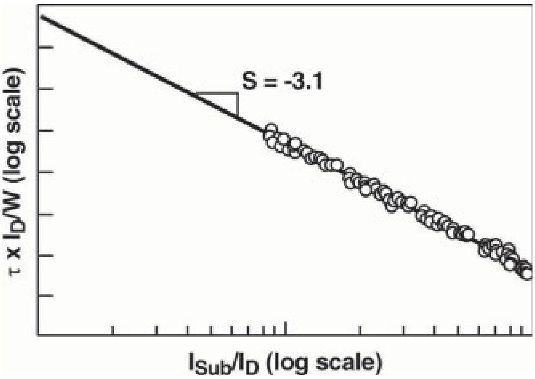


Figure 6.36. Lifetime projection of HCI degradation for several transistors (from [42]).

then you solve for chip POH. Remember that HCI damage occurs only in normal operation when the transistor is in saturation, and that occurs only during the logic transitions. The method is tedious, but does provide HCI estimated lifetimes. Power-on hours are typically 10 years, but will vary with product expectations. Typical HCI reliability goals are about 0.1–1.0 years of DC stress. Practical stresses of test structures are 10–100 hours.

6.4.3 Defect-Induced Oxide Breakdown

Foreign particulates or poor quality oxides typically lead to early breakdown of the oxide for lower applied voltages. These are called gate oxide shorts. The origins of particulate-driven gate shorts are different than those for HCI or wearout, but the end result can be similar to breakdown by wearout. The time to failure t_F is shorter for defect-related gate oxide shorts appearing as early as the production test or from field returns weeks to years later. Manufacturers try to eliminate defect-induced gate shorts by stressing the ICs at high voltages for short time periods (burn-in). Chapter 7 will analyze the electronic effects that occur when gate oxide shorts are present in an IC.

6.4.4 Process-Induced Oxide Damage

Plasma and ion implantation process techniques are manufacturing steps used for etching or doping some areas of the circuit during semiconductor device fabrication. During these steps, devices may be exposed to charges collected by aluminum or polysilicon conducting lines (called “antennas”) connected to the gate terminals of MOSFETs. As a result, gate oxides can be damaged, affecting V_t, the subthreshold slope, and the transconductance of transistors. The stress currents to which these charges give rise may eventually cause oxide breakdown. For oxides that do not experience breakdown, oxide trapped charges and interface states can be created. These damaging processes can be avoided by providing alternative conducting paths to the charge collected during these fabrication steps and thus protecting the gate oxides. Often, reverse-biased diodes are tied to the “antennas” to provide a charge path to ground.

6.4.5 Negative Bias Temperature Instability (NBTI)

A recent oxide reliability issue has appeared that impacts short-channel *p*MOS transistors with their *p*-doped polysilicon gates. It is called negative bias temperature instability (NBTI), and it is a wearout mechanism with positive charge buildup at the channel interface of *p*MOS transistors. It causes threshold voltage absolute magnitude increase and reduction in I_{Dsat}. The damage has been referred to as caused by “cold holes” [38], and is identified as getting worse with scaling of oxide thickness. The “instability” refers to the time variation in V_{tp} and I_{Dsat} [21]. The affected *p*MOS transistor has higher V_{tp} than a transistor in normal inversion.

Several mechanisms have been proposed for NTBI but, presently, the accepted one is a hole-assisted electrochemical reaction [38]. Interface states are formed when a hole breaks a silicon bond, forming trivalent silicon in the interface region. Hydrogen is believed to be the spun-off element, and as H diffuses away, a positive charge is left on the trivalent Si atom. H_2O and holes are believed to be reactants. The exact nature of the reaction is still not well understood. The necessary components for NBTI are holes, hydrogen (either as H_2, H^+, or H_2O moisture), high temperature (> 100°C), and oxide voltage [38].

NBTI occurs in the *p*MOS transistor when $V_{\text{in}} = 0$ V, but NBTI shows some recovery phenomena when $V_{\text{in}} = V_{\text{DD}}$, especially at high temperatures such as at burn-in. Chen et al. reported that for inverter experiments, NBTI was less during dynamic stressing than that of DC stressing, and that NBTI damage was overestimated for the DC studies [26]. NBTI is a significant problem in *p*MOSFET performance for advanced technologies [38]. Circuit design and lifetime projections must consider the competing degradation from hot-carrier injection.

6.5 CONCLUSION

Metal and oxide failure mechanisms were described that showed the relation between material properties and potential IC failure. Electromigration and stress voiding are constant challenges for deep-submicron transistor ICs. Oxide wearout, hot-carrier injection, oxide ruptures due to defects, and NBTI are other materials concerns. Engineers in the CMOS IC industry need to understand these reliability failure mechanisms.

ACKNOWLEDGMENTS

We thank Bill Filter, Joe Clement, Bill Miller, Ted Dellin, Dave Monroe, Duane Bowman and Steve Yazzie of Sandia National Labs, and Shannon Hawkins of Boeing Aircraft Corp. for critical reviews and suggestions. Bill Filter´s stress void lectures and Joe Clement's electromigration lectures in the University of New Mexico Microelectronics Reliability class were particularly beneficial. We also thank Steve Nelson and Rod Augur of Philips Semiconductors, Steve Mittle, Jim Lloyd, Tim Sullivan, and Bob Rosenberg of IBM, Cleston Messick of Fairchild Semiconductors, Tim Turner of Keithley Instruments, and Don Pierce and Eric Snyder of Sandia Technologies, who lectured in the University of New Mexico Microelectronics Reliability Series Lectures and influenced us.

BIBLIOGRAPHY

Materials Science

1. J. Ahn, H.-S. Kim, T.-J. Kim, H.-H. Shin, Y.-Ho Kim, D.-U. Lim, J. Kim, U. Chung, S.-C. Lee, and K.-P. Suh, "1GHz microprocessor integration with high performance transistor and low RC delay," in *Proceedings of International Electron Devices Meeting (IEDM),* pp. 28.5.1–28.5.4, December 1999.
2. R. A. Flinn and P. K. Trojan, *Engineering Materials and Their Applications,* Houghton Mifflin, 1986.
3. Annual MRS Symposium Proceedings (1991–present), Vols. (225, 265, 309, 338), titled *Materials Reliability Issues in Microelectronics,* from the Materials Research Society.
4. R. E. Reed-Hill and R. Abbaschian, *Physical Metallurgy Principles,* PWS-Kent, 1997.

Electromigration

5. J. Black, "Mass transport of aluminum by momentum exchange with conducting electrons," in *International Reliability Physics Symposium,* pp. 148–159, April 1967.

6. I. Blech and E. Meieran, "Electromigration in thin aluminum films," *Journal of Applied Physics, 40,* 2, 485–491, 1968.
7. J. Clement, "Electromigration modeling for integrated circuit interconnect reliability analysis," *IEEE Transactions on Device and Materials Reliability, 1,* 1, 33–42, March 2001.
8. J. Gill, T. Sullivan, S. Yankee, H. Barth, and A. von Glasow, "Investigation of via-dominated multi-modal electromigration failure distributions in dual damascene Cu interconnects with a discussion of the statistical implications," in *International Reliability Physics Symposium (IRPS),* pp. 298–304, April 2002.
9. A. Goel and Y. Au-Yeng, "Electromigration in the VLSI interconnect metallizations," in *32nd IEEE Midwest Symposium on Circuits and Systems,* pp. 821–824, 1989.
10. K-D. Lee, X. Lu, E. Ogawa, H. Matsuhashi, V. Blaschke, R. Augur, and P. Ho, "Electromigration study of Cu/low k dual damascene interconnects," in *International Reliability Physics Symposium (IRPS),* pp. 322–326, April 2002.
11. B. Li, T. Sullivan, and T. Lee, "Line depletion electromigration characteristics of Cu interconnects," in *International Reliability Physics Symposium (IRPS),* pp. 140–145, April 2003.
12. J. Lloyd, "Topical review: electromigration in thin film conductors," *Semiconductor Science Technology, 12,* 1,177–1,185, 1997.
13. E. Ogawa et al., "Stress-induced voiding under vias connected to wide Cu metal leads," in *International Reliability Physics Symposium (IRPS),* pp. 312–321, April 2002.
14. D. Pierce and P. Brusius, "Electromigration: A review," *Microelectronics Reliability, 37,* 7, 1,053–1,072, 1997.
15. W. Song et al., "Pseudo-breakdown events induced by biased-thermal-stressing of intra-level Cu interconnects-reliability and performance impact," in *International Reliability Physics Symposium (IRPS),* pp. 305–311, April 2002.
16. D. Pierce, E. Snyder, S. Swanson, and L. Irwin, "Wafer-level pulsed DC electromigration response at very high frequencies," in *International Reliability Physics Symposium (IRPS),* pp. 198–206, April 1994.

Stress Voiding

17. *American Institute of Physics (AIP) Conference Proceedings #263,* "Stress-Induced Phenomena in Metallization," (1991 to present, every two years).
18. M. F. Ashby and D. R. H. Jones, *Engineering Materials I: An Introduction to Their Properties and Applications,* 2nd ed., Butterworth-Heinemann, 1996.
19. K. Doong et al., "Stress-induced voiding and its geometry dependence characterization," in *International Reliability Physics Symposium (IRPS),* pp. 156–160, April 2003.
20. S. Rauch and T. Sullivan, "Modeling stress-induced void growth in Al–4 wt% Cu lines," *Proceedings of SPIE, 1,* 805, 197–208, 1993.

Oxide Reliability

21. W. Abadeer and W. Ellis, "Behavior of NBTI under ac dynamic circuit conditions," in *International Reliability Physics Symposium (IRPS),* pp. 17–22, April 2003.
22. M. Alan and R. Smith, "A phenomenological theory of correlated multiple soft-breakdown events in ultra-thin gate dielectrics," in *International Reliability Physics Symposium (IRPS),* pp. 406–411, April 2003.
23. G. Barbottin and A. Vapaille, *Instabilities in Silicon Devices, Vol. 2: Silicon Passivation and Related Instabilities,* North-Holland, 1989.
24. D. Buchanan, "Scaling the gate dielectric: Materials, integration and reliability," *IBM Journal of Research and Development, 43,* 3, 245–264, 1999.

25. P. Chatterjee, W. Hunter, A. Amerasekera, S. Aur, C. Duvvury, P. Nicollian, L. Ting, and P. Yang, "Trends for deep submicron VLSI and their implications for reliability," in *International Reliability Physics Symposium (IRPS),* April 1995.

26. G. Chen et al., "Dynamic NBTI of PMOS transistors and its impact on device lifetime," in *International Reliability Physics Symposium (IRPS),* pp. 196–207, April 2003.

27. F. Crupi, B. Kaczer, R. Degraeve, A. De Keersgieter, and G. Groeseneken, "Location and hardness of the oxide breakdown in short channel n- and p-MOSFETs," in *International Reliability Physics Symposium (IRPS),* pp. 55–59, April 2002.

28. R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, "Relation between breakdown mode and breakdown location in short channel nMOSFETs," in *International Reliability Physics Symposium (IRPS),* pp. 360–366, May 2001.

29. N. Dumin, K. Liu, and S.-H. Yang, "Gate oxide reliability of drain-side stresses compared to gate stresses," in *International Reliability Physics Symposium (IRPS),* pp. 60–72, April 2002.

30. D. Fleetwood, "Fast and slow border traps in MOS devices," *IEEE Transactions on Nuclear Science, 43,* 3, 779–786, June 1996.

31. C. R. Helms and B. E. Deal, *The Physics and Chemistry of SiO_2 and the Si–SiO_2 Interface,* Plenum, 1988.

32. C. Hawkins and J. Soden, "Electrical characteristics and testing considerations for gate oxide shorts in CMOS ICs," in *International Test Conference (ITC),* pp. 544–555, November 1985.

33. C. Hawkins and J. Soden, "Reliability and electrical properties of gate oxide shorts in CMOS ICs," in *International Test Conference (ITC),* pp. 443–451, September 1986.

34. C. Hu et al., "Hot-electron induced MOSFET degradation–Model, monitor and improvement," *IEEE Transactions on Electron Devices, ED-32,* 375–385, February 1985.

35. V. Huard, F. Monsieur, and S. Bruyere, "Evidence for hydrogen-related defects during NBTI stress in p-MOSFETs," in *International Reliability Physics Symposium (IRPS),* pp. 178–182, 2003.

36. Proceedings and Tutorials, *International Reliability Physics Symposium (IRPS),* Annual Spring Conference Sponsored by IEEE.

37. J. McKenna, E. Wu, and S-H Lo, "Tunneling current characteristics and oxide breakdown in p+ poly gate pFET capacitors," in *International Reliability Physics Symposium (IRPS),* pp. 16–20, April 2000.

38. G. La Rosa, "NBTI challenges in *p*MOSFETs of advanced CMOS technologies," *Tutorial in International Reliability Physics Symposium (IRPS),* Section 241, April 2003.

39. Y. Leblebici, *Hot-Carrier Reliability of MOS VLSI Circuits,* Kluwer Academic, 1993.

40. B. Linder, J. Statis, D. Frank, S. Lombardo, and A. Vayshenker, "Growth and scaling of oxide conduction after breakdown," in *International Reliability Physics Symposium (IRPS),* pp. 402–405, April 1998.

41. J. McPherson and H. Mogul, "Disturbed bonding states in SiO_2 thin-films and their impact on time-dependent dielectric breakdown," in *International Reliability Physics Symposium (IRPS),* pp. 47–56, April 1998.

42. S. Mittl, IBM Corp. Hot-carrier injection lecture given at the University of New Mexico in March 2000.

43. F. Monsieur, E. Vincent, D. Roy, S. Bruyere, G. Pananakakis, and G. Ghibaudo, "Time to breakdown and voltage to breakdown modeling for ultra-thin oxides (T_{ox} < 32 Å)," *Proceedings of IEEE International Reliability Workshop (IRW),* pp. 20–25 , October 2001.

44. F. Monsieur et al., "Evidence for defect-generation-driven wearout of breakdown conduction path in ultrathin oxides," in *International Reliability Physics Symposium (IRPS),* pp. 424–431, April 2003.

45. E. Nicollian and J. Brews, *MOS Physics and Technology,* Wiley, 1982.

46. P. Nicolean, W. Hunter, and J-C Hu, "Experimental evidence for voltage-driven breakdown models in ultrathin gate oxides," in *International Reliability Physics Symposium (IRPS),* pp. 7–15, April 2000.

47. V. Reddy, "Introduction to semiconductor reliability," *Tutorial at International Reliability Physics Symposium (IRPS),* Section 111, April 2002.

48. R. Rodriguez, J. Statis and B. Linder, "Modeling and experiemntal verification of the effect of gate oxide breakdown on CMOS geometry," in *International Reliability Physics Symposium (IRPS),* pp. 11–16, April 2003.

49. J. Segura, C. De Benito, A. Rubio, and C. Hawkins, "A detailed analysis of GOS defects in MOS transistors: testing implications at circuit level," in *International Test Conference (ITC),* pp. 544–550, October 1995.

50. J. Suehle and P. Chaparala, "Low electric field breakdown of thin SiO_2 films under static and dynamic stress," *IEEE Transactions on Electron Devices, 44,* 5, May 1995.

51. J. Suehle, "Ultrathin gate oxide reliability: Physical models, statistics, and characterization," *IEEE Transactions on Electron Devices, 49,* 6, 958–971, June 2002.

52. B. Weir et al., "Ultra-thin dielectrics: They break down, but do they fail?," in *International Electron Device Physics Symposium (IEDM),* pp. 41–44, December 1997.

EXERCISES

6.1. Why would a Cu alloy in an Al base interconnect (a substitutional interconnect) increase the resistivity over pure Al?

6.2. When ICs increase their operating temperature from about 85°C to 120°C, how does this impact metal reliability?

6.3. When electromigration time to fail is plotted against the width of metal samples, a typical curve looks like the one sketched in Figure 6.37. From your knowledge of electromigration and grain boundary models, why does t_F increase for very narrow and very wide interconnects?

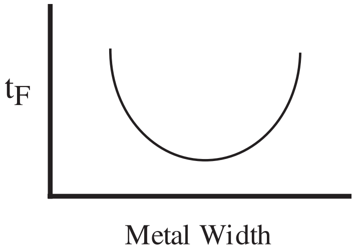


Figure 6.37. Time to failure versus interconnect width.

6.4. Aluminum metal lines using tungsten vias have increased sensitivity to electromigration at the Al/W interface due to the inert nature of W to electromigration. Do copper interconnect systems using Cu vias form a perfect interface, thus avoiding the W/Al problem?

6.5. The line that interconnects a 2NAND gate drive circuit to a load 2NOR gate input has certain electromigration design rules. If an *n*-channel transistor in the 2NOR gate connected to this line acquires a gate-to-source oxide rupture, discuss the electromigration risk.

6.6. Two walls in an airtight room move slowly inward and stop. The room pressure is now 10 atmospheres. There is no net motion of air molecules under this high stress. Why?

6.7. Why does electromigration failure rate decrease at higher clock frequencies?

6.8. An unpassivated interconnect line has a Blech constant of 3500 A/cm and a Blech length of 95 μm. If the same structural interconnect is passivated, the Blech constant goes to 6500 A/cm. How is the Blech length effected?

6.9. Given for an Al interconnect that $\alpha_{\text{Al}} = 23.5 \times 10^{-6}/°\text{C}$, $\alpha_{\text{SiO2}} = 0.5 \times 10^{-6}/°\text{C}$, $Y_m =$ 71.5 GPa, and $\nu = 0.35$. Compare the stress at 30°C when the passivation deposition temperature is lowered from 430°C to 400°C to 300°C.

6.10. Cu with its higher melting temperature of 1085°C was expected to offer total protection from electromigration. What prevented Cu from being the perfect electromigration metal?

6.11. Copper interconnects require barrier metal lines to confine Cu. Compare the via-resistive effect of a Ta liner that occupies 10% of the damascene space to that of a pure Cu via. The dimensions of a cylindrical interconnect are given in Figure 6.38 ($\rho_{\text{Cu}} = 1.7\ \mu\Omega \cdot \text{cm}$ and $\rho_{\text{Ta}} = 200\ \mu\Omega \cdot \text{cm}$).

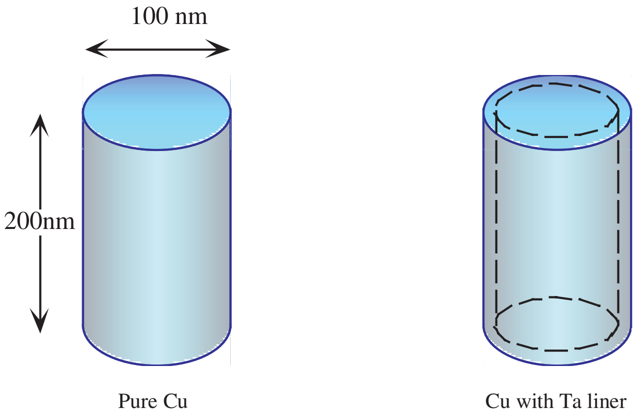


Figure 6.38.

6.12. An oxide dielectric has a shorter time to failure if the oxide has greater area and thinner dimensions. Use the percolation model to explain why.

6.13. Explain what effect large load capacitance has on hot-carrier injection.

6.14. Assume that an aluminum line of 0.5 μm thickness and 0.5 μm width is overlaid with TiN of 0.1 μm thickness. When the Al opens, it leaves a void of 0.5 μm length. The current density is $J = 20$ MA/cm^2. If the resistivities are $\rho_{\text{Al}} = 2.66$ μΩ · cm and $\rho_{\text{TiN}} = 2.66$ μΩ · cm, what is the power dissipation in the TiN section of the break? Give your answer in power per unit area (Watts/cm^2), where area is the bottom face of the TiN.

6.15. Electromigration T_{50} data: two metals evaluated from two different processes. Boltzman's constant = 1.38×10^{-23} eV/°K.

(a) Find the thermal activation energy for both metals and state which metal is better.

(b) Give reasons why the quality might be different for (A) and (B) in Figure 6.39.

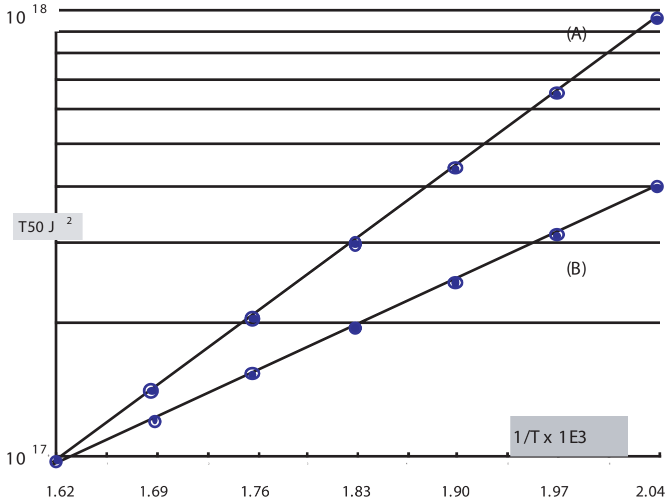


Figure 6.39.

6.16. An overall stress acceleration factor of 5×10^5 is desired for a particular defect in a qualification test. The thermal activation energy is 1 eV, normal temperature is 55°C, normal voltage is 5 V, stress voltage is 7 V, and the oxide voltage acceleration constant B = 400 Å/T_{ox}, where T_{ox} = 100 Å. Calculate the stress temperature T_s in °C to provide the acceleration factor of 5×10^5.

6.17. A company had been using an oxide defect thermal activation of $E_a = 0.3$ eV and a calculated failure rate of 500 FITs (1 FIT = 10^{-9} fails/hour). From new extrapolated data, they found that $E_a = 0.6$ eV. What would the failure rate calculation be if $E_a =$ 0.6 eV? The experimental temperatures were $T_1 = 55$ °C and $T_2 = 125$ °C.

6.18. An aluminum stress experiment is conducted in which $T_1 = 27$ °C, $T_2 = 227$ °C, and $J_1 = J_2 = 10^7$ A/cm^2. The average times to fail are: $t_{\text{F1}} = 5000$ hour and $t_{\text{F2}} = 15$ hour.

(a) Calculate the Al activation energy E_a.

(b) If J_2 increases to 3×10^7 A/cm^2, calculate the expected failure time t_{F2}.

CHAPTER 7

BRIDGING DEFECTS

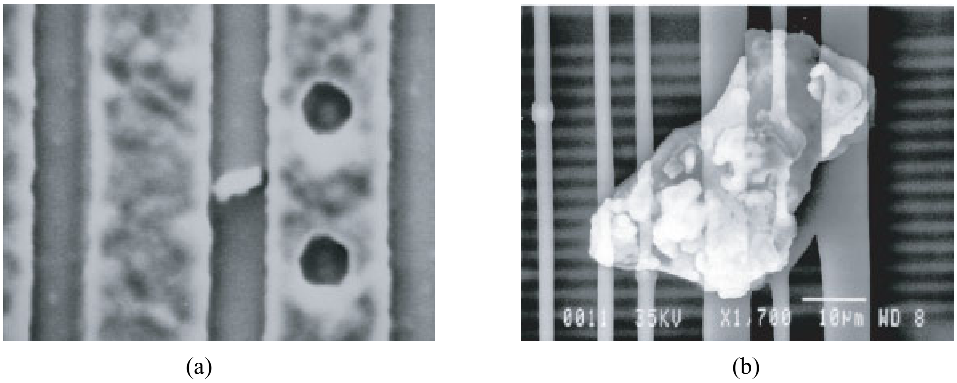
7.1 INTRODUCTION

The previous chapter showed failure mechanisms resulting in shorts between IC conducting paths. A bridge or shorting defect is an unintentional connection between two or more circuit nodes. Bridges in ICs induce abnormal electrical behaviors that depend on certain circuit parameters and the resulting circuit topology. The major bridge defect variables are

- Ohmic or nonlinear
- Intragate-connections across transistor internal nodes
- Connections across the I/O nodes of separate logic gates
- Power rail to ground rail
- Combinational or sequential resulting circuit topology
- Interconnect material types—metal, polysilicon, diffusion region
- Critical resistance—transistor drive strength and *W*/*L* ratios

Ohmic bridge defects can be metal slivers bridging two interconnections (Figure 7.1(a), large amounts of material shorting more than one interconnect (Figure 7.1(b)), or certain forms of transistor gate oxide shorts. Gate oxide short defects are ruptures of the transistor thin oxide that electrically connect the gate to the silicon structures underneath. Gate shorts are well controlled in some fabrication processes and a plague in others. Bridging defects in memory cells or flip-flops may or may not show responses different than those in combinational circuits. Power rail shorts between V_{DD} and GND are common and though they do not involve signal paths of the IC, they need to be recognized and con-

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.



(a) (b)

Figure 7.1. (a) Metal sliver. (b) Metal blob. (Reproduced by permission of Jerry Soden, Sandia National Labs.)

trolled. Low-power and battery-powered products cannot sustain predicted life if defect-induced power supply leakage occurs.

This chapter characterizes bridge defect behavior at the circuit level, showing how to calculate signal node voltages and power supply currents that are altered. We introduce the parameters related to bridging defects, and then analyze their effect at the circuit level for combinational and sequential circuits.

7.2 BRIDGES IN ICs: CRITICAL RESISTANCE AND MODELING

7.2.1 Critical Resistance

Critical resistance may be the most important concept in bridge defect electronics. It relates the defect resistance to the electrical properties of the surrounding circuitry and its induced logic behavior [18]. The critical resistance is the defect resistance value above which the circuit will not functionally fail. Contamination particles like the ones shown in Figure 7.1 cause shorts between two or more lines, resulting in bridging resistance between the shorted nodes.

Figure 7.2(a) shows one inverter (I_1) driving a second inverter (I_2) with an ohmic defect bridge that shorts the output of I_1 (V_2) to the V_{DD} rail. When the input to I_1 is at 0 V, the

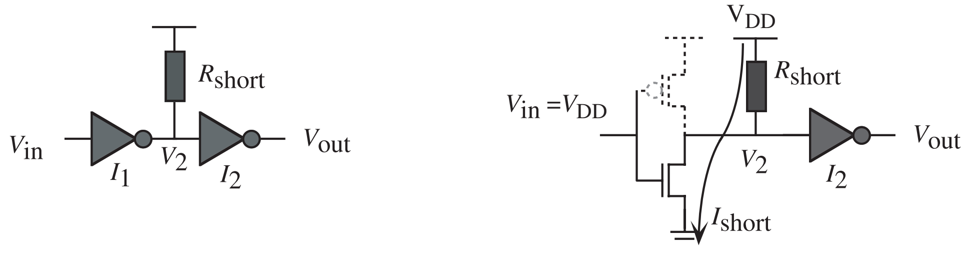


Figure 7.2. Power rail to signal node bridge defect.

*n*MOS transistor is off, whereas the *p*MOS is on and V_2 is pulled to a logic one state. Since there is no voltage drop across the bridging defect, no current passes through the short. In this situation, it is said that the defect is not activated, since it has no effect on the circuit in this logic state. When the input gate is at logic one, the *p*MOS transistor shuts off and the *n*MOS transistor turns on. The *p*MOS transistor is not conducting, and current is drawn through the resistor (I_{short} in Figure 7.2(b)) with a subsequent voltage drop from the V_{DD} rail. The defect is said to be activated for the logic one input signal.

When the defect is activated in Figure 7.2, there is a competition between the *n*MOS transistor of the first inverter that pulls node V_2 to ground and the short resistance that pulls this node to V_{DD}. The final node voltage will depend on the relationship between the value of the short resistance and the current drive of the inverter *n*MOS transistor. Figure 7.3(a) plots a set of voltage measurements at this node versus the input voltage V_{in} for different resistance values obtained for an inverter chain taken from [18]. Figure 7.3(b) shows the power supply current versus the input voltage of the circuit for each resistance value. Figure 7.3(a) shows that the higher the resistance value, the closer the transfer voltage curve to the fault-free one (note that a defect of infinite resistance would be equivalent to a fault-free circuit). When the voltage at the input is at V_{DD}, the *n*MOS transistor transconductance takes its maximum value, and the inverter output is reduced for each decreasing defect resistance value.

When the defect resistance is 1 kΩ, the lowest possible voltage at the I_1 inverter output is around 4.3 V. Clearly, the circuit will exhibit a logic error for this defect resistance value, and also for any other short resistance in the 0–5 kΩ range. For larger defect resistance values, the voltage in node V_2 is correct but logically weak. The successive gates can recover the full logic response so that the static logic behavior remains correct, but the logic gate whose input is near the defect has little protection against electrical noise spikes. Since the critical resistance is defined as the defect resistance value beyond which the logic operation is correct, this example shows a value of about 5 kΩ. All bridges except power rail shorts have a critical resistance whose value depends on the strength of the contending transistor(s) across that bridge.

The impact of a given bridging defect on the logic behavior of a circuit does not depend only on the resistance value and the transconductance of the transistors that compete

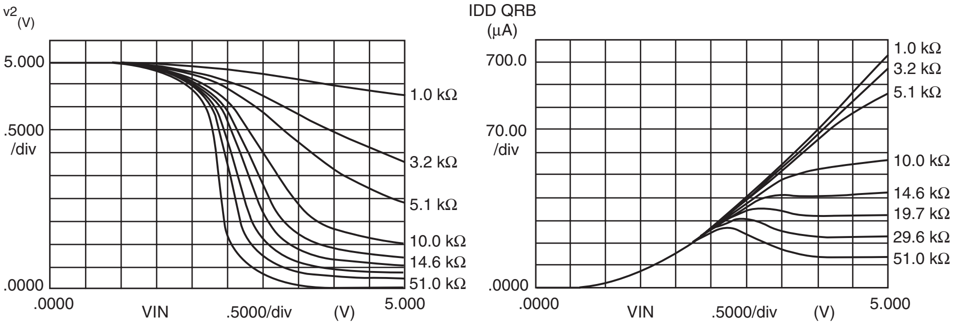


Figure 7.3. (a) Voltages and (b) consumption current for different resistance values of an internal node to V_{DD} short for an inverter chain [18].

with the shorting resistance. The transfer characteristic of the gates driven by the weak node will determine the impact of the defect, since they will interpret the logic value that corresponds to each intermediate voltage. The driver gate may have a symmetrical V_{out} versus V_{in} transfer curve, or it could be skewed to the left or right depending upon the individual *p*- and *n*-channel transistor drive strength (*W*/*L* ratio).

Functional failure is defined by a parameter called the logic threshold voltage V_{TL} and it is measured on the inverter voltage transfer curve at the point where $V_{\text{in}} = V_{\text{out}}$. V_{TL} is easily found as the intersection of a 45° line from the origin and the V_{out} versus V_{in} transfer curve. V_{TL} is typically about $V_{\text{DD}}/2$ which is a convenient parameter to define failure for the critical resistance calculations that follow. Some examples illustrate critical resistance calculations.

■ EXAMPLE 7.1

For the defective circuit in Figure 7.2(a), $K_n = 75\ \mu\text{A/V}^2$, $W/L = 4$, $V_{\text{DD}} = 1.5$ V, $V_{tn} = 0.4$ V, and the logic threshold voltage is 0.75 V. Find the current and voltage V_2 when $R_{\text{short}} = 100\ \Omega$ and (b) $R_{\text{short}} = 4\ \text{k}\Omega$.

(a) Figure 7.2(b) shows the equivalent circuit for analysis. The *p*MOS transistor is removed and the input activates the defect. The saturated state equation for the transistor is

$$I_{\text{Dn}} = K_n \frac{W}{L}(V_{\text{DD}} - V_{tn})^2 = 75\frac{\mu\text{A}}{\text{V}^2}\ 4(1.5 - 0.4)^2 = 363\ \mu\text{A}$$

KVL gives

$$V_{\text{DS}} = V_2 = V_{\text{DD}} - I_{\text{Dn}}R_{\text{short}} = 1.5 - (363\ \mu\text{A})(100) = 1.46\ \text{V}$$

Check the bias state:

$$V_{\text{GS}} < V_{\text{DS}} + V_{tn} \qquad 1.5\ \text{V} < 1.46 + 0.4$$

The transistor is saturated, so the answer is correct and $V_2 = 1.46$ V. This is an error voltage since $V_{\text{in}} = 1.5$ V should produce V_2 of about 0 V, or at least below the logic threshold voltage $V_{\text{TL}} = 0.75$ V.

(b) R_{short} increases to 4 kΩ and the saturated state equation again gives $I_{\text{Dn}} =$ 363 μA, as in (a) and

$$V_{\text{DS}} = V_2 = V_{\text{DD}} - I_{\text{Dn}}R_{\text{short}} = 1.5 - (363\ \mu\text{A})(4\ \text{k}\Omega) = 48\ \text{mV}$$

Check the bias state:

$$V_{\text{GS}} < V_{\text{DS}} + V_{tn} \qquad 5\ \text{V} < 1 + 0.4$$

The transistor is in the nonsaturated state so we must try again by combining the nonsaturated state equation:

$$I_{\text{Dn}} = 75\frac{\mu\text{A}}{\text{V}^2}\ 4[2(V_{\text{in}} - V_{tn})V_2 - V_2^2] = 300\frac{\mu\text{A}}{\text{V}^2}[2(1.5 - 0.4)V_2 - V_2^2]$$

with KVL:

$$I_{\text{Dn}} = \frac{V_{\text{DD}} - V_2}{R_{\text{short}}} = \frac{1.5 - V_2}{4\ \text{k}\Omega}$$

Solving both equations simultaneously:

$$V_2 = 0.492\ \text{V, or } 2.54\ \text{V}$$

the solution is

$$V_2 = 0.492\ \text{V}$$

V_2 is below the logic threshold value of 0.75 V; therefore, a correct logic value would be read. 1 kΩ is above the critical resistance, but this defective circuit has lost significant noise immunity at the V_2 node, making it vulnerable to logic upset by noise spikes. Also, $I_{\text{DDQ}} = (1.5 - 0.492)\ \text{V}/4\ \text{k}\Omega = 252\ \mu\text{A}$ is considerably above the normal quiescent power supply current value for the gate. ■

Critical resistance was evident in these two examples. The 100 Ω defect resistance was small, and V_2 failed its intended logic value. The larger 4 kΩ resistance did not cause V_2 to fail. The critical resistance R_{crit} lies between 100 Ω and 4 kΩ. The next example calculates its exact value.

■ EXAMPLE 7.2

Calculate the exact value of the critical resistance $R_{\text{short}} = R_{\text{crit}}$ for the circuit of Figure 7.2(a), using $K_n = 75\ \mu\text{A/V}^2$, $W/L = 4$, $V_{tn} = 0.4$ V, $V_{\text{DD}} = 1.5$ V, and the logic threshold voltage $V_{\text{TL}} = 0.75$ V.

This problem is not as bad as it looks. The logic threshold of $V_{\text{TL}} = 0.75$ V defines the V_2 voltage point where failure just occurs. If $V_0 = 0.75$ V, then the *n*MOS transistor is in the nonsaturated state since

$$V_{\text{in}} > V_0 + V_{tn} \qquad 1.5\ \text{V} > 0.75\ \text{V} + 0.4\ \text{V}$$

The nonsaturated equation gives

$$I_{\text{Dn}} = 75\frac{\mu\text{A}}{\text{V}^2}\ 4[2(1.5 - 0.4)0.75 - 0.75^2] = 326\ \mu\text{A}$$

$$R_{\text{crit}} = \frac{V_{\text{DD}} - V_0}{I_{\text{Dn}}} = \frac{(1.5 - 0.75)\ \text{V}}{326\ \mu\text{A}} = 2.3\ \text{k}\Omega$$

The circuit will functionally fail if the bridge defect resistance is below 2.3 kΩ. ■

■ EXAMPLE 7.3

Calculate the critical resistance for the bridge defect in Figure 7.4, given $K_p = 300\ \mu\text{A/V}^2$, $K_n = 500\ \mu\text{A/V}^2$, $W/L = 1$ $V_{tn} = 0.35$ V, $V_{tp} = -0.35$ V, $V_{\text{DD}} = 1.2$ V, and the logic threshold for high/low distinction is $V_{\text{TL}} = 0.6$ V.

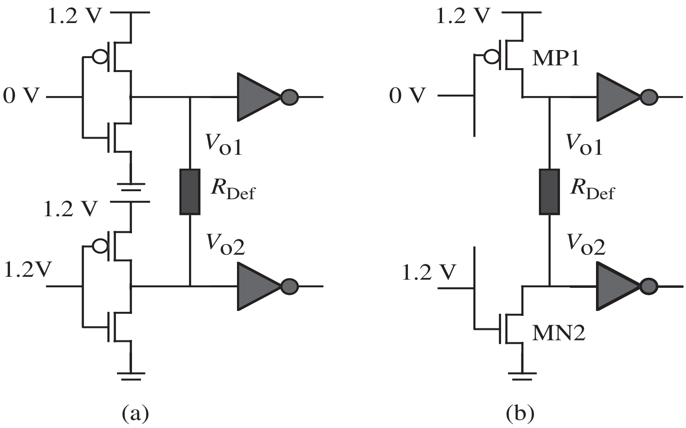


Figure 7.4.

Figure 7.4(b) shows the equivalent circuit when the off-transistors are removed. We must first choose which node attached to R_{Def} will fail first. If V_{o1} drops below 0.6 V, then it fails, or if V_{o2} rises above 0.6 V, then it fails. The clue is that MP1 has weaker drive strength than MN2 and, therefore, cannot win the contest with MN2. The drain at MP1 will fail first. At failure, we can assign V_{DP1} = V_{TL} = 0.6 V. We then know all node voltages except the drain of MN2. Since V_{GP1} = 0 V, V_{DD} = 1.2 V, and V_{DP1} = 0.6 V, you can verify that MP1 is in the nonsaturated state. The drain current through both transistors and the defect is

$$I_{DP1} = 300\frac{\mu A}{V^2}[2(1.2 - 0.35)0.6 - 0.6^2] = 198\ \mu A$$

We can find the voltage V_{o2} and R_{crit} is found by Ohm's law. We know that $V_{o2} <$ 0.6 V, since we deduced that it had not failed. MN2 is then in the nonsaturated state since

$$V_{in} > V_{o2} + V_{tn} \qquad 1.2\ V > 0.6\ V + 0.35\ V$$

The nonsaturated equation allows calculation of V_{o2}:

$$I_{DN1} = 500\frac{\mu A}{V^2}[2(1.2 - 0.35)V_{o2} - V_{o2}^2] = 198\ \mu A$$

where the valid solution is V_{o2} = 279 mV.

R_{crit} is

$$R_{crit} = \frac{V_{o1} - V_{o2}}{I_{DN1}} = \frac{(0.6 - 0.279)\ V}{198\ \mu A}$$

$$R_{crit} = 1.62\ k\Omega$$

■

Self-Exercise 7.1.

For the circuit and problem in Figure 7.4, all parameters are the same except

(a) Calculate R_{crit} if $K_p = 150\ \mu\text{A/V}^2$ and $K_n = 250\ \mu\text{A/V}^2$.

(b) What is the ratio of K_p/K_n when R_{crit} goes to zero; what does that mean?

These problems deepen insight into critical resistance properties. The single transistor bridged to a power rail has a critical resistance that is dependent upon the current drive strength of the transistor. When the defect resistance connects a pull-up to a pull-down transistor, the R_{crit} is a function of the current drive mismatch and the current drive strengths. The signal node with the weaker current drive will fail first. The weaker a transistor's current drive, the larger the resistance needed to protect that node from failure.

R_{crit} increases as the mismatch in pull-up and pull-down current strength gets larger, but R_{crit} goes to zero when the pull-up and pull-down transistor current strengths are equal. The product significance is that real IC bridging defects have a range of resistance or impedance values, and many ICs will still function for bridge defects with quite low resistances. It is not known a priori whether a particular bridge will fail or pass the IC when voltage-based tests are applied.

Different logic combinations affect R_{crit} since pull-up and pull-down strengths vary with the logic input signals. Figure 7.5 shows a 3NAND contending with a 2NAND. The critical resistance values were simulated from a standard cell library for various input combinations, and results are in Table 7.1. The *p*MOS transistor widths were double the *n*MOS transistor widths. The minimum R_{crit} of 150 Ω occurred when a single *p*MOS transistor contended with two *n*MOS transistors in series. The pull-down was only slightly stronger than the pull-up strength for this logic state. The maximum R_{crit} of 1750 Ω occurred for the worst-case contention of three *n*-channel series pull-down transistors (weak) against two parallel *p*-channel pull-ups (strong). This example shows the strong influence of input logic states on electrical response in the presence of a bridging defect. These bridging defects cause weak node voltages or logic failure, but the quiescent power supply current I_{DDQ} is always elevated.

Self-Exercise 7.2

Two inverters have a bridge defect connection at their output terminals. The nominal $K'_n = 100\ \mu\text{A/V}^2$ and inverter current drives (K'_n, K'_p) are matched within 10% of the worst case. What is the range of R_{crit} if $V_{tn} = 0.5$ V, $V_{tp} = -0.5$ V, $V_{\text{TL}} = 0.75$ V, and $V_{\text{DD}} = 1.5$ V?

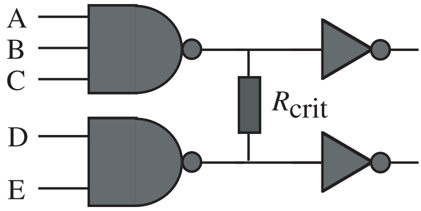


Figure 7.5. Bridging defect between two NAND gate outputs.

Table 7.1. Critical Resistance as a Function of Logic State

A	B	C	D	E	$R_{\text{crit}}(\Omega)$
0	0	0	1	1	950
0	0	1	1	1	725
0	1	1	1	1	150
1	1	1	0	1	1250
1	1	1	0	0	1750

■ EXAMPLE 7.4

Use the short channel *n*MOS transistor curves from Figure 3.34(b) in Figure 7.2(a) to graphically find the critical resistance when that transistor has a bridge defect tied between its drain and V_{DD} rail.

We repeat the transistor curves in Figure 7.6 for convenience. The supply voltage for this technology is 2.5 V, and we will assume that the value for the gate logic threshold voltage is $V_{TL} = V_{DD}/2 = 1.25$ V. To compute the critical resistance value, we must find the resistor load curve that intersects the drain current curve with $V_{GS} = V_{DD}$ at $V_{DS} = 1.25$ V.

We plot a vertical line at $V_{DS} = 1.25$ V and look at the intersection with the topmost drain current curve (corresponding to $V_{GS} = V_{DD}$). The intersection point ($V_{DS} = 1.25$ V, $I_{DS} = 12.8$ mA), gives the drain current that will pass through the *n*MOS device and the resistor short for the critical resistance value. We can compute this critical resistance from the slope of the line joining the $V_{DS} = V_{DD}$, $I_{DS} =$ 0 point to the intersection point found ($I_{DS} = 12.8$ mA), or from the Ohm's law using the obtained current value, i.e:

$$R_{\text{crit}} = \frac{0.5V_{DD}}{I_D(V_{GS} = V_{DD}, V_{DS} = V_{DD}/2)} = \frac{1.25\text{ V}}{12.8\text{ mA}} = 97.6\ \Omega$$

■

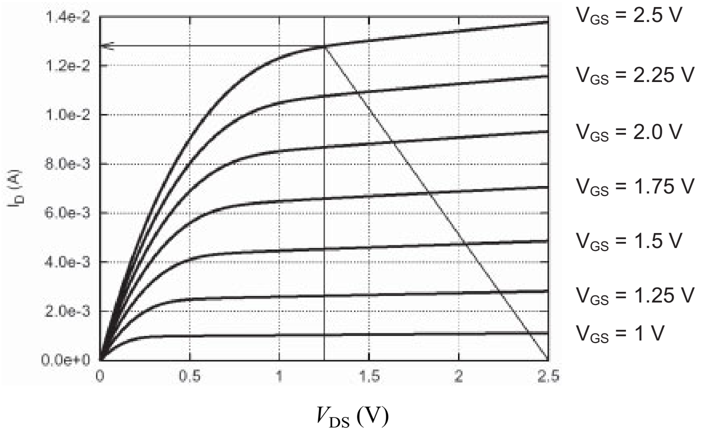


Figure 7.6. Transistor curves and resistor load to compute the critical resistance.

7.2.2 Fault Models for Bridging Defects on Logic Gate Nodes (BF)

We saw that when a bridge defect is activated, it causes intermediate voltages at the shorted nodes. An accurate description of the induced behavior may use analog-based circuit simulators, such as SPICE, to determine the operation region of the affected transistors, the intermediate voltages at the short nodes, and the induced power supply current increase. A test and diagnosis goal is to predict the faulty behavior of the circuit to find appropriate circuit stimuli (called test vectors) that can expose the bridging fault (BF).

Most circuit test pattern generators derive test vectors from a logic description (net list) of the IC. Unfortunately, logic circuit simulators cannot describe the behavior induced by a BF since the altered node voltages may not fit defined logic values. Analog simulators take too long to calculate the state of large circuits, and in many cases a detailed circuit analysis by analog simulators is not required. Several methods try to overcome this difficulty using logic fault models. A logic fault model is an approximation at the logic level of the behavior induced by a defect within an IC.

Generally, fault models try to overcome the gap between the logic description and the analog behavior induced by defects in the circuit. Since logic fault models are widely used in industry to calculate test vectors, we describe some bridging fault models intended for bridge defects at the logic gate I/O level. These models include:

- Stuck-at
- Pseudo stuck-at
- Logic-wired AND/OR
- Voting
- Biased voting

A brief view of each logic model will be given with more details found in [1, 8, 10, 21].

Stuck-at Fault Model (SAF). The stuck-at model, further discussed in Chapter 10, is the simplest and most used logic fault model in the industry. It came from the bipolar transistor IC era, and was accurate for that technology. Many authors showed that stuck-at faults are inadequate for CMOS technology, but its substitution by more accurate models has been slow, given the SAF's easy computational efficiency and its established practice.

Stuck-at fault models applied to bridging defects at the logic level assume that one signal node is permanently tied to the power rail or to ground, and is therefore referred to as stuck-at–1 or stuck-at–0. As shown in previous examples, the SAF model can only quantify detection of low-resistance (sub-R_{crit}) BFs between a signal node and the power/ground rails. Literally, the SAF models a zero Ohm bridge defect to one of the power rails. The inefficiency of the SAF model in detecting BFs motivated the search for more accurate logic fault models.

Pseudo Stuck-at Fault Model. This fault model, initially proposed in [7], exploits the leakage current mechanism induced by bridging faults to simplify the computation effort for ATPG and increase fault coverage. Schematically, this fault model targets stuck-at faults at primitive logic gate inputs, and considers a defect to be detected if its effect is propagated to the output of such a gate. The effect of the fault does not need to be propagated to the IC primary outputs, since it will be detected at the circuit power supply pin by measuring the quiescent current. The advantage of this model is at the computational level, since only small changes to conventional stuck-at tools are required to adopt this fault model.

Logic-Wired AND/OR Model. Logic-wired models for BFs were taken from bipolar technologies (ECL and TTL), in which defective shorted logic gate outputs are logically equivalent to a logic OR or logic AND gate. This behavior appears in technologies where one of the logic levels is always stronger than the other. Therefore, when two nodes are shorted, the stronger node overrides the weaker. Although this logic fault model is more versatile than the stuck-at model, it is not well suited for CMOS technologies, since CMOS ICs have no logic value always stronger than the other. We know that the voltage at a shorted node depends on the relative sizing of the gates involved, the bridge resistance, and the input logic states. An experimental study of wired-logic BFs showed poor correlation between this fault model and real defects [3]. A detailed analysis of logic wired fault models for BFs is in [1].

Voting Model. The voting model was a step forward in modeling BFs [2]. It observes that when shorted nodes are driven to opposite voltages, there is a competition between conducting *p*MOS and *n*MOS transistors. The model assumes that the set of drivers having the largest driving strength (i.e., those driving more current) will decide the final logic value.

The voting model does not account for the nonzero bridging resistance value or the logic threshold voltage of the fan-out gates. Nonzero BFs allow the output of the shorted gates to be at different voltages (because of the voltage drop at the bridge), and therefore be interpreted as different logic values for the subsequent gates. Maxwell showed that two logic gates whose inputs are shorted could interpret the bridge-defect-related analog voltage differently [8].

Biased Voting. This model overcomes the limitations for circuits with variable logic gate thresholds [8]. The biased voting model finds the conductances of the transistors involved in the bridge, taking into account the particular voltage of the bridged node. The voting model assumes a fixed initial voltage to calculate the conductance of the involved transistors. Therefore, the biased voting model accounts for the nonlinear transistor characteristics in calculating the driving strength of a device. Additionally, it takes into account the different thresholds of the logic gates connected at the shortened outputs.

Mixed Description. Since the electrical behavior induced by a bridge defect is analog, an accurate model of the induced behavior uses a mixed description. The whole circuit is described logically, except for the fault site that is described with an analog simulator. This method, described in [13], joins the accuracy of analog simulators for the defect site with the efficiency of simulating large circuits with logic-based tools.

7.3 GATE OXIDE SHORTS (GOS)

This section describes a form of intratransistor bridge defect caused by hard transistor oxide breakdown from particles or oxide imperfections. Chapter 6 described oxide wearout and rupture, and hot-carrier degradation of oxide material that inherently had no defects. The particle-induced oxide failures described here are found at production test, or during the infant mortality phase of the product life cycle.

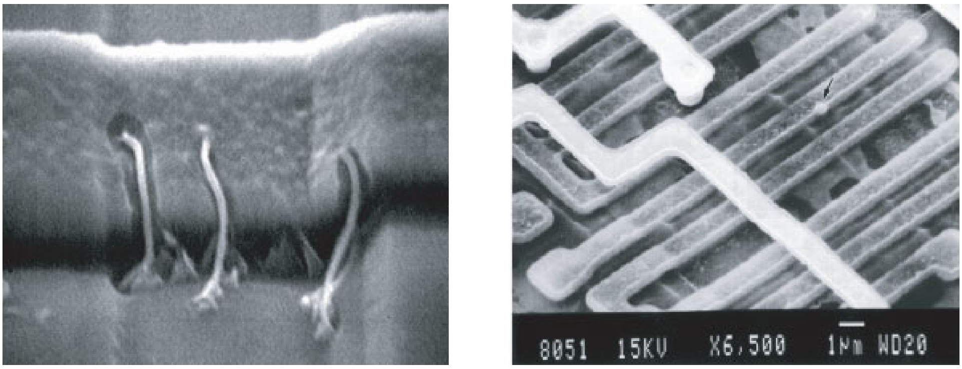
Gate oxide shorts (GOS), or gate shorts, have troubled MOS technology since its beginning in the mid-1960s. The thin oxide of the MOSFET is the control region in which a transistor modulates charge population in the channel. A gate oxide short is a rupture in the thin silicon dioxide (SiO_2) between the polysilicon gate and any of the silicon structures beneath the oxide. The undamaged regions of the thin oxide generally still show normal charge inversion. In some cases, the transistor may still support functionality, although I_{Dsat} may be degraded.

Figure 7.7 shows two forms of gate oxide shorts. Figure 7.7(a) shows thermal filament growth on the gate edge, caused by high overvoltage on the gate. The electric field is higher on the edge of the gate, causing breakdown, and filament growth between gate and source. Figure 7.7(b) shows a gate short to the *p*-well caused by a small particle. The electrical response is different for the two types of gate shorts.

7.3.1 Gate Oxide Short Models

Figure 7.8 shows an inverter cross section in an *n*-well technology. Gate drain/source oxide shorts have simple electrical models, depending on the relative doping of the shortened terminals. There are six places where a gate short acquires a distinct parasitic connection when the gate material merges with the substrate material. Since gate oxide shorts connect the gate polysilicon with the drain, source, or bulk of the device, the electrical properties of the contact depend on the doping type of the terminals being shorted. If the gate and diffusion are of the same doping type, then the electrical model is a resistor between both terminals. If the shorted region has opposite doping, the electrical model is a *pn* junction diode. A detailed description of all the possibilities is given next.

n*MOS Transistor Gate–Drain/Source Oxide Shorts.* An Ohmic connection forms in an *n*MOS transistor when the rupture is from the *n*-doped polysilicon gate to the *n*-doped drain or source (Figure 7.8) [4, 19]. The result is similar to that of an external resistor connected between the gate to drain/source terminals. Figure 7.9 shows an I-V



(a) (b)

Figure 7.7. (a) Thermal filaments across gate to source. (b) Particle-induced gate short from gate to *p*-well [4, 19].

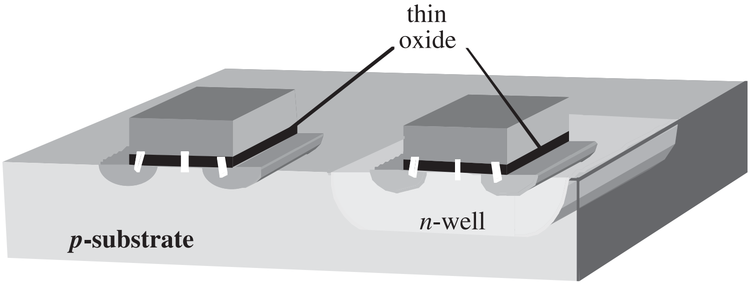


Figure 7.8. Possible GOS defects in a CMOS inverter.

curve (a) taken with probes placed across the gate and source terminals. This signature is an Ohmic gate short between the *n*-doped gate to *n*-doped source region. These forms of gate shorts can result from weak oxides or, preferentially, because electric fields are higher at the edges of the gate structure than in the middle. Typical resistances for *n*MOS transistor gate drain/source shorts formed in the lab ranged from about 1 kΩ to 20 kΩ, putting them above most critical resistances. The gate short resistance in Figure 7.9, curve (a), is about 20 kΩ.

n*MOS Transistor Gate–Substrate Oxide Shorts.* An *n*MOS transistor gate short between the *n*-doped polysilicon gate to substrate doping results in a diode with its cathode, or negative end, at the gate. Under normal positive biasing on the gate, this parasitic diode is reversed-biased and might never conduct, but its depletion region behaves as an additional "parasitic" drain diffusion region, taking electrons from the channel. The dam-

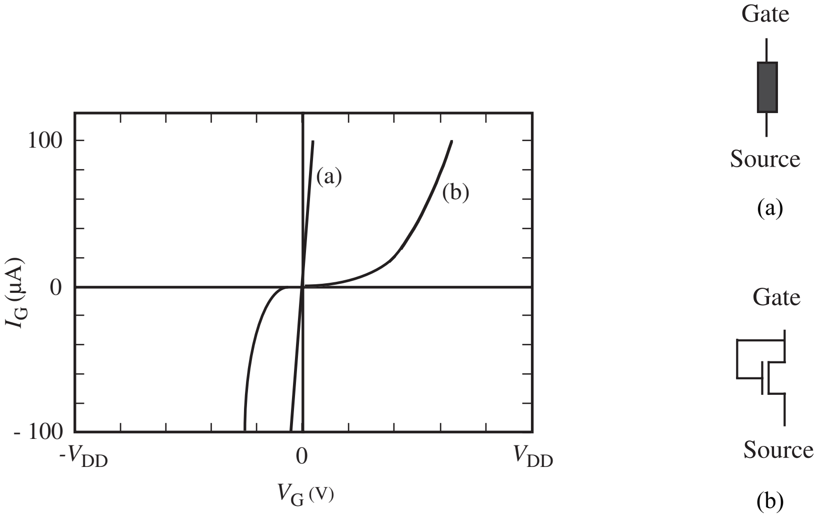


Figure 7.9. I-V curves for different parasitic elements forms of *n*MOS transistor gate oxide shorts. (a) Gate to source. (b) Gate to *p*-well short [sketched from 4].

age-induced depletion region is surrounded by an inversion layer of electrons in the nondamaged portion of the transistor. A parasitic *n*MOS transistor forms, with its gate and drain connected (Figure 7.9(b)).

Chapter 3 showed that a transistor with its gate connected to its drain is always in the saturated state when $V_{GS} > V_t$. Therefore the I_G versus V_G curve is always that of a saturated *n*MOS transistor. Curve (b) in Figure 7.9 shows this quadratic curve for a gate short between the *n*-doped polysilicon gate and the *p*-well. Quadrant I is the normal bias operation of the transistor. Quadrant III shows the parasitic diode response in forward bias as the gate voltage is reversed. The defect can be modeled with two parameters describing the defect position within the transistor and its effective resistance [19].

> ***Self-Exercise 7.3***
>
> An inverter has an *n*MOS transistor with a gate short whose I-V properties are shown in Figure 7.9, curve (a). Given $K_n = 200\ \mu\text{A/V}^2$, $K_p = 150\ \mu\text{A/V}^2$, $W/L = 1$, $V_{TL} = 1.25$ V, $V_{tn} = 0.5$ V, $V_{tp} = -0.5$ V, and $V_{DD} = 2.5$ V, show by calculation whether the circuit will functionally fail. Hint: you must include the relevant driving transistor.

p*MOS Transistor Gate–Drain/Source Oxide Shorts.* Short-channel technology changed the *p*MOS transistor polysilicon gate doping from *n*-doped arsenic to *p*-doped boron to reduce short-channel effects. We will review responses of *n*-doped polysilicon gates taken from measurements, and relate them to expected behavior in *p*-doped polysilicon gates.

*p*MOS transistor gate shorts form diodes when the rupture is across an *n*-doped polysilicon gate to a *p*-drain or *p*-source. Such a diode gate short at the source clamps the *p*MOS transistor gate voltage at one diode drop below V_{DD} when a preceding pull-down transistor attempts to pull the gate node to a logic zero. A gate short to the drain is a little more complicated since the diode acts as a nonlinear feedback element from the drain (output node) to the gate (input node). A *p*-doped polysilicon gate short to the drain/source diffusion regions causes an Ohmic short.

p*MOS Transistor Gate–Substrate Oxide Shorts.* When the defect appears between the *n*-doped polysilicon gate and substrate of a *p*MOS transistor, the GOS is a low-resistive Ohmic contact to the device substrate since they have the same doping type [19]. However, when power is applied, the total *p*MOS transistor structure combines with the defect to form a parasitic *pnp* (bipolar) transistor. The parasitic GOS resistance allows base terminal current injection to the *pnp* transistor. The gate current acts as the base current of a bipolar transistor, and the resulting device characteristics mix MOSFET and bipolar transistor current characteristics. When the gate voltage drops toward logic zero, the *p*MOS transistor now provides an impedance path to the previous logic gate *n*MOS transistor. This action raises the *p*MOS transistor gate voltage, weakening its correct signal. Figure 7.10 draws the parasitic bipolar structure.

An important circuit effect occurs within a single-well CMOS structure. The parasitic bipolar device biased by the defect is connected to other parasitic bipolar devices inherent to the structure that may cause latchup (See Chapter 5). Figure 7.11(a) shows a GOS defect in a *p*MOS transistor causing light emission. Figure 7.11(b) shows the subsequent

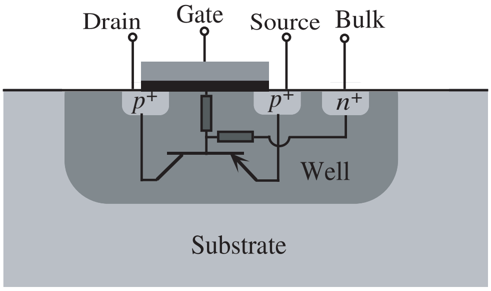


Figure 7.10. Electrical equivalent of a gate–substrate GOS for a *p*MOS transistor.

measured latchup behavior at the circuit level. When sufficient current passes through the latchup structure, a negative I-V slope region rapidly locks the structure into a high-current state.

General Electrical Model Equivalents for Hard Gate Oxide. All combinations of transistor type, defect location, and polysilicon doping type lead to a generalized gate oxide rupture model with 12 subcircuits. The electrical principles are: gate–drain/source shorts form diodes or resistors, depending on whether the relative doping type is the same (short) or opposite (diode). For gate–substrate shorts, defects connecting regions of different doping type create a parasitic MOSFET, whereas shorts between same doping type regions activate the parasitic bipolar transistor. Figure 7.12 summarizes the parasitic electrical elements. The *n*MOS transistor equivalent defect circuits are shown in the first row, and the *p*MOS transistors in the second row.

The *p*-doped polysilicon gate to *n*-well short creates a *pn* junction diode with its anode at the *p*MOS transistor gate. This situation is similar to the *n*MOS transistor gate to *p*-well short, where a parasitic MOSFET is formed. In the *p*MOS transistor gate short, the con-

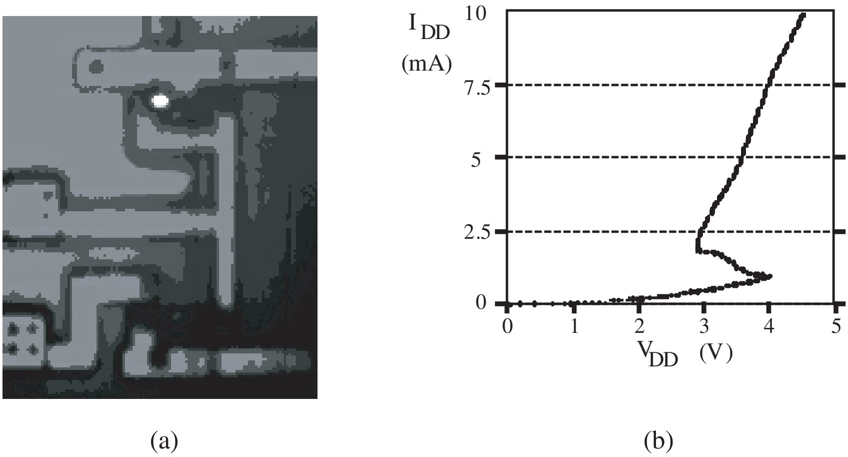


Figure 7.11. (a) A *p*MOS transistor GOS emitting light. (b) Latchup current response at the circuit level.

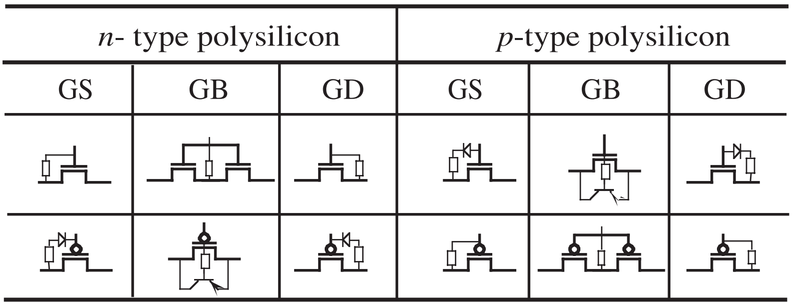


Figure 7.12. Generalized electrical model for gate oxide short defects.

nection forms a parasitic *p*MOS transistor that is always in the saturated state when gate drive is greater than threshold voltage. In all gate short cases, the power supply quiescent current, I_D, is elevated, and logic voltages are either weakened or erroneous. ICs that have GOSs, but still pass functional testing are reliability risks [5].

Soft Gate Oxide Shorts in Ultrathin Oxides. Recent studies show that wearout and breakdown in the ultrathin oxides below 30 Å shows different properties than for the thicker oxides in large-channel-length transistors [23, 24, 25, 26]. The older oxides were well characterized, emphasizing the elevation of I_{DDQ} and the reliability risk when gate shorts passed through the test process. Ultrathin oxides show a soft breakdown in addition to the hard breakdowns of thicker transistor oxides. Soft breakdown is an irreversible damage to the oxide, the most significant effect of which is the increase in noise of the gate signal. I_{DDQ} is not elevated for the soft gate–substrate ruptures of ultrathin oxides. The noise can show up to four orders of magnitude increase after soft breakdown, and this is the only certain evidence of the irreversible damage to ultrathin oxides.

Degraeve et al. found that uniformly stressed 24 Å gates oxides have a uniform area breakdown in transistors [23]. Breakdown over the channel had a high resistance from 10^5–10^9 Ω, whereas breakdowns over the drain and source had resistances of 10^2–10^4. Since the drain and source gate overlap area was much smaller than the area over the channel, most of the breakdowns occurred over the channel. By measuring breakdowns in 200, 180, and 150 nm transistors, they found that the percentage of gate to drain and source breakdowns increased as the transistors became smaller. The drain and source gate overlap area becomes a larger fraction of the transistor gate area. The gate-to-drain or source breakdowns were likely to cause hard failure, whereas the gate-to-channel breakdown showed no failure effects. A 41-stage ring oscillator with seven transistor gate ruptures continued to function with a 15% decrease in oscillator frequency [25]. Part of the frequency reduction was due to hot-carrier damage during a prestress. Significantly, the ring oscillator did not fail despite having several gate oxide breakdowns present.

Henson et al. also found that gate oxide breakdown was more severe at the drain and source than over the channel [24]. They fabricated three channel-length transistors at 4.65 μm, 1.3 μm, and 0.45 μm, where the percentages of gate-to-drain–source breakdowns were 30%, 25%, and 75%. The data show that failure may occur for gate-to-drain or source breakdowns, and the shrinking size of the transistors will begin to shift a greater percentage of breakdowns from the channel region to the critical drain–source region.

Detection of softly ruptured ultrathin oxide at production does not appear possible at this time, nor is the reliability status clear. The normal functioning of the transistor with an ultrathin oxide is not as effected as the killer ruptures of the thicker oxides. The recent ultrathin oxide experiments indicate that test escape and subsequent reliabilities may not be as risky as for breakdown in older technologies. These different properties of the transistor oxide demand more studies at the circuit level to assess the implications of test escapes.

Percolation paths are high-impedance paths in the oxides connecting the gate to the substrate regions. This negates the hard breakdown diode and low resistance gate-diffusion paths. It also replaces parasitic MOSFETs with parasitic bipolar transistors with the path percolation path forming a high-resistance base lead. There is need for more research of these types of shorts at the logic gate level.

Three Other Transistor Node Shorts. There are also drain-to-bulk, source-to-bulk, and gate-to-well (substrate) shorts. These can form *pn* junctions with soft or hard breakdowns between diffusion regions and the well (substrate), or from drain to source. Drain–source shorts can arise from channel punchthrough, or mask particle-related shortened L_{eff}, causing increased leakage.

7.4 BRIDGES IN COMBINATIONAL CIRCUITS

At the circuit level, we distinguish between intragate and intergate BFs. Intragate BFs are shorts between internal nodes of a gate, whereas intergate shorts are between two or more logic gates. This distinction is important for automatic test pattern generation (ATPG) tools since the circuit description (at the gate level or at the transistor level) determines which set of BFs is targeted. The behavior induced by BFs in combinational and sequential circuits is different. BFs in combinational circuits are of two types: nonfeedback bridging faults and feedback bridging faults.

7.4.1 Nonfeedback Bridging Faults

A nonfeedback bridging fault is a short that does not create a logic path between a gate output and a node that can determine the value of any of its inputs. This distinction is made because nonfeedback bridging faults in combinational circuits cannot induce sequential behavior. This is the simplest bridging fault, and corresponds to the BFs discussed to this point. When the defect is activated, the shorted nodes assume intermediate voltages whose values depend on the bridge resistance and the current drive strength of the transistors involved. The logic behavior depends on the critical resistance of the shorted nodes and the logic threshold of fan-out gates.

When the defect is activated, the resulting logic value may or may not be correct, but the power supply current will be elevated. This significantly impacts the coverage achieved with different voltage-based test techniques, as explained in Chapter 10.

7.4.2 Feedback Bridging Fault

A feedback bridging fault is a short in which some logic input combinations to the circuit create a logic path from a gate output to one of its inputs. Specifically, a BF between two

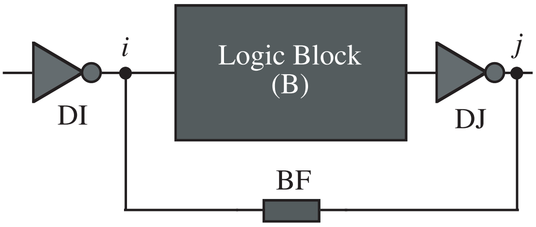


Figure 7.13. Feedback bridging fault.

nodes *i* and *j* is a feedback bridging fault if the logic value of *j* depends on the logic value of *i*. Node *i* is called *predecessor,* while node *j* is called *successor.* Figure 7.13 shows such a bridge between inverters DI and DJ. The output of DJ is connected to the output of inverter DI, and this may determine the logic state of DJ.

Feedback bridging faults are complex, since they can induce sequential behavior in a combinational circuit. The behavior induced by a feedback-bridging fault depends on the logic connecting the shorted gates (Block B in Figure 7.13). For a given input to the circuit, three situations may appear:

Case A. The input to the circuit is such that the logic value of node *j* is independent of the logic value of node *i*.

Case B. The logic value of node *j* is equal to the logic value of node *i*.

Case C. The logic value of node *j* is opposite to the logic value of node *i*.

The first case is equivalent to a nonfeedback bridging fault since the logic values of the shorted nodes are not related. The second case creates a noninverted feedback bridging fault, whereas the third is referred to as an inverted feedback bridging fault. The two latter cases are discussed after an example.

■ EXAMPLE 7.5

Determine the logic values of A, B, and C in Figure 7.14 that categorize the circuit as Case A, Case B, or Case C. Assume that the NAND gate is always stronger than the inverter.

We redraw the circuit to better identify the role of each gate in the circuit (Fig-

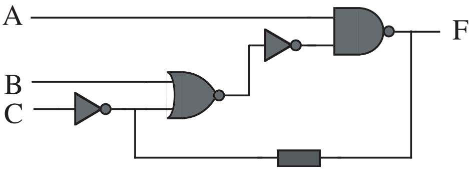


Figure 7.14. A combinational circuit with a BF.

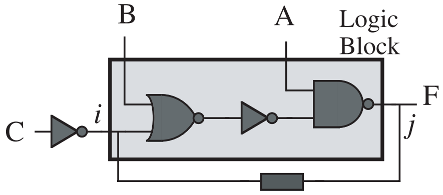


Figure 7.15.

ure 7.15). Node i is the output of the first inverter, whereas node j is the output of the NAND gate (predecessor and successor, respectively).

Case A. When A = 0, the NAND gate output is $j = 1$, independent of any other input to the circuit. When B = 1, the output of the NOR gate is 0 for all value of i, thus making the value of j again independent of i. These two conditions cover 6 of the 8 possible combinations.

Case B. In the fault free case, j can be put in terms of i: $j = \overline{A} + (\overline{B} \cdot \bar{i})$. There is no condition under which $j = i$.

Case C. From the expression derived in the previous case, when A = 1 and B = 1, then $j = \bar{i}$. This covers the two remaining cases. ■

Noninverted Feedback Bridging Fault. These faults were traditionally thought to be redundant, since the shorted nodes are at the same logic level. This is only true when the strength of the predecessor gate (DI in Figure 7.13) is much stronger than the driving capability of the successor gate. If the successor gate (DJ) has a stronger driving capability than the predecessor gate (DI), then the value of node i is determined by DJ. If both drivers have the same logic output, then there is no conflict, and the defect is not activated. If DI makes a transition against the stronger DJ gate, then node i will not change and a competition between gates will appear. This causes an intermediate voltage and a quiescent supply current elevation. The situation can be described as a latched state.

A subtle effect appears when gates DI and DJ have similar driving capabilities, with DI slightly stronger than DJ. This case causes a significant delay because of a transient behavior appearing while the drivers are competing to set the final node value.

Inverted Feedback Bridging Faults. Inverted feedback BFs appear when the logic values of the shorted gates are opposite. Two behaviors may occur, depending on which gate is stronger. When the gate driving the predecessor (gate DI in Figure 7.13) is stronger than the successor (gate DJ), then the defect causes a logic error and current elevation, since driver DI overrides the output of DJ.

If the successor gate is stronger than the predecessor one, then the defect causes oscillation in the circuit. The period of oscillation is related to the delay of the logic connecting the predecessor output to the successor (logic block B in Figure 7.13).

Figure 7.16 shows experimental data from a test circuit. When the input was at logic zero, then the predecessor dominated and no voltage oscillations were observed. When the circuit input was set to a high logic value, the successor driver was dominant, causing oscillations.

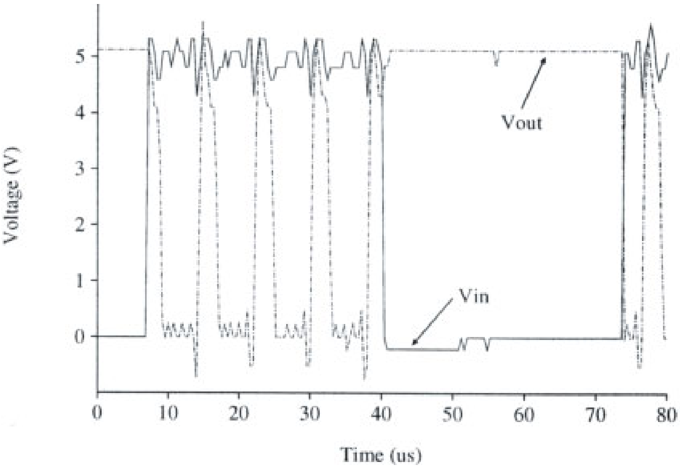


Figure 7.16. Oscillating behavior caused by an inverted feedback BF.

7.5 BRIDGES IN SEQUENTIAL CIRCUITS

The electrical behavior induced by BFs in sequential circuits may differ from combinational ones, since the conditions to detect a BF in combinational circuits may not hold for sequential ones. This depends on the circuit topology and defect site location. Lee introduced the concept of *control loops* to designate groups of transistors that control each other to create memory elements [6]. A control loop is said to be in a *floating state* when it is in a memory state, so that its value cannot be changed by the driving logic. If the memory element is driven by its preceding logic, then it is said to be in a *forced state.* In some situations, a control loop in a floating state may prevent quiescent current elevation since the internal nodes cannot be forced to the values required to elevate this magnitude. Since these defects swap memory values, they may be detected with logic-voltage-based testing, depending on the circuit implementation (shown in Chapter 10). The next section discusses the effect of BFs in flip-flops as the basic memory element in sequential circuits, and also BFs in semiconductor memories.

7.5.1 Bridges in Flip-Flops

The behavior induced by BFs in flip-flops depends on the circuit implementation. Flip-flops can be designed with a standard NAND gate configuration, but when using CMOS technology, tri-state inverter gates or pass gate transistor-based designs are often used. The induced BF behavior differs with the design.

Rodriquez et al. showed that flip-flops implemented with tri-state inverters and pass gate transistors may exhibit behavior in which the defect is not current-testable because of the floating state control loop [15]. These BFs change the logic state of the memory element and are therefore logically testable. Sachdev proposed a design modification using an additional inverter to make these cells current-testable [16].

Metra et. al. showed that flip-flops may have BFs that do not change the logic behavior of the gate, and do not elevate the quiescent current of the circuit [9]. The defect impacts the timing parameters of the cell, changing the set-up or hold time of the cell.

7.5.2 Semiconductor Memories

Bridging defects in SRAMs were analyzed by several researchers [12, 16, 22]. Bridging defects at SRAM cells may or may not cause quiescent current elevation. The main reason for not elevating the quiescent current is similar to the effect in flip-flops, and is due to floating control loops. BFs between cells are not generally current-testable because the cells that are not accessed do not maintain their logic value, and do not create a logic conflict required to elevate quiescent current. These defects may be detected using logic testing.

Sachdev showed that internal timing restrictions inherent to asynchronous operations in the memory may prevent current elevation for a desired period of time [17]. This significantly impacts current-based test techniques since the time window required to measure the quiescent current is not guaranteed. The proposed solution requires design modifications to improve current-testability coverage.

A detailed analysis of the behavior induced by gate oxide shorts in SRAMS was given in [20]. Although gate oxide shorts (GOS) are inherent to a single transistor, they may involve more than one single SRAM cell. GOS defects are responsible for many failures in RAM cells such as data retention (a memory cell looses its value a short time after it was written), coupling faults (more than one cell is accessed simultaneously), sense amplifier recovery (the memory sense amplifier does not respond at the normal frequency of operation), and stuck-at behavior.

Sachdev analyzed BFs in DRAMs [17]. DRAM cells made of a storage capacitor (typically 50 fF) need refreshing because of leakage currents. BFs in these cells cause logic errors, but the quiescent current is not elevated since there is no voltage stress or drive. Moreover, since the charge stored in a DRAM cell is very small, even a high-impedance defect can seriously affect the storage capability since there is no feedback structure to compensate for the defect-induced leakage.

7.6 BRIDGING FAULTS AND TECHNOLOGY SCALING

Technology scaling does not have a direct impact on the bridging defect's inherent characteristics. The physical mechanisms for these defects that appear during IC manufacturing or operation generally remain invariant. The same holds for the electrical properties of the short.

Since the effect of a short on IC behavior is related not only to the defect itself, but also to the surrounding circuitry characteristics, the impact of these defects is expected to change with technology scaling. The main parameters affected are the critical resistance and the quiescent current detectability of the defect. As technology scales down, the driving strength of transistors increases. Therefore, following the analysis made at the initial part of the chapter, the critical resistance values are expected to decrease, thus making logic testing even less effective in detecting BFs.

On the other hand, traditional quiescent current testing based on a single pass/fail threshold (this technique will be introduced in detail in Chapter 10) is becoming less effective in submicron technologies due to elevated current leakage values and variability.

From this perspective, BFs are expected to be a challenge in submicron technologies since more sophisticated techniques based on monitoring the delay induced by this defect or variations in current leakage will be required.

7.7 CONCLUSION

This section described the many variables affecting the response of a circuit with a bridge defect. The critical resistance concept is perhaps the most important, as it explains why flagrant visual bridging defects may not cause a functional test of the circuit to fail. Bridges between logic gate transistor nodes and bridges between signal nodes of two or more gates have slightly different responses. Gate oxide shorts are a major source of the intratransistor node bridges. They have linear and nonlinear properties depending on relative doping levels and location of the gate short. The bridge models developed to assist the test vector generation problem were described. Finally, the behavior of bridging defects in memory circuits was explained. Bridging defects cause I_{DDQ} elevation and either a correct but weakened node voltage or functional failure occurs. An exception to I_{DDQ} elevation is found in certain forms of memory bridges.

REFERENCES

1. M. Abramovici, M. Breuer, and A. Friedman, *Digital Systems Testing and Testable Design,* IEEE Press, 1993.
2. J. Acken and S. Millman, “Accurate modelling and simulation of bridging faults,” in *Custom Integrated Circuits Conference (CICC),* pp. 17.4.1–17.4.4, 1991.
3. R. Aitken, “Finding defects with fault models,” in *IEEE International Test Conference (ITC),* pp. 498–505, October 1995.
4. C. Hawkins and J. Soden, “Electrical characteristics and testing considerations for gate oxide shorts in CMOS ICs,” in *IEEE International Test Conference (ITC),* pp. 544–555, November 1985.
5. C. Hawkins and J. Soden, “Reliability and electrical properties of gate oxide shorts in CMOS ICs,” in *IEEE International Test Conference (ITC),* pp. 443–451, September 1986.
6. K. J. Lee and M. Breuer, “Design and test rules for CMOS circuits to facilitate I_{DDQ} testing of bridging faults,” *IEEE Transactions on Computer-Aided Design, 11,* 5, 659–670, May 1992.
7. Y. K. Malaiya and S. Y. H. Su. “A new fault model and testing technique for CMOS devices,” in *IEEE International Test Conference (ITC),* pp. 25–34, October 1982.
8. P. Maxwell and R. Aitken, “Biased voting: A method for simulating CMOS bridging faults in the presence of variable gate logic thresholds,” in *IEEE International Test Conference (ITC),* pp. 63–72, October 1993.
9. C. Metra, M. Favalli, P. Olivo, and B. Ricco, “Testing of resistive bridging faults in CMOS flip-flop,” in *European Test Conference,* pp. 392–396, 1993.
10. S. Millman, E. McCluskey, and J. Acken, “Detecting bridging faults with stuck-at test sets,” in *IEEE International Test Conference (ITC),* pp. 773–783, October 1990.
11. T. Miller, J. Soden, and C. Hawkins, “Diagnosis, analysis, and comparison of 80386EX I_{DDQ} and functional failures,” *IEEE I_{DDQ} Workshop,* Washington DC, October 1995.
12. Naik, F. Agricola, and W. Maly, “Failure analysis of high density CMOS SRAMs using realistic defect modeling and I_{DDQ} testing,” in *IEEE Design & Test of Computers,* pp. 13–23, June 1993.

13. J. Rearick and J. Pate, "Fast and accurate CMOS bridging fault simulation," in *IEEE International Test Conference (ITC),* pp. 54–60, October 1993.

14. R. Rodriguez-Montanes, E. Bruls, and J. Figueras, "Bridge defects resistance measurements," in *IEEE International Test Conference (ITC),* pp. 892–899, September 1992.

15. R. Rodriguez and J. Figueras, "Analysis of bridging defects in sequential CMOS circuits and their current testability," in *European Design and Test Conference,* pp. 356 –360, 1994.

16. M. Sachdev, "I_{DDQ} and voltage testable CMOS flip-flop configurations," in *IEEE International Test Conference (ITC),* pp. 534–543, October 1995.

17. M. Sachdev, "Reducing CMOS RAM test complexity with I_{DDQ} and voltage testing," *Journal of Electronic Testing: Theory and Applications, 6,* 2, pp. 191–202, 1995.

18. J. Segura, V. Champac, R. Rodríguez, J. Figueras, and A. Rubio, "Quiescent current analysis and experimentation of defective CMOS circuits," *Journal of Electronic Testing: Theory and Applications, 3,* 337–348, 1992

19. J. Segura, C. De Benito, A. Rubio, and C. Hawkins, "A detailed analysis of GOS defects in MOS transistors: Testing implications at circuit level," in *IEEE International Test Conference (ITC),* pp. 544–550, October 1995.

20. J. Segura, C. De Benito, A. Rubio, and C. Hawkins, "A detailed analysis and electrical modeling of gate oxide shorts in MOS transistors," *Journal of Electronic Testing: Theory and Application, 8,* 229–239, 1996.

21. T. Storey and W. Maly, "CMOS bridging fault detection," in *IEEE International Test Conference (ITC),* pp. 842–850, October 1990.

22. H. Yokoyama, H. Tamamoto, and Y. Narita, "A current testing for CMOS static RAMs," in *IEEE International Workshop on Memory Technology, Design And Testing,* pp. 137–142, August 1993.

23. R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, "Relation between breakdown mode and breakdown location in short channel NMOSFETs," in *International Reliability Physics Symposium (IRPS),* pp. 360–366, May 2001.

24. W. Henson, N. Yang, and J. Wortman, "Observation of oxide breakdown and its effects on the characteristics of ultra-thin *n*MOSFET's," *IEEE Electron Device Letters, 20,* 12, 605–607, December 1999.

25. B. Kaczer, R. Degraeve, G. Groseneken, M. Rasras, S. Kubieck, E. Vandamme, and G. Badenes, "Impact of MOSFET oxide breakdown on digital circuit operation and reliability," in *IEEE International Electron Device Meeting,* 553–556, December 2000.

26. J. Segura, A. Keshavarzi, J. Soden, and C. Hawkins, "Parametric failures in CMOS ICs—A defect-based analysis," in *IEEE International Test Conference (ITC),* October 2002.

27. J. Suehle, "Ultrathin gate oxide reliability: Physical models, statistics, and characterization," in *IEEE Transactions on Electron Devices, 49,* 6, 958–971, June 2002.

28. B. Weir et al., "Ultra-thin gate dielectrics: they break down, but do they fail?," in *International Electron Device Meeting (IEDM),* pp. 73–76, 1997.

EXERCISES

7.1. A 2 kΩ defect resistance connects the inverter output to ground (Figure 7.17). $K_n' = 200\ \mu\mathrm{A/V}^2$, $K_p' = 100\ \mu\mathrm{A/V}^2$, $V_{tn} = 0.6$ V, $V_{tp} = -0.6$ V, and the logic threshold voltage is $V_{\mathrm{TL}} = 1.5$ V. Will the circuit logically fail?

7.2. (a) As V_{DD} is reduced, explain why the critical resistance increases and how this relates to more or less defect detection sensitivity by voltage-based testing.

(b) As temperature is reduced, carrier mobility increases and normal circuits run

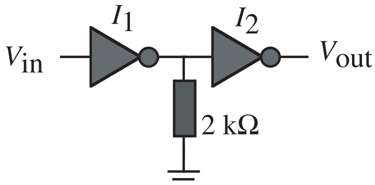


Figure 7.17.

faster. Assume that the defect does not change resistance, What affect would lowering temperature have on the values of critical resistance. Is the defect easier to detect at low temperature for any given value of V_{DD}?

7.3. Figure 7.1(a) shows a metal sliver between two metal interconnections.

(a) Calculate the sliver resistance if the height, width, and length are 0.5 μm, 0.4 μm, and 1 μm. Assume that an Al sliver has a resistivity of 3.4 μΩ · cm.

(b) A physical measurement of the resistance between the two metals interconnects gives a resistance of about 500 Ω. Why is there a discrepancy between the calculated and measured resistance?

(c) What is the impact on detection of this defect by a conventional voltage test that tests for functionality?

7.4. Repeat the calculation for the critical resistance of the circuit shown in Figure 7.4(a) when the pull-up and pull-down drive strengths are closer. Let K_p = 300 μA/V^2 and K_n = 310 μA/V^2 when $W/L = 1$, $V_{tn} = 0.35$ V, $V_{tp} = -0.35$ V, $V_{\text{DD}} = 1.2$ V, and $V_{\text{TL}} = 0.6$ V. Comment on the significance of this.

7.5. (a) An integrated circuit is suspected of having a hard gate oxide short in an *n*-channel transistor. With measurements at the IC pin level, can you distinguish whether this short is to the channel or diffusion regions (drain, source)?

(b) Repeat (a) for an ultrathin-oxide transistor.

7.6. (a) When a *n*MOSFET hard gate short occurs between a *n*-doped gate to a *p*-doped substrate or a *p*MOSFET *p*-doped gate to a *n*-doped substrate, then a parasitic MOSFET appears. When power is applied to an IC with such a short, explain why these parasitic transistors are in either the saturated or nonsaturated state bias.

(b) What is the effect when a soft-gate oxide rupture occurs on an ultrathin oxide?

7.7. (a) Assume that all transistors in the gates of Figure 7.14 have $K_n(W_n/L_n) = K_p(W_p/K_p)$, and $V_{tn} = -V_{tp}$, except for the NOR gate that has $V_{\text{TL}} = V_{\text{DD}}/3$ for the input driven by the inverter. If the defect resistance is negligible, determine which input combinations will lead to oscillations. *Hint: the circuit needs to be in Case C.*

(b) If the transistor sizes of the inverters are doubled with respect to (a), and for all devices V_t is 20% of V_{DD}, compute the critical resistance that will lead the inverter override the NAND output.

7.8. Compute the critical resistance trend for the circuit in Exercise 7.4 when scaling the technology if V_{DD} is reduced by 16%, V_t's are reduced by 28.55%, and the transistor drive current (K_n and K_p) increases by 15%.

7.9. An IC shows no functional failure, but I_{DDQ} is abnormally elevated for every test vector measurement. What type of defect might this indicate? What are the reliability implications?

7.10. A functionally failing IC shows abnormal I_{DDQ} for certain test vectors. What might you conclude about this IC?

CHAPTER 8

OPEN DEFECTS

8.1 INTRODUCTION

Open circuit defects are unintended breaks or electrical discontinuities in IC interconnect lines occurring in metal, polysilicon, or diffusion regions.* Figure 8.1 shows two CMOS open circuit defects. Figure 8.1(a) is a via not well bonded to the metal liner in the via hole. Figure 8.1(b) shows locations in the metal where two vias are missing. This caused an open circuit to two inverter logic gates. Open defects have greater behavioral complexity than bridge defects.

The major open circuit defect variables are

- Size of the open defect. Is the crack wide or narrow?
- Defect location:
  —Open gate to a single transistor
  —Open drain or source
  —Open to a logic gate input affecting a CMOS complementary transistor pair
- Open on a metal line driving several gates
- Capacitive coupling of open node to surrounding circuit nodes

Deep-submicron CMOS technologies use metal line widths of 130 μm or less and via height-to-width ratios of more than 5:1. These dimensions, when coupled with IC via

*In the context of this chapter, an open is a complete disconnect. Resistive or weak opens are described in Chapter 9 as extrinsic parametric failures.

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

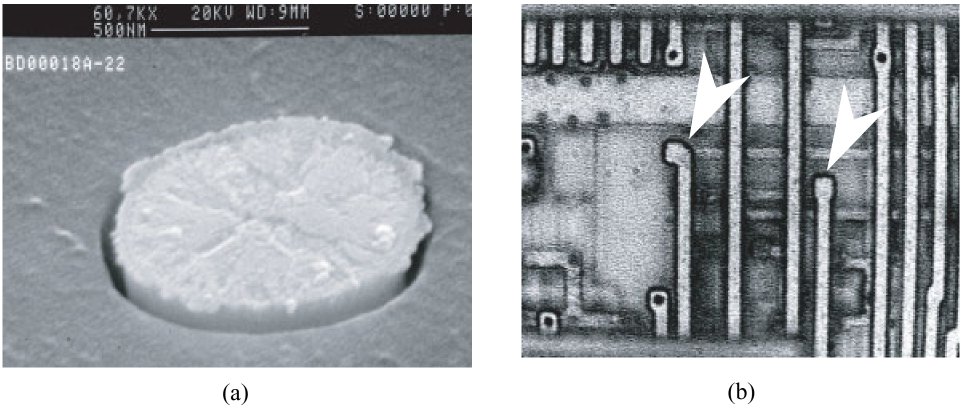


Figure 8.1. (a) A resistive open with poor bonding between the via metal and via liner (reproduced by permission of Bruce Draper of Sandia National Laboratories), (b) Missing vias (arrows) [6].

counts from hundreds of millions to over a billion and total metal lengths of several kilometers, make via- and contact-related open defects more probable than before. Open defects are unavoidable, and their detection is sometimes nearly impossible. We will start by modeling the behavior of a floating node within an IC, and then analyze its impact on circuit behavior.

8.2 MODELING FLOATING NODES IN ICs

The main effect of an open IC signal line is that one circuit node is no longer driven by any gate, but may be left in a floating or high-impedance (high-Z) state. The node does not have a conducting path to V_{DD} or ground through a low impedance connection. The voltage on the floating node depends on the properties and topology of the surrounding circuitry. Two primary variables determine the final voltage value of a floating node: (1) the size of the crack and (2) the amount of charge present at the floating node.

The size of the crack determines if electrons can tunnel across the open, thereby controlling the amount of charge injected from the original driver (the gate that should drive the node in the fault-free circuit) toward the floating node. The charge at the floating node also depends on the capacitive coupling to the surrounding nodes, and the charge at the gate and drain terminals of the transistors to which the node may be connected. It is important to emphasize that the complete problem is often a complex combination of all these factors. We will describe each effect and then summarize with a model that includes all the effects.

8.2.1 Supply–Ground Capacitor Coupling in Open Circuits

Capacitor voltage dividers appear in some MOSFET open circuits. Figure 8.2(a) shows an open on a metal line over a field oxide (floating node V_2). Part of the floating metal line runs over the IC substrate (tied to ground), and part of it runs over the well area (connected to V_{DD}). The metal–oxide–semiconductor structure is a capacitor, so that the floating

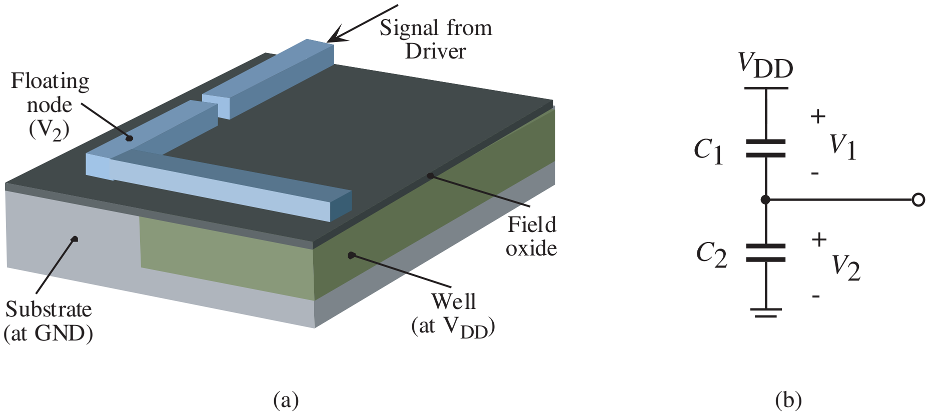


Figure 8.2. (a) Open crack in a metal line and (b) its electrical equivalent: a capacitor voltage divider.

node is capacitatively coupled to ground and supply. The values of the coupling capacitors to V_{DD} and to ground depend on the length of the metal track running over the well and the substrate.

The equivalent circuit of the floating metal has two capacitors in series (Figure 8.2(b)). The voltages V_1 and V_2 are functions of V_{DD} and the values of C_1 and C_2. From Chapter 1, we review

$$C_1 = \frac{Q}{V_1} \quad \text{and} \quad C_2 = \frac{Q}{V_2} \tag{8.1}$$

then

$$V_{DD} = V_1 + V_2 = \frac{Q}{C_1} + \frac{Q}{C_2} = Q\left(\frac{1}{C_1} + \frac{1}{C_2}\right) \tag{8.2}$$

Substituting

$$V_{DD} = V_2 C_2\left(\frac{1}{C_1} + \frac{1}{C_2}\right) \tag{8.3}$$

or

$$V_2 = V_{DD}\left(\frac{C_1}{C_1 + C_2}\right) \tag{8.4}$$

In general, the voltage at the floating node can be expressed as

$$V_2 = \alpha V_{DD} \tag{8.5}$$

where α is a constant ranging between 0 and 1.

Self-Exercise 8.1

In Figure 8.2, (a) If $V_{\text{DD}} = 2.8$ V, $C_1 = 10$ fF, and $C_2 = 18$ fF, calculate V_2. (b) If $V_2 = 0.768$ V, $C_2 = 25$ fF, and $C_1 = 11$ fF, calculate V_{DD}.

The circuit in Figure 8.2(b) is used in open circuit defect analysis. For example, when the transistor gate is open, it is in a high-impedance state but the parasitic capacitive environment can often be reduced to a form of Figure 8.2(b). The defective circuit responds to transistor gate voltages set up by the capacitor voltage divider.

8.2.2 Effect of Surrounding Lines

Typically, an IC metal line is surrounded by other conducting lines. When two conducting lines at different metal levels cross over, there is a parasitic capacitance that couples both nodes. The value of this capacitance depends on the metal lines' intersecting area, which in this case is not large. A similar situation occurs for two conducting lines running parallel on the same metal level for a given distance, but in this case the coupling capacitance is bigger.

Figure 8.3(a) shows two metal lines running adjacent on the same metal level (V_{m} and V_{F} nodes), while a third metal line (V_3 node) on a higher level crosses both metals. One of the lower metal lines has an open defect, and it is floating. The coupling capacitance from the adjacent line in the same metal level will be higher than the coupling capacitance to the upper metal, so that the influence of the adjacent line is much stronger.

The equivalent electrical circuit for the floating node in Figure 8.3(a) is shown in Figure 8.3(b). Neglecting the coupling influence of the metal-2 line (which is much smaller than the influence of the metal-1 line), the voltage of the floating node (V_{F}) will depend on the logic value of the metal-1 line (V_{m}). An analysis similar to that developed previously leads to

$$\begin{aligned} V_{\text{F}} &= V_{\text{DD}}\left(\frac{C_1 + C_{\text{m}}}{C_1 + C_2 + C_{\text{m}}}\right) && \text{if } V_{\text{m}} = H \\ V_{\text{F}} &= V_{\text{DD}}\left(\frac{C_1}{C_1 + C_2 + C_{\text{m}}}\right) && \text{if } V_{\text{m}} = L \end{aligned} \tag{8.6}$$

In general, when n nodes with voltage V_i are each coupled to the floating metal line, the final voltage is expressed as

$$V_{\text{F}} = \alpha V_{\text{DD}} + \sum_{i=1}^{n} \alpha_i V_i \tag{8.7}$$

where α and α_i are constants ranging between 0 and 1.

In a real IC, the signal nodes that influence the floating node experience logic transitions, thus changing the floating voltage with time. We can compute the floating voltage after a transition of the influencing metal, using charge conservation. If the influencing metal (V_{m} node) makes a high-to-low transition, then the initial voltage at the floating line before the transition (V_{Fin}) is computed from

$$(C_{\text{m}} + C_1)(V_{\text{DD}} - V_{\text{Fin}}) = C_2 V_{\text{Fin}} \tag{8.8}$$

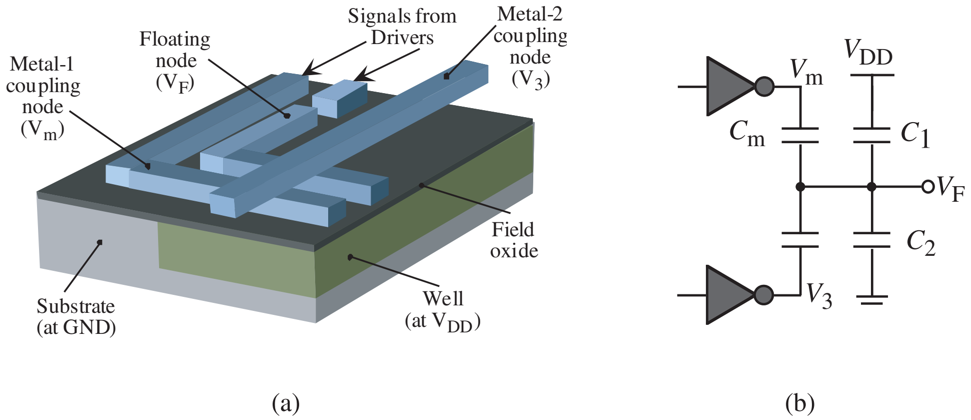


Figure 8.3. (a) Two metal-1 signal lines crossing one metal-2 line. The coupling capacitance of each metal-1 line to the other will be higher than the coupling capacitance to the metal-2 line. (b) Equivalent electrical model with the signal drivers shown as inverters.

which gives

$$V_{\text{Fin}} = \left(\frac{C_{\text{m}} + C_1}{C_{\text{m}} + C_1 + C_2} \right) V_{DD} \tag{8.9}$$

Once the transition is finished, the final voltage at the floating line (V_{Fend}) is computed from

$$C_1(V_{\text{DD}} - V_{\text{Fend}}) = (C_{\text{m}} + C_2)V_{\text{Fend}} \tag{8.10}$$

which gives

$$V_{\text{Fend}} = \left(\frac{C_1}{C_{\text{m}} + C_1 + C_2} \right) V_{DD} \tag{8.11}$$

Subtracting Equation (8.9) from Equation (8.11), the floating voltage change ΔV_{F} is

$$\Delta V_{\text{F}} = \left(\frac{C_{\text{m}}}{C_{\text{m}} + C_1 + C_2} \right) V_{DD} \tag{8.12}$$

Equation (8.12) shows that the influence of the surrounding metals depends proportionally on the coupling capacitance value.

■ EXAMPLE 8.1

Consider the dynamic NAND gate shown in Figure 8.4, where part of the metal-1 interconnect layout is shown (the interconnect between metal-1 and the transistor gate runs in polysilicon and is not illustrated). Determine the output value in the evaluation phase (clock = 1) if transistor N2 input makes a transition from 0

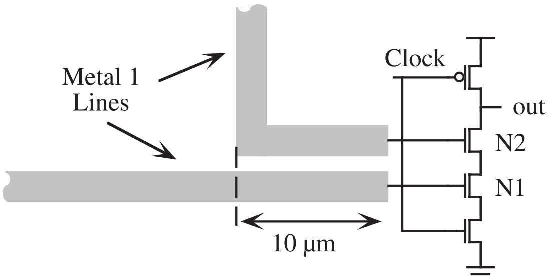


Figure 8.4. Dynamic NAND gate.

to V_{DD} when the metal-1 input to N1 has an open (a) 5 μm away from the polysilicon, (b) 280 μm away from the polysilicon. Discuss the results in both cases.

The metal lines shown are at the minimum distance. Consider that the input gate transistor capacitor and the polysilicon-substrate capacitor are negligible. The circuit is fabricated with a technology that has $V_{tn} = 0.35$ V, $V_{DD} = 1$ V, a metal-1 to bulk capacitance of 0.05 fF/μm, and a metal-1 to metal-1 capacitance of 0.35 fF/μm for minimum distance metal-1 lines.

(a) The metal-1 substrate capacitance of the metal portion connected to the N1 gate is $C_{ms} = 5\ \mu\text{m} \times 0.05\ \text{fF}/\mu\text{m} = 0.25$ fF, whereas the capacitance to the metal line driving the N2 transistor input is $C_{mm} = 5\ \mu\text{m} \times 0.35\ \text{fF}/\mu\text{m} = 1.75$ fF. Assuming that the gate of N1 is initially grounded, the final voltage of the N1 gate input node will be given by the voltage of a capacitor divider between $V_{\text{Dmetal-1 to metal-1}}$ (N2 input) and ground.

$$V_{GN1} = V_{DD}\left(\frac{C_{mm}}{C_{mm} + C_{ms}}\right) = 1\ \text{V}\left(\frac{1.75\ \text{fF}}{1.75\ \text{fF} + 0.25\ \text{fF}}\right) = 0.875\ \text{V}$$

This voltage will turn on N1, since $V_{tn} = 0.35$ V. The circuit function will be correct if the gate driving N1 in the fault-free circuit also makes a low-to-high transition, but an erroneous logic behavior would occur if N1 was off in the fault-free circuit.

(b) The N1 input to ground capacitance is $C_{ms} = 280\ \mu\text{m} \times 0.05\ \text{fF}/\mu\text{m} = 14$ fF, and $C_{mm} = 10\ \mu\text{m} \times 0.35\ \text{fF}/\mu\text{m} = 3.5$ fF, since both metal lines run parallel for 10 μm. The N1 gate voltage after the N2 transition is

$$V_{GN1} = 1\ \text{V}\left(\frac{3.5\ \text{fF}}{3.5\ \text{fF} + 14\ \text{fF}}\right) = 0.2\ \text{V}$$

In this case, N1 will remain off and the gate output will stay high. This will lead to a logic error if N1 input is supposed to switch in the fault-free circuit.

In general, the lower the interconnect level and the farther away the defect from the gate input, the more unperturbed the value of the floating node. ■

8.2.3 Influence of the Charge from MOSFETs

In many cases, the floating node drives the gate of one or more transistors. Renovell and Cambon [8], Champac et al. [1], and Johnson [4] analyzed the device terminal's charge influence on the gate voltage. The charge stored at the gate terminal of a MOSFET has a strong dependence on the coupling from the drain and channel terminals. Hence, the drain voltage plays an important role on the final gate voltage value. Neglecting the effects from other nodes, the floating gate voltage can be expressed as [4]

$$V_{\mathrm{FG}} = \frac{Q_{\mathrm{FG}}}{C_{\mathrm{G}}} + \alpha V_{\mathrm{DS}} \tag{8.13}$$

where Q_{FG} is the charge at the floating gate, C_{G} its capacitance, and α ranges between zero and one. A detailed expression for Q_{FG} can be found in [1].

Experimental results [1, 4] report floating voltage values up to 3 V for a 5 V technology, demonstrating that the device can conduct significant drain current while its gate is floating.

8.2.4 Tunneling Effects

Narrow cracks can appear in metal lines or in a contact or via. Figure 8.5 shows such a defect whose electrical behavior is another unusual feature of open defects. The crack is narrow enough to support quantum mechanical electron tunneling when an applied voltage creates an electric field [3]. Barriers must be on the order of less than 100 Å to support a probability of significant tunneling. Another study proposed tunneling conduction mechanisms caused by incomplete etching or native (room temperature grown) oxidation after etching [5].

Tunneling is a quantum effect by which a small particle can cross through a finite potential barrier as a consequence of the wave–particle duality. There are different tunneling mechanisms, one of which is electric-field dominated. This field-dependent tunneling is called Fowler–Nordheim tunneling, and the current density J_e associated with this mechanism can be quantitatively described by

$$J_e = \alpha \cdot \mathscr{E}^2 \cdot e^{-\beta/\mathscr{E}} \tag{8.14}$$

where J_e is the current density in A/m^2, and α and β are constants depending on the physical properties of the structures through which tunneling takes place. The electric field in the metal void depends on the size of the open, the applied voltage, and the morphology.

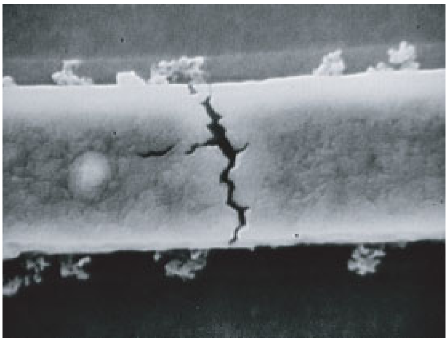


Figure 8.5. SEM photo of metal tunneling open [3].

As the crack narrows, direct tunneling increases. The direct tunneling current relation to the electric field is more complex than Equation (8.14). An important observation is that metal has a high thermal coefficient of expansion (TCE) so that tunneling efficiency increases as the metal is heated and expands to close the crack. An IC will show faster operation at higher temperatures when metal cracks are present. No defect-free circuit ever shows this behavior.

Electron tunneling across opens enables a logic gate to function at low frequencies but fail at high frequencies, depending on the size of the open. Therefore, detection of these defects strongly depends on the careful application of speed testing discussed later in Chapter 10.

■ EXAMPLE 8.2

A metal-1 line of width 0.3 μm has a crack of d = 25 Å located 10 μm away from the transistor gate that the line is driving. The node line has been at 0 V for a long time, thus completely discharging the transistor gate node. Determine the final voltage at the transistor gate if the node driver connects its output at V_{DD} for 1 μs. Assume that the polysilicon capacitance is negligible and that for this technology $V_{DD} = 3.5$ V, metal-1 height t = 5500 Å, $V_{tn} = 0.35$ V, $\alpha = 1$ μA/V^2, and $\beta = 3 \times 10^8$ V/cm.

The initial voltage difference across the crack is 3.5 V, while the crack size is 25 Å indicating that the dominant conduction mechanism will be Fowler–Nordheim tunneling [Equation (8.14)]. Figure 8.6 illustrates the problem.

Initially, C is discharged and when the node driver switches to a logic 1, the voltage across the crack is $V = V_{DD}$. This creates an initial electric field of value $\mathscr{E} = V/d$ = 3.5 V/25 Å = 14 MV/cm, which creates a tunneling current that charges the capacitor and reduces the electric field. Hypothetically, the charging mechanism would continue until the capacitor is charged at V_{DD}. In practice, when the voltage across the metal crack goes below around 3 V, direct tunneling also takes place and Equation (8.14) is no longer valid.

The current through the crack is given by

$$I = J_e \cdot s = s \cdot \alpha \cdot \mathscr{E}^2 \cdot e^{-\beta/\mathscr{E}} = s \cdot \alpha \cdot \frac{V^2}{d^2} \cdot e^{-\beta \cdot d/V}$$

where s is the crack section area given by

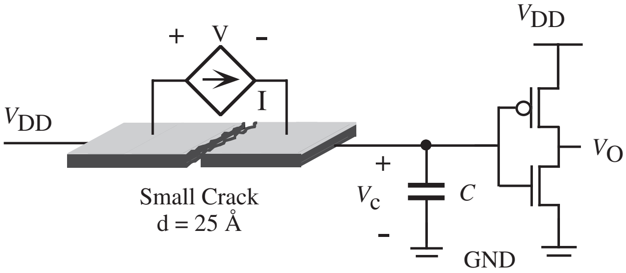


Figure 8.6. Tunneling through a metal crack.

$$s = w \cdot t = 0.3\ \mu\text{m} \cdot 5500\ \text{Å} = 165\ \text{fm}^2$$

The current–voltage equation at the parasitic capacitor is

$$I = C\left(\frac{dV_\text{C}}{dt}\right)$$

since

$$V_\text{DD} = V + V_\text{C} \Rightarrow V_\text{C} = V_\text{DD} - V$$

then

$$I = -C\left(\frac{dV}{dt}\right)$$

Equating this current to the current through the crack, the equation describing the problem is

$$-C\left(\frac{dV}{dt}\right) = s \cdot \alpha \cdot \frac{V^2}{d^2} \cdot e^{-(\beta \cdot d)/V}$$

that can be rewritten as

$$\left(\frac{1}{V^2}\right)e^{(\beta \cdot d)/V}\, dV = -\left(\frac{\alpha \cdot s}{C \cdot d^2}\right)dt$$

To solve this equation we define

$$y = e^{(\beta \cdot d)/V}$$

and rewrite

$$dy = \left(\frac{\beta \cdot s \cdot \alpha}{C \cdot d}\right)dt$$

Integrating

$$\int_{y(t=0)}^{y(t=T)} dy = \frac{\beta \cdot s \cdot \alpha}{C \cdot d}\int_{t=0}^{t=T} dt \Rightarrow e^{(\beta \cdot d)/V_\text{f}} - e^{(\beta \cdot d)/V_\text{DD}} = \left(\frac{\beta \cdot \alpha \cdot s}{C \cdot d}\right)T$$

where V_f is the final voltage at $t = T = 1\ \mu$s in this case. The solution is:

$$V_\text{f} = \frac{\beta \cdot d}{\ln\left[e^{(\beta \cdot d)/V_{DD}} + \left(\frac{\beta \cdot \alpha \cdot s}{C \cdot d}\right)T\right]}$$

Note that the solution is consistent since for $T = 0$, $V_\text{f} = V_\text{DD}$, whereas if $T \rightarrow \infty$, then V_f goes to zero. Substituting the interconnect and technology values, and making $T = 1\ \mu$s, $V_\text{f} = 3.3$ V. Since the voltage across the crack is beyond 3 V, the equation is valid during the charging process. The voltage at the gate capacitor is $V_C = 0.17$ V, which is not enough to turn on the *n*MOS device. ■

8.2.5 Other Effects

Additional charge components may be present when a floating node due to an open defect affects a transistor gate. These components come from the charge induced during the fabrication processes such as plasma etching or ion implantation. Although this charge can be removed or masked during subsequent annealing steps, the oxide traps created by these reactive process steps may manifest later due to hot-carrier injection during circuit operation [14]. The final trapped charge due to this mechanism depends on the amount of charge injected during the reactivation process step, the amount of charge removed during successive annealing, and, finally, the amount of charge trapped in the oxide during circuit operation. This makes the contribution of these effects on the total charge in the floating node difficult to compute if an open defect occurs, thus further complicating the exact modeling of opens in ICs.

8.3 OPEN DEFECT CLASSES

This section organizes the diverse behavior of CMOS open circuit defects into six classes or models. Each class is described with supporting evidence. The outward effect of an open defect in an IC can give immediate clues to the type of defect. Importantly, the clues also eliminate what the possible defect is not. The impact of an open on circuit behavior depends on the transistors driven by such a floating gate and the gate topology to which they belong.

Six general classes of opens are identified from failure analysis and test research:

1. Transistor on
2. Transistor pair on
3. Transistor pair on/off
4. Delay
5. Memory (transistor off)
6. Sequential

Although the names are awkward, they roughly describe open defect class behavior. The first five open categories appear in combinational logic circuits, and in certain instances in sequential circuit open defect behavior. We will look at each type and the supporting behavioral data.

8.3.1 Transistor-On Open Defect

An open gate to a single transistor is called a transistor-on defect class, and it has an unusual response [1, 2, 7]. Figure 8.7(a) shows a CMOS test inverter whose polysilicon gate to metal was removed at the mask level. Its static transfer curve is shown in Figure 8.7(b). Surprisingly, the curve shows functionality despite the loss in noise margin caused by the shift of the voltage curve to the right [2].

When $V_{\text{in}} - V_{\text{DD}} \ll V_{tp}$, the *p*MOS transistor in Figure 8.7(a) turns on strongly into the

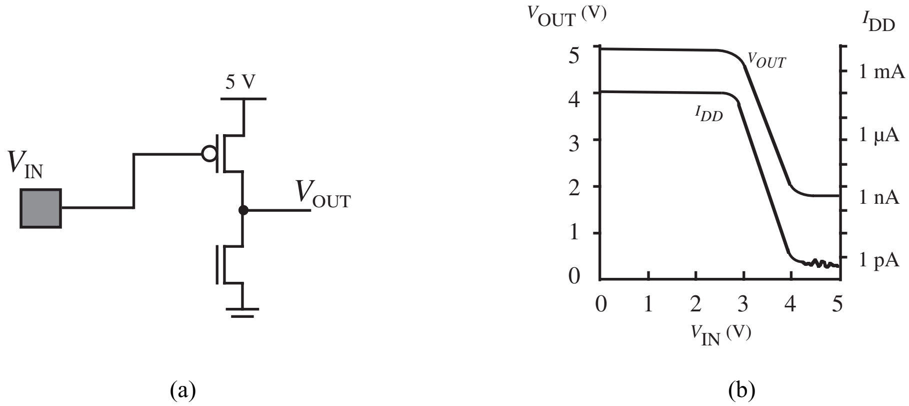


Figure 8.7. Transistor-on open defect class. (a) Open transistor in inverter. (b) Transfer characteristic [2].

nonsaturated state and 5 V is fed to the output drain. The *n*MOS transistor initially has no gate drive and is off. When the drain voltage rises and approaches V_{DD}, the DC capacitive coupling to the *n*MOS transistor from drain to gate and from gate to source creates a capacitive voltage divider to the gate. When the capacitively induced voltage at V_{GSn} is larger than V_{tn} the *n*MOS transistor turns on, drawing current from the *p*MOS transistor.

The static transfer curve in Figure 8.7(b) shows $V_{out} = 4.96$ V for $V_{in} < 2.5$ V in a technology with $V_{tn} \approx |V_{tp}| \approx 0.8$ V. Above $V_{in} = 2.5$ V, the *p*MOS transistor begins to turn off, but charge continues to pass from the output node to the *n*MOS transistor, as shown by the I_{DD} curve. V_{out} drops until the attenuated signal V_{GSn} is below threshold. The *n*MOS transistor turns off when $V_{out} \approx 1.8$ V, and the output node is in a high-impedance state with slow discharge leakage through the reverse-bias junctions of the MOSFETs. The flat portion of the curve at the lower-right-hand side reflects this slow discharge. Power supply current is strongly elevated at 96 µA for $0 < V_{in} < \approx 2.8$ V. It is elevated in only one logic state, which is typical for many defects that elevate I_{DDQ}. It is important to observe that contrary to electrical intuition, open circuit defects often elevate power supply current.

■ EXAMPLE 8.3

The circuit in Figure 8.7(a) approximates the DC conditions of a transistor gate open. If $V_{DD} = 5$ V, $V_{tn} = 0.8$ V, $V_{tp} = -0.8$ V, $K_n = 200$ µA/V², $K_p = 125$ µA/V², $W/L = 1$, $C_{gd} = 20$ fF, and $C_{gs} = 9$ fF, what is V_{out} when the *n*MOS transistor just turns off?

$$V_G = V_D\left(\frac{C_{gd}}{C_{gs} + C_{gd}}\right)$$

$$V_{out} = V_D = V_G\left(\frac{C_{gs} + C_{gd}}{C_{gd}}\right)$$

$$V_{out} = (29/20) \times 0.8\text{ V} = 1.160\text{ V}$$

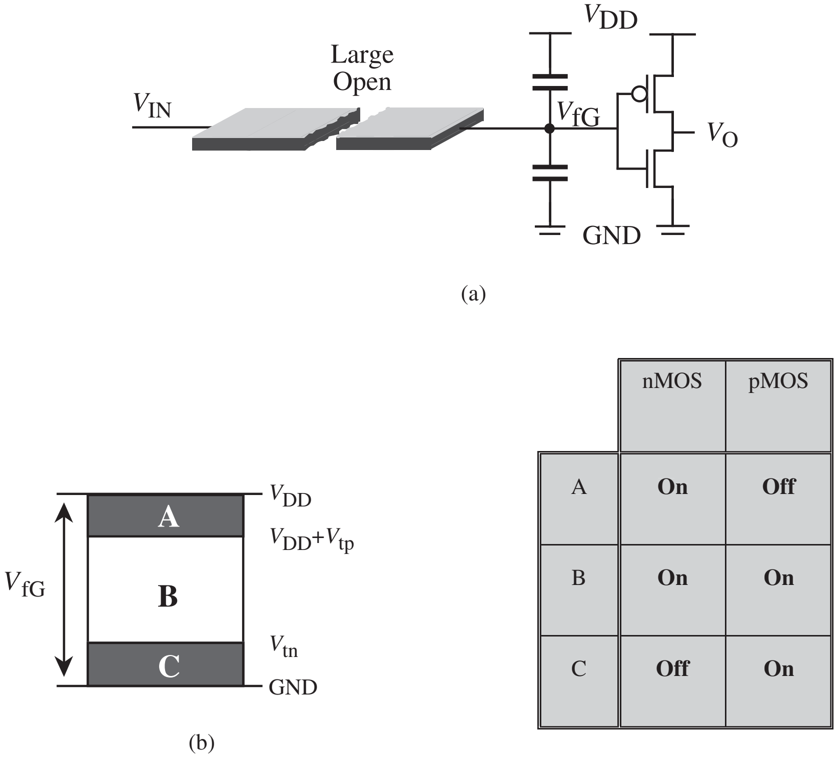
■

A dynamic transfer curve would include the effects of the C_{gd} capacitance from the *p*MOS transistor. C_{gd} would allow charge injection quickly out of the output drain node, causing the voltage on that node to drop quicker. This would tend to shift the voltage transfer curve to the left of the static curve shown in Figure 8.7, providing some assistance to the logic change.

8.3.2 Transistor Pair-On and Transistor Pair-On/Off

A wide open-circuit interconnect defect on a logic gate input affects two transistors—the *p*MOS transistor and its complement *n*MOS transistor. Figure 8.8(a) shows such a defect situation. The affected node floats in a high-impedance state, searching for a voltage that satisfies its environment. The node electrically lies between V_{DD} and ground coupled by parasitic capacitances. The local topology will settle the node somewhere between V_{DD} and 0 V. It is typically not predictable where a floating node finds a steady state value, therefore we expect a range, and that range has three distinct regions. Figure 8.8(b) illustrates the variable region where the floating node can settle. If the gate floats to a value in region B, then both transistors turn on, V_0 is at a weak stuck voltage, and quiescent power supply current I_{DDQ} is elevated.

If the node voltage is near the power rails at a value outside region B, then only one



(a)

	nMOS	pMOS
A	**On**	**Off**
B	**On**	**On**
C	**Off**	**On**

(b)

Figure 8.8. Transistor pair on and pair on/off open defects. (a) Large open defect and (b) its floating gate voltage range.

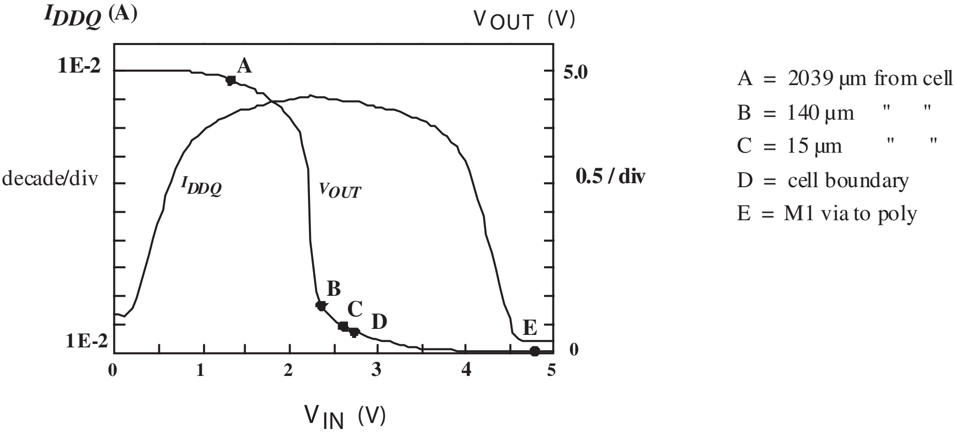


Figure 8.9. Floating node response to decreasing gates capacitance (A–E) [2].

transistor is on and the other off (regions A and C). V_0 is then a strong stuck voltage, and I_{DDQ} is not elevated in regions A and C. Thus, a wide open defect to a logic gate can cause (1) a weak stuck voltage and I_{DDQ} elevation (transistor pair on), or (2) a strong stuck voltage and no I_{DDQ} elevation (transistor pair on/off). The table in Figure 8.8 summarizes this.

Supporting data for these two open defect classes are shown in Figure 8.9 [2]. Points A–E are results from five CMOS inverter test structures in which the length of the metal interconnect floating node to the inverter was varied from 2039 μm (A) down to an open polysilicon gate to metal missing contact (E). The longer floating metal structures had a larger capacitance to ground, pulling the floating gate voltage lower (an effect similar to the one analyzed in Example 8.1).

The measured output voltages and I_{DDQ} were superimposed on a normal inverter transfer curve from the same die to estimate the floating gate voltage on the x-axis. Experimental points A–D are measured from transistor pair on open defects (region B in Figure 8.8) and Point E is a transistor pair on/off defect (region A in Figure 8.8). Circuits A–D showed weak stuck output voltage behavior and significant elevated I_{DDQ}. Circuit E showed strong stuck output voltage behavior and no I_{DDQ} elevation. The floating gate voltage for structure E was $V_{Gfl} \approx 4.8$ V. This turns off the *p*MOS transistor clamping V_{out} at 0 V. These distinctions are important when understanding symptoms in failure analysis or devising test detection methods.

■ EXAMPLE 8.4

Assume that the two open defects in Figure 8.10 lie (a) 50 μm to the right and (b) 50 μm to the left of the lower-left node. What electronic behavior differences would you expect?

The defect in Figure 8.10(a) affects one transistor. It is a transistor-on defect whose response is linked to capacitive coupling between its drain and source. It will probably pass a functional test, but have elevated I_{DDQ} in one logic state (AB = 10). Figure 8.10(b) shows a transistor pair on or pair on/off defect. If the transistors connected to node B float between the threshold voltages, then I_{DDQ} will

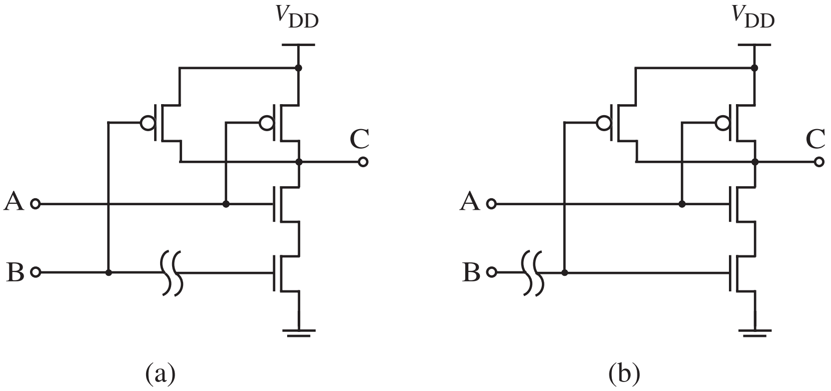


Figure 8.10. Open defect difference. (a) Open defect to transistor gate. (b) Open defect to logic gate.

elevate, and a weak stuck V_C appears for the AB = 10, 11 states. If the floating nodes clamp hard to one of the rails, then I_{DDQ} is not elevated and the output is a strong stuck voltage. ■

8.3.3 The Open Delay Defect

This effect was discussed earlier in the chapter where tunneling was the main conduction mechanism through the open [3]. Figure 8.11 shows an equivalent circuit with a tunneling current density J_e crossing the crack. ICs with these defects can operate into the hundreds of MHz, depending on the small dimension of the crack.

This metal defect has an interesting temperature property that relates to test and failure

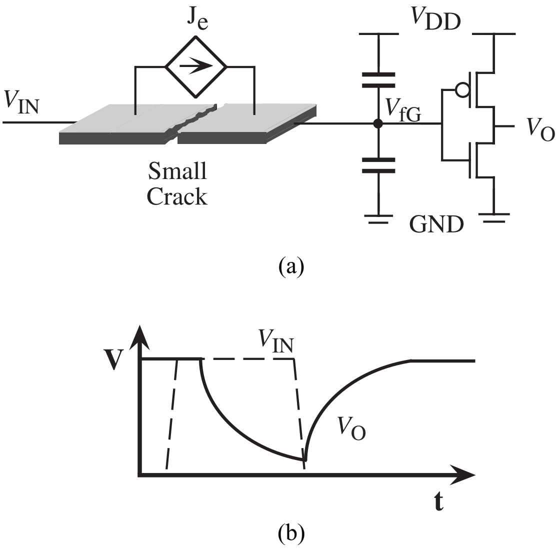


Figure 8.11. (a) Tunneling open and (b) response to electron tunneling.

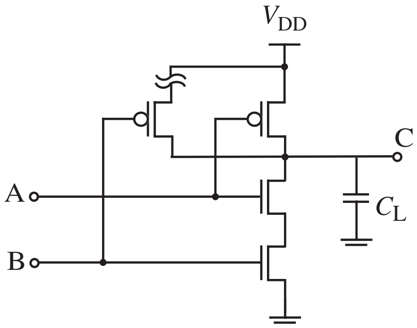
analysis. When metal is cooled, it contracts and the crack widens. Fewer tunneling electrons are supported, and the maximum operating frequency drops. This is an unusual property for an IC since, normally, ICs increase their maximum operating frequency as the circuit is cooled. It is a clue to the presence of incomplete metal arising, perhaps, from stress voiding, electromigration, or a flaw in fabrication to a via or contact. These defects are often referred to as tunneling defects. They can be difficult to detect in test or locate in failure analysis. Resistive interconnect opens (specifically, high-resistive vias) are discussed further in Chapter 9, as they are also classed as parametric delay defects.

8.3.4 CMOS Memory Open Defect

Figure 8.12 shows a 2NAND with an open defect in a *p*MOS transistor source—no current can pass through this transistor. The truth table in Figure 8.12 shows correct logic results for the first two rows. These are expected. When AB = 00, both *n*-channel pull-down transistors are off, and the good *p*-channel transistor pulls the output to a correct logic one. When AB = 01, the pull-down transistor is again blocked, but the good *p*-channel transistor pulls node C to a correct logic one. However, row 3 gives a correct result that is subtle. The AB = 10 puts node A in a noncontrolling logic state, but the result seems to depend on conduction through the defective *p*MOS transistor. The reality is that the AB = 10 vector puts the output node in a floating or high-impedance state, and there is no outlet for rapid discharge of its high logic voltage charge from the previous logic state. The previous logic high state was stored in the load capacitance C_L and a correct value was read. The fourth vector AB = 11 pulls node C down to its expected 0 V.

This is a dramatic defect causing no error even though the logic gate was stepped through the whole truth table, and one transistor was incapable of conduction. The same result would have occurred if the defective transistor were removed from the circuit. The clue to potential error is the sequence of vectors. If the AB = 11 state was followed by the AB = 10 state, then an error of 0 V is measured on the second vector (last two rows in table).

We could examine an IC for this defect by testing each transistor with a sequence of



A	B	C
0	0	1
0	1	1
1	0	1
1	1	0
⋮	⋮	⋮
1	1	0
1	0	0

Figure 8.12. Stuck-open defect.

two-vectors [15]. This test may become numerically intractable for large circuits not designed for easy testing. Design for testability (DfT) circuits such as the scan circuits described in Chapter 10 partition the larger circuit into smaller combinational logic blocks. In this case, two-vector-sequenced patterns may be feasible.

Figure 8.13 shows a logic gate output voltage and I_{DDQ} timing response when a memory defect was placed in its high-impedance state at room temperature [13]. The circuit was a 2NOR driving a ROM row decoder inverter whose V_0 node was probed. At $t = 12$ s, the circuit was put in the high-impedance state of the memory defect. The output voltage shows a drift and change in logic state over about 6–8 s. This altered logic state under normal operating conditions might occur many millions of clock cycles after setting the high-impedance state.

I_{DDQ} responds with about a 200 ms step followed by instability to about 30 s and then a slow decline. $I_{DD}(t)$ reflects the biasing of one or more transistors as the floating node drifts through a range of values. In practice this defect is usually caught by an accidental sequence of voltage-based test patterns or by I_{DDQ} as shown here. The memory defect is difficult for test detection and failure analysis.

8.3.5 Sequential Circuit Opens

Figure 8.14 shows several wide open circuit defect locations in a sequential circuit. Open-1 (O_1) affects the *n*MOS and *p*MOS transistor gates of the two transmission gates. O_1 will cause those nodes to float to the same voltage. If the node floats to an intermediate voltage that turns on the *n*- and *p*-channel transistors, then the latch loses its signal isolation. Transistor contention or conflict occurs since both transmission gates are always on. The incoming signal conflicts with the stored feedback signal $\overline{Q}$. The latch will functionally fail, and I_{DDQ} will elevate during contention.

If the O_1 node floats to near V_{DD}, then the *n*-channel pass transistor is permanently on, and the *p*-channel transistor is permanently off. The feedback transmission gate still works with a single *n*-channel pass transistor, but conflict again occurs when the input transmission gate is supposed to be off. A similar situation occurs when the O_1 node floats near ground. The *p*-channel transistor of the feedback transmission gate is always on, and the *n*-channel transistor of the input transmission gate always off. Signal contention oc-

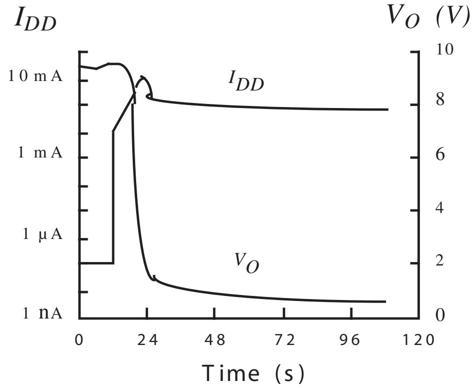


Figure 8.13. Timing response of 2NOR memory open defect [13].

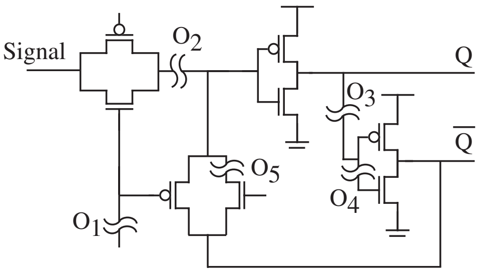


Figure 8.14. Open circuit defects in a sequential circuit.

curs when the feedback transmission gate is supposed to be off. I_{DDQ} elevation and signal error occur during the contention. The transistor contention is between one of the transistors of the second inverter ($\overline{Q}$) through the transmissions gates to the transistors driving the signal input.

Defect O_2 will float the input to the inverter when the feedback transmission gate is off and the response will be a transistor pair on or pair on/off. The returned signal in the latch will alter the floating node voltage and try to drive it to a strong voltage when the CLK signal turns on the latch T-gates. When the latch T-gate is off, then the floating node will again seek its steady-state level. The latch is dysfunctional with O_2 and an error is detected when applying a functional voltage test. I_{DD} may or may not be elevated. Defect O_3 also gives a response similar to that described for transistor pair on or pair on/off; the latch is again dysfunctional.

Defect O_4 exhibits the transistor-on open response of Figure 8.14 that should elevate I_{DDQ}, and it may or may not fail a functional test. O_5 opens the *p*MOS transistor so the T-gate will pass weak logic low voltages and unattenuated logic high voltages. Overall, the open circuit responses in the flip-flop are similar to the open behaviors described in Section 8.3. The failing symptom can be either that of a functional (voltage-based) test or an I_{DDQ} failure, or both simultaneously. All these defects affect the critical setup and hold times.

8.4 SUMMARY

The variety of open circuit responses is due to the sensitivity to (1) open defect location (drain, source, gate), (2) open to logic gate input, (3) sequential circuit, or (4) narrow cracks in flat metal or vias. Local topography determines capacitive coupling, and temperature changes will affect the dimensions of some opens. Although these responses are more complex than bridging defects, they are electronically understood so that later we can intelligently apply test methods that target each form of open circuit behavior.

REFERENCES

1. V. Champac, A. Rubio, and J. Figueras, “Electrical model of the floating gate defect in CMOS IC’s: Implications on I_{DDQ} Testing,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 13.* 3, 359–369, March 1994.

2. C. Hawkins, J. Soden, A. Righter, and J. Ferguson, "Defect classes—An overdue paradigm for testing CMOS ICs," in *IEEE International Test Conference (ITC),* pp. 413–424, October 1994.
3. C. Henderson, J. Soden, and C. Hawkins, "The behavior and testing implications of CMOS IC Open Circuits," in *IEEE International Test Conference (ITC),* pp. 302–310, October 1991.
4. S. Johnson, "Residual charge on the faulty floating gate MOS transistor," in *IEEE International Test Conference (ITC),* pp. 555–560, October 1994.
5. J. Li and E. McCluskey, "Testing for tunneling opens," in *IEEE International Test Conference (ITC),* pp. 85–94, October 2000.
6. T. Miller, J. Soden, and C. Hawkins, "Diagnosis, analysis, and comparison of 80386EX I_{DDQ} and functional failures," *IEEE I_{DDQ} Workshop,* Washington D.C., October 1995.
7. W. Maly, P. Nag, and P. Nigh, "Testing oriented analysis of CMOS ICs with opens," in *International Conference on Computer Aided Design (ICCAD),* pp. 344–347, 1988.
8. M. Renovell and G. Cambon, "Electrical analysis and modeling of floating-gate fault," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 11,* 1450–1458, Nov. 1992.
9. M. Renovell, A. Ivanov, Y. Bertrand, F. Azais, and S. Rafiq, "Optimal conditions for Boolean and current detection of floating gates," in *IEEE International Test Conference (ITC),* pp. 477–486, October 1999.
10. W. Riordan, R. Miller, J. Sherman, and J. Hicks, "Microprocessor performance as function of die location for a 0.25 μm five layer metal CMOS logic process," in *International Reliability Physics Symposium (IRPS),* pp. 1–11, April 1999.
11. R. Rodriquez-Montanes, J. Segura, V. Champac, J. Figueras, and A. Rubio, "Current vs. logic testing of gate oxide shorts, floating gates, and bridging failures in CMOS," in *IEEE International Test Conference (ITC),* pp. 510–519, October 1991.
12. A. Singh, H. Rasheed, and W. Weber, "I_{DDQ} testing of CMOS opens: An experimental study," in *IEEE International Test Conference (ITC),* pp. 479–489, October 1995.
13. J. Soden, R. Treece, M. Taylor, and C. Hawkins, "CMOS IC stuck-open fault electrical effects and design considerations," pp. 423–430, in *IEEE International Test Conference (ITC),* pp. 302–310, August 1989.
14. R. Tu, J. King, H. Shin, and C. Hu, "Simulating process-induced gate oxide damage in circuits," *IEEE Transactions on Electron Devices, 44,* 9, 1393–1400, September 1997.
15. R. Wadsack, "Fault modeling and logic simulation of CMOS and MOS integrated circuits," *Bell Systems Technical Journal,* 1449–1488, May-June 1978.

EXERCISES

8.1. Observe the schematic and transfer curve in Figure 8.7. Would the shape of the curve change if the input were swept from $V_{in} = 5 - 0$ V instead of being swept from $V_{in} = 0 - 5$ V as shown in Figure 8.7(b)?

8.2. A 2NOR gate has an open defect in one of the pull-down source leads (Figure 8.15). Verify whether this defective circuit will
(a) Correctly pass an ordered truth table beginning with 00
(b) Correctly pass an ordered truth table beginning with 11

8.3. The stuck-open fault behavior was described as a behavior related to the parallel transistors in a logic gate. In theory, a stuck-open behavior can occur in a logic gate

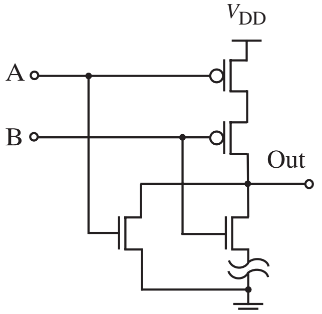


Figure 8.15.

series stack of transistors initiated by events occurring during power-up of the circuit. Figure 8.16 illustrates this with a single series stack (an inverter) in which an open defect appears in the drain of the *n*-channel transistor. Describe how a stuck-open behavior could occur in this series stack circuit.

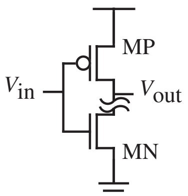


Figure 8.16.

8.4. An integrated circuit has the following failure symptoms: It fails a functional (Boolean) test and I_{DDQ} is elevated. If you suspect an open circuit defect is causing this, what type of open defect class would it be? Explain.

8.5. An integrated circuit has the following failure symptoms: The maximum clock frequency of the IC declines as the temperature is increased. There is no Boolean failure at slow clock frequency, nor is there I_{DDQ} elevation. If you suspect an open circuit defect is causing this, what type of open defect class would it be? Explain.

8.6. An integrated circuit has the following failure symptoms: It fails a functional (Boolean) test and I_{DDQ} is not elevated. If you suspect an open circuit defect is causing this, what type of open defect class would it be? Explain.

8.7. A metal-6 interconnect that is 3 μm wide has a 27 Å crack located 50 μm away from its end where it goes down to a gate node. Assuming that the floating portion of the metal is at 0 V, compute the final node voltage when the portion of the line connected to its driver is settled to V_{DD} during 1 μs. For this technology, the metal-6 height is $t = 1.04$ μm, its capacitance to the substrate is 0.008 fF/μm, and $V_{DD} = 3.5$ V. Determine the dominant conduction mechanism and the validity of the equation used.

8.8. The dynamic NAND shown in Figure 8.17 has been unpowered for a long time. Determine the voltage at the gate of N1 2 μs after the signal lines driving the gates of N1 and N2 simultaneously make a 0 to 1 transition. Assume the technology values given in Example 8.2 and a crack of 20 Å.

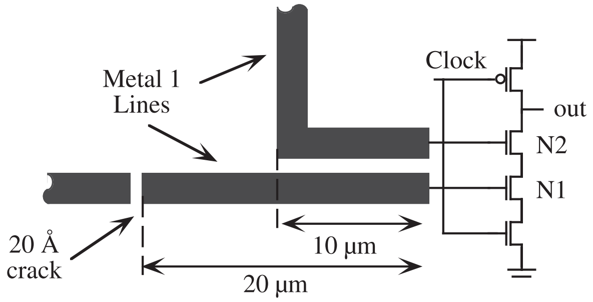


Figure 8.17.

8.9. Two different failure symptoms were found upon speed testing a microprocessor at 30°C and at 100°C. If you suspect an open defect, describe which it might be if:

(a) The parts run faster at 100°C than at 30°C.

(b) The parts run slower at 100°C than at 30°C.

8.10. An ASIC showed peculiar failure symptoms. The failure was isolated to a subcircuit that failed certain verification tests and not others. It did not seem to make sense. What type of an open defect might it be?

APPENDIX A

SOLUTIONS TO SELF-EXERCISES

A.1 CHAPTER 1. ELECTRICAL CIRCUIT ANALYSIS

Self-Exercise 1.1
$V_0 = 1.65$ V
$V_P = 52.33$ V

Self-Exercise 1.2
$R_{eq} = 1200 \parallel 950 \parallel \mathrm{R3}$
$250 = [1/1200 + 1/950 + 1/\mathrm{R3}]^{-1}$
$1/R_3 = 1/250 - 1/1200 - 1/950 \Rightarrow R_3 = 473.0\ \Omega$

Self-Exercise 1.3
$R_{in} = 8$ kΩ
$I_{BB} = 1.25$ mA
$V_0 = 7.5$ V

Self-Exercise 1.4
(a) $R_{in} = 667\ \Omega$
(b) $R_{in} = 2$ kΩ
(c) $R_{in} = 990.1\ \Omega$
(d) $R_{in} = 999\ \Omega$

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

Self-Exercise 1.5

$R_{\mathrm{in}} = 11.5\ \mathrm{k\Omega}$
$I_{\mathrm{BB}} = 86.96\ \mu\mathrm{A}$
$V_0 = 652.2\ \mathrm{mV}$

Self-Exercise 1.6

(a) $\mathrm{I}_3 = 100\ \mu\mathrm{A} - 50\ \mu\mathrm{A} - 10\ \mu\mathrm{A} = 40\ \mu\mathrm{A}$

(b) $V_{\mathrm{R3}} = 40\ \mu\mathrm{A} \times 50\ \mathrm{k\Omega} = 2.0\ \mathrm{V}$, so that

$R_1 = 2.0\ \mathrm{V}/50\ \mu\mathrm{A} = 40\ \mathrm{k\Omega}$

$R_2 = 2.0\ \mathrm{V}/10\ \mu\mathrm{A} = 200\ \mathrm{k\Omega}$

Self-Exercise 1.7

$V_{\mathrm{R3}} = V_0 = 200\ \mu\mathrm{A} \times 8\ \mathrm{k\Omega} = 1.6\ \mathrm{V}$

$R_1 = [3.3 - 1.6]/650\ \mu\mathrm{A} = 2.615\ \mathrm{k\Omega}$

$I_2 = 650\ \mu\mathrm{A} - 200\ \mu\mathrm{A} = 450\ \mu\mathrm{A}$

$R_2 = 1.6/450\ \mu\mathrm{A} = 3.556\ \mathrm{k\Omega}$

Self-Exercise 1.8

$R_1 = 7.125\ \mathrm{k\Omega}$

Self-Exercise 1.9

(a) $R_{\mathrm{eq}} = 1\ \mathrm{M\Omega} \parallel 2.3\ \mathrm{M\Omega}$; that is, $R_{\mathrm{eq}} = 697.0\ \mathrm{k\Omega}$

(b) $R_{\mathrm{eq}} = 75\ \mathrm{k\Omega} \parallel [150\ \mathrm{k\Omega} + 35\ \mathrm{k\Omega}]$; that is, $R_{\mathrm{eq}} = 53.37\ \mathrm{k\Omega}$

Self-Exercise 1.10

$R_{\mathrm{eq}} = R_1 + R_3 \parallel (R_2 + R_4)$

$R_{\mathrm{eq}} = R_1 + R_3 \parallel (R_2 + R_4) + R_5$

$R_{\mathrm{eq}} = R_1 \parallel [R_2 + R_4 \parallel (R_3 + R_5)]$

Self-Exercise 1.11

(a) $R_{\mathrm{eq}} = 31.98\ \mathrm{k\Omega}$

(b) $R_{\mathrm{eq}} = 36.98\ \mathrm{k\Omega}$

(c) $R_{\mathrm{eq}} = 10.32\ \mathrm{k\Omega}$

Self-Exercise 1.12

$V_0 = (12\mathrm{k} \parallel 20\mathrm{k})/[4\mathrm{k} + (12\mathrm{k} \parallel 20\mathrm{k})] \times 1\ \mathrm{V} = 0.652\ \mathrm{V}$

$V_{4\mathrm{k}} = (4\mathrm{k})/[4\mathrm{k} + (12\mathrm{k} \parallel 20\mathrm{k})] \times 1\ \mathrm{V} = 0.348\ \mathrm{V}$

$V_{\mathrm{BB}} = 0.652 + 0.348 = 1.00\ \mathrm{V}$

$R_{\mathrm{in}} = 4\mathrm{k} + (12\mathrm{k} \parallel 20\mathrm{k}) = 11.5\ \mathrm{k\Omega}$

$I_{\mathrm{BB}} = V_{\mathrm{BB}}/R_{\mathrm{in}} = 87.0\ \mu\mathrm{A}$

Self-Exercise 1.13

$R_{\mathrm{in}} = 45\mathrm{k}\|[8\mathrm{k} + (18\mathrm{k}\|30\mathrm{k})] = 13.48\ \mathrm{k\Omega}$

$V_0 = (18\mathrm{k}\|30\mathrm{k})/[8\mathrm{k} + (18\mathrm{k}\|30\mathrm{k})] \times 2.5\ \mathrm{V} = 1.461\ \mathrm{V}$

$I_{\mathrm{BB}} = 185.4\ \mu\mathrm{A}$

Self-Exercise 1.14

$I_{12\mathrm{k}} = [20/(20 + 12)] \times 87.0\ \mu\mathrm{A} = 54.38\ \mu\mathrm{A}$

$I_{20\mathrm{k}} = [12/(20 + 12)] \times 87.0\ \mu\mathrm{A} = 32.63\ \mu\mathrm{A}$

(Notice that the KCL is satisfied: $87.0\ \mu\mathrm{A} = 54.38\ \mu\mathrm{A} + 32.63\ \mu\mathrm{A}$.)

Self-Exercise 1.15

I_{45k} = [8k + (18k || 30k)]/[45k + 8k + (18k || 30k)] × 185.4 μA = 55.55 μA

I_{8k} = (45k)/[45k + 8k + (18k || 30k)] × 185.4 μA = 129.9 μA

(Notice that I_{8k} is more easily obtained from KCL, where I_{8k} = 185.4 μA – 55.55 μA = 129.9 μA.)

I_{18k} = (30k)/(30k + 18k) × 129.9 μA = 81.16 μA

I_{30k} = (18k)/(30k + 18k) × 129.9 μA = 48.69 μA

V_{BB} = 2.5 V

Self-Exercise 1.16

(a) I_{10k} = 46.15 μA
I_{15k} = 30.77 μA
I_{20k} = 23.08 μA
(b) $R_{20k} \Rightarrow$ 114 kΩ

Self-Exercise 1.17

(a) V_0 = 3.60 mV
(b) I_2 = 714.3 μA, I_9 = 400 μA

Self-Exercise 1.18

R_{in} = 250 + [4k + (3k ||1k)] || [2k || (750 + 1.5k)] + 250 = 1.366 kΩ

$I_{1.5k}$ = [5 V/R_{in}] × [2k || (4k +(3k 1k))]/[2k || (4k + (3k||1k) + (750 + 1.5k)]
= 1.409 mA

Self-Exercise 1.19

(a) C_{eq} = 38.1 nF
(b) C_{eq} = 25.9 fF

Self-Exercise 1.20

$$V_1 = \frac{C_2}{C_1 + C_2} \times 100\text{ mV}$$

$$= \frac{75}{45 + 75} \times 100\text{ mV} = 62.5\text{ mV}$$

$$V_2 = \frac{45}{45 + 75} \times 100\text{ mV} = 37.5\text{ mV}$$

Self-Exercise 1.21

$$V_2 = \frac{C_1}{C_1 + C_2} \times V_D$$

$$= \frac{30}{30 + 25} \times V_D = 0.7\text{ V}$$

$$V_D = \frac{55}{30} \times 0.7\text{ V} = 1.28\text{ V}$$

Self-Exercise 1.22

(a) $V_{\mathrm{D}} = \left(\dfrac{86.17\ \mu\mathrm{eV/K}}{1\ \mathrm{eV}}\right)(298\ \mathrm{K})\ \ln\left(\dfrac{200\ \mathrm{nA}}{1\ \mathrm{nA}} + 1\right) = 136.2\ \mathrm{mV}$

(b) $I_{\mathrm{D}} = 1\ \mathrm{nA}(e^{400/26}) = 4.80\ \mathrm{mA}$

Self-Exercise 1.23

$I_{\mathrm{D}} = 185\ \mu\mathrm{A}$
$V_{\mathrm{D}} = 315.3\ \mathrm{mV}$

Self-Exercise 1.24

(a) 5.179 V, –0.179 V
(b) 5 V

Self-Exercise 1.25

$V_{\mathrm{D1}} = 18.02\ \mathrm{mV}$
$V_0 = 4.982\ \mathrm{V}$

Self-Exercise 1.26

$I_{\mathrm{D}} = 26\ \mathrm{mV}/1\ \mu\mathrm{A} \Rightarrow 26\ \mathrm{k\Omega}$
$I_{\mathrm{D}} = 26\ \mathrm{mV}/100\ \mu\mathrm{A} \Rightarrow 260\ \Omega$
$I_{\mathrm{D}} = 26\ \mathrm{mV}/1\ \mathrm{mA} \Rightarrow 26\ \Omega$
$I_{\mathrm{D}} = 26\ \mathrm{mV}/10\ \mathrm{mA} \Rightarrow 2.6\ \Omega$
Dynamic diode resistance can vary over a wide range, depending upon the bias current.

A.2 CHAPTER 3. MOSFET TRANSISTORS

Self-Exercise 3.1

(a) Ohmic
(b) Ohmic
(c) Saturated
(Hint: watch your terminals.) The source terminal always has a lower or equal voltage than the drain in a conducting *n*MOS transistor.

Self-Exercise 3.2

$I_{\mathrm{D}} = 12.5\ \mu\mathrm{A}$ using saturated state model.
$V_{\mathrm{D}} = 9.375\ \mathrm{V}$

Self-Exercise 3.3

$I_{\mathrm{D}} = 184.8\ \mu\mathrm{A}$
$V_{\mathrm{D}} = 0.761\ \mathrm{V}$

Self-Exercise 3.4

$V_{\mathrm{GS}} = 3.33\ \mathrm{V}$
M1 in saturation

Self-Exercise 3.5

$R_1 = 180\ \mathrm{k\Omega}$

Self-Exercise 3.6

$R_0 = 537.6\ \Omega$

Self-Exercise 3.7
(a) Saturated state
(b) Ohmic state
(c) Boundary point of both saturated and ohmic state

Self-Exercise 3.8
$I_D = 29.4\ \mu A$, $V_0 = 2.94$ V

Self-Exercise 3.9
$I_D = 48.46\ \mu A$, $V_0 = 4.846$ V

Self-Exercise 3.10
$I_D = 153.6\ \mu A$, $V_0 = -1.848$ V

Self-Exercise 3.11
R = 239.0 kΩ

Self-Exercise 3.12
$W/L > 6024$

Self-Exercise 3.13
$I_{SAT} = 6$ mA

Self-Exercise 3.14
$I_{SAT} = 0.66$ mA

A.3 CHAPTER 4. CMOS BASIC GATES

Self-Exercise 4.1
$W_p/W_n = 3.36$

Self-Exercise 4.2
(a) 61.4%
(b) 71.7%; you must derive the equation for this from the V_{in} equation in Example 4.2.
(c) 10.4%

Self-Exercise 4.3
$\Delta V_{out}/\Delta V_{in} \approx -17$

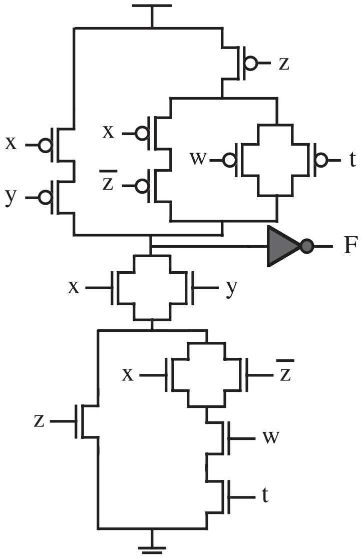
Self-Exercise 4.4
$\tau_D = 103.1$ ps at $V_{DD} = 2.5$ V
$\tau_D = 186.0$ ps at $V_{DD} = 1.8$ V
The extra delay is 83.0 ps, which represents an increase of 80.4%.

Self-Exercise 4.5
(a) $I_1I_2\mathrm{I}_3I_4 = 11X0$
(b) $I_1I_2\mathrm{I}_3I_4 = 1X00$

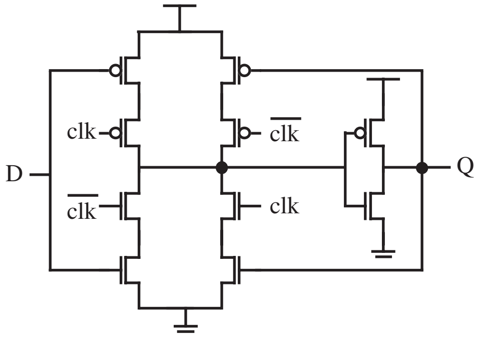
Self-Exercise 4.6
(a) $I_1I_2\mathrm{I}_3I_4 = 1100$
(b) $I_1I_2I_4I_5 = 1101$
(c) $I_2\mathrm{I}_3I_4I_5 = 1001$

A.4 CHAPTER 5. CMOS BASIC CIRCUITS

Self-Exercise 5.1



Self-Exercise 5.2
The gate level design of a D-latch shown in Figure 5.16(a) has four transistors in each NOR gate (thus, a total of 16 devices), plus two transistors for the inverter, resulting in 18 transistors. A transistor-level design of the circuit in Figure 5.16(b) is



This design uses 10 transistors shown plus two more to invert the clock for a total of 12 transistors. The circuit loads data when clk is high, and holds that data to output *Q* when the clk goes negative.

A.5 CHAPTER 6. FAILURE MECHANISMS IN CMOS ICs

Self-Exercise 6.1

(a) 29.7% Reduction in current density
(b) 50.7% Reduction in lifetime at the higher temperature

Self-Exercise 6.2
(a) $P = 1.2\ \text{W/cm}^2$
(b) $P = 12\ \text{kW/cm}^2$. Since the passivation material is highly insulating, you can get a hot region. 10 mA is typical for electromigration test structure experiments and 100 μA is typical for IC operation.
(c) The 100 μA current drops 300 μV and the 10 mA current drops 30 mV. These drops would typically not alter logic function.

Self-Exercise 6.3
250 μm

Self-Exercise 6.4
(a) 8.658×10^{-6} lb, 3.936×10^{-6} kg
(b) 0.1329 μm = 132.9 nm

A.6 CHAPTER 7. BRIDGING DEFECTS

Self-Exercise 7.1
(a) $R_{\text{crit}} = 3.246\ \text{k}\Omega$
(b) $K_p/K_n = 1$ and both transistor drains are at the logic threshold voltage.

Self-Exercise 7.2
$R_{\text{crit}} = 1.433\ \text{k}\Omega - 1.722\ \text{k}\Omega$

Self-Exercise 7.3
$V_{\text{G}} = 2.30$ V, so gate passes function (barely).

A.7 CHAPTER 8. OPEN DEFECTS

Self-Exercise 8.1
(a) $V_2 = 1$ V
(b) $V_{\text{DD}} = 2.513$ V

A.8 CHAPTER 10. DEFECT BASED TESTING

Self-Exercise 10.1
Node-c SA1 has three possible test vectors:
01 11 11
01 11 10
01 11 01
Node-m SA0 has nine possible test vectors including:
11 00 10
11 00 01
11 00 11

INDEX

CMOS Electronics: How It Works, How It Fails. By Jaume Segura and Charles F. Hawkins
ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

ABOUT THE AUTHORS

Jaume Segura is an associate professor of Electronic Technology in the Electronic Technology Group at the Balearic Islands University, Spain. He teaches graduate courses in microelectronic analog and digital design, microprocessor-based design, test engineering, and reliability. He also teaches short courses on these topics in Europe, the United States, and South America. Professor Segura has done research in these topics for more than 10 years, establishing research collaborations with several universities and companies. He has been a visiting researcher at Philips Semiconductors in New Mexico and Intel Corporation in Oregon.

Dr. Segura has been a guest editor for the *IEEE Design & Test of Computers* and was the general chair of the IEEE International On-Line Testing Workshop in 2000. Dr. Segura is a member of the executive committee of the Design Automation and Test in Europe (DATE) conference and a member of the organizing committee of the IEEE VLSI Test Symposium (VTS). He was director of the IEEE-TTTC Summer Courses on Design and Test Topics from 1999–2002. He also served as program committee member for several international conferences including the IEEE International Test Conference (ITC), IEEE VLSI Test symposium (VTS), the IEEE International On-Line Testing Symposium (IOLTS), the IEEE European Test Symposium (ETS), and the IEEE Workshop on Defect Based Testing (DBT). He is the chairman of the IEEE-CAS Spanish Chapter. He received his Ph.D. in Electronic Technology from the Polytechnic University of Catalonia in 1992, and a Physics degree from the Balearic Islands University in 1989. In 1994, he was a visiting researcher at the University of New Mexico.

Charles F. Hawkins is a professor in the Electrical and Computer Engineering Department at the University of New Mexico, Albuquerque. He teaches graduate courses in Microelectronics Design, Reliability, Test Engineering, and Failure Analysis. For more than

ISBN 0-471-47669-2 © 2004 Institute of Electrical and Electronics Engineers, Inc.

20 years, Professor Hawkins has also taught industry short courses on these topics in the United States, Canada, Europe, and Australia. His research in these topics includes a 20-year research collaboration with the Microelectronics Group at Sandia National Laboratories in Albuquerque, New Mexico; a four-year collaboration with the AMD Corporation and Sandia Labs in failure analysis of timing paths in high-speed microprocessors; and a four-month sabbatical with Intel Corporation in Rio Rancho, New Mexico.

He was the editor of the *ASM Electron Device Failure Analysis Magazine (EDFA)* from 1999–2003 and the general and program chair of the International Test Conference (ITC) in 1996 and 1994, respectively. He co-shared eight Best or Outstanding Paper conference awards with colleagues at Sandia Labs, Intel, and AMD Corp at ITC and at the International Symposium on Test & Failure Analysis (ISTFA). In addition to this book, he has coauthored two books on electronics. He received his Ph.D. in Bioengineering from the University of Michigan, a Master degree in Electrical Engineering from Northeastern University, and a BEE from the University of Florida. He was the associate dean of the School of Engineering at the University of New Mexico from 1980–1982. He held summer faculty appointments at the University of the Balearic Islands, Spain from 1998–2002.