THE UNIVERSITY OF MICHIGAN
LIBRARIES

# VALUES OF MATERIAL AND PHYSICAL CONSTANTS

| Name | Symbol | Value | Units |
|---|---|---|---|
| Room temperature | $T$ | 300 (= 27°C) | K |
| Boltzman constant | $k$ | $1.38 \times 10^{-23}$ | J/K |
| Electron charge | $q$ | $1.6 \times 10^{-19}$ | C |
| Thermal voltage | $\phi_T = kT/q$ | 26 | mV (at 300 K) |
| Intrinsic Carrier Concentration (Silicon) | $n_i$ | $1.5 \times 10^{10}$ | $cm^{-3}$ (at 300 K) |
| Permittivity of Si | $\varepsilon_{si}$ | $1.05 \times 10^{-12}$ | F/cm |
| Permittivity of $SiO_2$ | $\varepsilon_{ox}$ | $3.5 \times 10^{-13}$ | F/cm |
| Resistivity of Al | $\rho_{Al}$ | $2.7 \times 10^{-8}$ | Ω-m |
| Resistivity of Cu | $\rho_{Cu}$ | $1.7 \times 10^{-8}$ | Ω-m |
| Magnetic permeability of vacuum (similar for $SiO_2$) | $\mu_0$ | $12.6 \times 10^{-7}$ | Wb/Am |
| Speed of light (in vacuum) | $c_0$ | 30 | cm/nsec |
| Speed of light (in $SiO_2$) | $c_{ox}$ | 15 | cm/nsec |

UNIVERSITY OF MICHIGAN LIBRARIES

# FORMULAS AND EQUATIONS

## Diode

$$I_D = I_S(e^{V_D/\phi_T} - 1) = Q_D/\tau_T$$

$$C_j = \frac{C_{j0}}{(1 - V_D/\phi_0)^m}$$

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)} \times [(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}]$$

## MOS Transistor

$$V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|})$$

$$I_D = \frac{k'_n}{2}\frac{W}{L}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \text{ (sat)}$$

$$I_D = \upsilon_{sat} C_{ox} W\left(V_{GS} - V_T - \frac{V_{DSAT}}{2}\right)(1 + \lambda V_{DS}) \text{ (velocity sat)}$$

$$I_D = k'_n \frac{W}{L}\left((V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right) \text{ (triode)}$$

$$I_D = I_S e^{\frac{V_{GS}}{nkT/q}}\left(1 - e^{-\frac{V_{DS}}{kT/q}}\right) \text{ (subthreshold)}$$

## Deep Submicron MOS Unified Model

$$I_D = 0 \text{ for } V_{GT} \le 0$$

$$I_D = k'\frac{W}{L}\left(V_{GT}V_{min} - \frac{V_{min}^2}{2}\right)(1 + \lambda V_{DS}) \text{ for } V_{GT} \ge 0$$

with $V_{min} = \min(V_{GT}, V_{DS}, V_{DSAT})$

and $V_{GT} = V_{GS} - V_T$

## MOS Switch Model

$$R_{eq} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}(1 + \lambda V_{DD})} + \frac{V_{DD}/2}{I_{DSAT}(1 + \lambda V_{DD}/2)}\right) \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1 - \frac{5}{6}\lambda V_{DD}\right)$$

## Inverter

$$V_{OH} = f(V_{OL})$$

$$V_{OL} = f(V_{OH})$$

$$V_M = f(V_M)$$

$$t_p = 0.69 R_{eq} C_L = \frac{C_L(V_{swing}/2)}{I_{avg}}$$

$$P_{dyn} = C_L V_{DD} V_{swing} f$$

$$P_{stat} = V_{DD} I_{DD}$$

## Static CMOS Inverter

$$V_{OH} = V_{DD}$$

$$V_{OL} = GND$$

$$V_M \approx \frac{rV_{DD}}{1 + r} \quad \text{with} \quad r = \frac{k_p V_{DSATp}}{k_n V_{DSATn}}$$

$$V_{IH} = V_M - \frac{V_M}{g} \qquad V_{IL} = V_M + \frac{V_{DD} - V_M}{g}$$

with $g \approx \dfrac{1 + r}{(V_M - V_{Tn} - V_{DSATn}/2)(\lambda_n - \lambda_p)}$

$$t_p = \frac{t_{pHL} + t_{pLH}}{2} = 0.69 C_L\left(\frac{R_{eqn} + R_{eqp}}{2}\right)$$

$$P_{av} = C_L V_{DD}^2 f$$

## Interconnect

Lumped $RC$: $t_p = 0.69\ RC$

Distributed $RC$: $t_p = 0.38\ RC$

$RC$-chain:

$$\tau_N = \sum_{i=1}^{N} R_i \sum_{j=i}^{N} C_j = \sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j$$

Transmission line reflection:

$$\rho = \frac{V_{refl}}{V_{inc}} = \frac{I_{refl}}{I_{inc}} = \frac{R - Z_0}{R + Z_o}$$

# DIGITAL INTEGRATED CIRCUITS

A DESIGN PERSPECTIVE

**Prentice Hall Electronics and VLSI Series**

Charles S. Sodini, Series Editor

LEE, SHUR, FIELDLY, YTTERDAL *Semiconductor Devices Modeling for VLSI*

LEUNG *VLSI for Wireless Communications*

PLUMMER, DEAL, GRIFFIN *Silicon VLSI Technology: Fundamentals, Practice, and Modeling*

RABAEY, CHANDRAKASAN, NIKOLIĆ *Digital Integrated Circuits: A Design Perspective*, Second Edition

# DIGITAL INTEGRATED CIRCUITS

## A DESIGN PERSPECTIVE

SECOND EDITION

JAN M. RABAEY
ANANTHA CHANDRAKASAN
BORIVOJE NIKOLIĆ

PRENTICE HALL ELECTRONICS AND VLSI SERIES
CHARLES G. SODINI, SERIES EDITOR

Pearson Education

Pearson Education, Inc.
Upper Saddle River, New Jersey 07458

**Library of Congress Cataloging-in-Publication Data on file.**

Vice President and Editorial Director, ECS: *Marcia J. Horton*
Publisher: *Tom Robbins*
Editorial Assistant: *Eric Van Ostenbridge*
Vice President and Director of Production and Manufacturing, ESM: *David W. Riccardi*
Executive Managing Editor: *Vince O'Brien*
Managing Editor: *David A. George*
Production Editor: *Daniel Sandin*
Director of Creative Services: *Paul Belfanti*
Creative Director: *Carole Anson*
Art and Cover Director: *Jayne Conte*
Art Editor: *Greg Dulles*
Manufacturing Manager: *Trudy Pisciotti*
Manufacturing Buyer: *Lisa McDowell*
Marketing Manager: *Holly Stark*

*About the Cover*: Detail of "Wet Orange," by Joan Mitchell (American, 1925–1992). Oil on canvas, 112 × 245 in. (284.5 × 622.3 cm). Carnegie Museum of Art, Pittsburgh, PA. Gift of Kaufmann's Department Store and the National Endowment for the Arts, 74.11. Photograph by Peter Harholdt, 1995.

© 2003, 1996 by Pearson Education, Inc.
Pearson Education, Inc.
Upper Saddle River, NJ 07458

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher shall not be liable in any event for incidental and consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-597444-5

Pearson Education Ltd., *London*
Pearson Education Australia Pty, Ltd., *Sydney*
Pearson Education Singapore, Pte. Ltd.
Pearson Education North Asia Ltd., *Hong Kong*
Pearson Education Canada Inc., *Toronto*
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan, *Tokyo*
Pearson Education Malaysia, Pte. Ltd.
Pearson Education Inc., *Upper Saddle River, New Jersey*

*To Kathelijn, Karthiyayani, Krithivasan, and our Parents*

*"Qu'est-ce que l'homme dans la nature?*
*Un néant a l'égard de l'infini,*
*un tout al l'égard du néant,*
*un milieu entre rien et tout."*

*"What is man in nature?*
*Nothing in relation to the infinite,*
*everything in relation to nothing,*
*a mean between nothing and everything."*

Blaise Pascal, *Pensées*, n. 4, 1670.

# Preface

## What is New?

Welcome to second edition of "*Digital Integrated Circuits: A Design Perspective.*" In the six years since the publication of the first, the field of digital integrated circuits has gone through some dramatic evolutions and changes. IC manufacturing technology has continued to scale to ever-smaller dimensions. Minimum feature sizes have scaled by a factor of almost ten since the writing of the first edition, and now are approaching the 100 nm realm. This scaling has a double impact on the design of digital integrated circuit. First of all, the complexity of the designs that can be put on a single die has increased dramatically. Dealing with the challenges this poses has led to new design methodologies and implementation strategies. At the same time, the plunge into the deep-submicron space causes devices to behave differently, and brings to the forefront a number of new issues that impact the reliability, cost, performance, and power dissipation of the digital IC. Addressing these issues in-depth is what differentiates this edition from the first.

A glance through the table of contents reveals extended coverage of issues such as deep-sub micron devices, circuit optimization, interconnect modeling and optimization, signal integrity, clocking and timing, and power dissipation. All these topics are illustrated with state-of-the-art design examples. Also, since MOS now represents more than 99% of the digital IC market, older technologies such as silicon bipolar and GaAs have been deleted (however, the interested reader can find the old chapters on these technologies on the web site of the book). Given the importance of methodology in today's design process, we have included *Design Methodology Inserts* throughout the text, each of which highlights one particular aspect of the design process. This new edition represents a major reworking of the book. The biggest change is the addition of two co-authors, Anantha and Bora, who have brought a broader insight into digital IC design and its latest trends and challenges.

## Maintaining the Spirit of the First Edition

While introducing these changes, our intent has been to preserve the spirit and goals of the first edition—that is, to bridge the gap between the **circuit and system visions** on digital design. While starting from a solid understanding of the operation of electronic devices and an in-depth analysis of the nucleus of digital design—the inverter—we gradually channel this knowledge into the design of more complex modules such as gates, registers, controllers, adders, multipliers, and memories. We identify the compelling questions facing the designers of today's



Dell Ex. 1025
Page 10

complex circuits: What are the dominant design parameters, what section of the design should he focus on and what details could she ignore? Simplification is clearly the only approach to address the increasing complexity of the digital systems. However, oversimplification can lead to circuit failure since global circuit effects such as timing, interconnect, and power consumption are ignored. To avoid this pitfall it is important to design digital circuits with both a circuits and a systems perspective in mind. This is the approach taken in this book, which brings the reader the knowledge and expertise needed to deal with complexity, using both analytical and experimental techniques.

## How to Use This Book

The core of the text is intended for use in a *senior-level digital circuit design class*. Around this kernel, we have included chapters and sections covering the more advanced topics. In the course of developing this book, it quickly became obvious that it is difficult to define a subset of the digital circuit design domain that covers everyone's needs. On the one hand, a newcomer to the field needs detailed coverage of the basic concepts. On the other hand, feedback from early readers and reviewers indicated that an in-depth and extensive coverage of advanced topics and current issues is desirable and necessary. Providing this complete vision resulted in a text that exceeds the scope of a single-semester class. The more advanced material can be used as the basis for a *graduate class*. The wide coverage and the inclusion of state-of-the-art topics also makes the text useful as a reference work for professional engineers. It is assumed that students taking this course are familiar with the basics of logic design.

The organization of the material is such that the chapters can be taught or read in many ways, as long as a number of precedence relations are adhered to. The core of the text consists of Chapters 5, 6, 7, and 8. Chapters 1 to 4 can be considered as introductory. In response to popular demand, we have introduced a short treatise on semiconductor manufacturing in Chapter 2. Students with a prior introduction to semiconductor devices can traverse quickly through Chapter 3. We urge everyone to do at least that, as a number of important notations and foundations are introduced in that chapter. In addition, an original approach to the modeling of deep-submicron transistors enabling manual analysis, is introduced. To emphasize the importance of interconnect in today's digital design, we have moved the modeling of interconnect forward in the text to Chapter 4.

Chapters 9 to 12 are of a more advanced nature and can be used to provide a certain focus to the course. A course with a focus on the circuit aspects, for example, can supplement the core material with Chapters 9 and 12. A course focused on the digital system design should consider adding (parts of) Chapters 9, 10, and 11. All of these advanced chapters can be used to form the core of a graduate or a follow-on course. Sections considered *advanced* are marked with an *asterisk* in the text.

A number of possible paths through the material for a senior-level class are enumerated below. In the *instructor documentation,* provided on the book's web site, we have included a number of complete syllabi based on courses run at some academic institutions.

*Basic circuit class (with minor prior device knowledge):*
1, 2.1–3, 3, 4, 5, 6, 7, 8, (9.1–9.3, 12).

*Somewhat more advanced circuit coverage:*
1, (2, 3), 4, 5, 6, 7, 8, 9, 10.1–10.3, 10.5–10.6, 12.

*Course with systems focus:*
1, (2, 3), 4, 5, 6, 7, 8, 9, 10.1–10.4, 11, 12.1–12.2.

The *design methodology inserts* are, by preference, covered in concurrence with the chapter to which they are attached.

In order to maintain a consistent flow through each of the chapters, the topics are *introduced* first, followed by a detailed and in-depth discussion of the ideas. A *Perspective* section discusses how the introduced concepts relate to real world designs and how they might be impacted by future evolutions. Each chapter finishes with a *Summary,* which briefly enumerates the topics covered in the text, followed by *To Probe Further* and *Reference* sections. These provide ample references and pointers for a reader interested in further details on some of the material.

As the title of the book implies, one of the goals of this book is to stress the design aspect of digital circuits. To achieve this more practical viewpoint and to provide a real perspective, we have interspersed actual *design examples* and layouts throughout the text. These case studies help to answer questions, such as "How much area or speed or power is really saved by applying this technique?" To mimic the real design process, we are making extensive use of design tools such as circuit- and switch-level simulation as well as layout editing and extraction. Computer analysis is used throughout to verify manual results, to illustrate new concepts, or to examine complex behavior beyond the reach of manual analysis.

Finally, to facilitate the learning process, there are numerous examples included in the text. Each chapter contains a number of *problems or brain-teasers* (answers for which can be found in the back of the book), that provoke thinking and understanding while reading.

## The Worldwide Web Companion

A worldwide web companion (***http://bwrc.eecs.berkeley.edu/IcBook/index.htm***) provides fully worked-out design problems and a complete set of overhead transparencies, extracting the most important figures and graphs from the text.

In contrast to the first edition, we have chosen **NOT to include problems sets** and **design problems** in the text. Instead we decided to make them available **on the book's web site**. This gives us the opportunity to dynamically upgrade and extend the problems, providing a more effective tool for the instructor. More than 300 challenging *exercises* are currently provided. The goal is to provide the individual reader an independent gauge for his understanding of the material and to provide practice in the use of some of the design tools. Each problem is keyed to the text sections it refers to (e.g., <1.3>), the design tools that must be used when solving the problem (e.g., SPICE) and a rating, ranking the problems on difficulty: (E) easy, (M) moderate, and

(C) challenging. Problems marked with a (D) include a design or research elements. Solutions to the problem sets are available only to instructors of academic institutions that have chosen to adopt our book for classroom use. They are available through the publisher on a password-protected web site.

Open-ended *design problems* help to gain the all-important insight into design optimization and trade-off. The use of design editing, verification and analysis tools is recommended when attempting these design problems. Fully worked out versions of these problems can be found on the web site.

In addition, the book's web site also offers samples of hardware and software laboratories, extra background information, and useful links.

## Compelling Features of the Book

- Brings both circuit and systems views on design together. It offers a profound understanding of the design of complex digital circuits, while preparing the designer for new challenges that might be waiting around the corner.
- Design-oriented perspectives are advocated throughout. Design challenges and guidelines are highlighted. Techniques introduced in the text are illustrated with real designs and complete SPICE analysis.
- Is the first circuit design book that *focuses solely on deep-submicron devices.* To facilitate this, a simple transistor model for manual analysis, called the *unified MOS model*, has been developed.
- Unique in showing how to use the latest techniques to design complex high-performance or low-power circuits. Speed and power treated as equal citizens throughout the text.
- Covers crucial real-world system design issues such as signal integrity, power dissipation, interconnect, packaging, timing, and synchronization.
- Provides unique coverage of the latest design methodologies and tools, with a discussion of how to use them from a designers' perspective.
- Offers perspectives on how digital circuit technology might evolve in the future.
- Outstanding illustrations and a usable design-oriented four-color insert.
- To Probe Further and Reference sections provide ample references and pointers for a reader interested in further details on some of the material.
- Extensive instructional package is available over the internet from the author's web site. Includes design software, transparency masters, problem sets, design problems, actual layouts, and hardware and software laboratories.

## The Contents at a Glance

A quick scan of the table of contents shows how the ordering of chapters and the material covered are consistent with the advocated design methodology. Starting from a model of the semiconductor devices, we will gradually progress upwards, covering the inverter, the complex logic gate (NAND, NOR, XOR), the functional (adder, multiplier, shifter, register) and the system

module (datapath, controller, memory) levels of abstraction. For each of these layers, the dominant design parameters are identified and simplified models are constructed, abstracting away the nonessential details. While this layered modeling approach is the designer's best handle on complexity, it has some pitfalls. This is illustrated in Chapters 9 and 10, where topics with a global impact, such as interconnect parasitics and chip timing, are discussed. To further express the dichotomy between circuit and system design visions, we have divided the book contents into two major parts: Part II (Chapters 4–7) addresses mostly the circuit perspective of digital circuit design, while Part III (Chapters 8–12) presents a more system oriented vision. Part I (Chapters 1-4) provides the necessary foundation (design metrics, the manufacturing process, device and interconnect models).

**Chapter 1** serves as a global *introduction*. After a historical overview of digital circuit design, the concepts of hierarchical design and the different abstraction layers are introduced. A number of fundamental metrics, which help to quantify cost, reliability, and performance of a design, are introduced.

**Chapter 2** provides a short and compact introduction to the *MOS manufacturing process*. Understanding the basic steps in the process helps to create the three-dimensional understanding of the MOS transistor, which is crucial when identifying the sources of the device parasitics. Many of the variations in device parameters can also be attributed to the manufacturing process as well. The chapter further introduces the concept of design rules, which form the interface between the designer and the manufacturer. The chapter concludes with an overview of the chip packaging process, an often-overlooked but crucial element of the digital IC design cycle.

**Chapter 3** contains a summary of the primary design building blocks, *the semiconductor devices*. The main goal of this chapter is to provide an intuitive understanding of the operation of the MOS as well as to introduce the device models, which are used extensively in the later chapters. Major attention is paid to the artifacts of modern submicron devices, and the modeling thereof. Readers with prior device knowledge can traverse this material rather quickly.

**Chapter 4** contains a careful analysis of the *wire,* with interconnect and its accompanying parasitics playing a major role. We visit each of the parasitics that come with a wire (capacitance, resistance, and inductance) in turn. Models for both manual and computer analysis are introduced.

**Chapter 5** deals with the nucleus of digital design, the *inverter*. First, a number of fundamental properties of digital gates are introduced. These parameters, which help to quantify the performance and reliability of a gate, are derived in detail for two representative inverter structures: the static complementary CMOS. The techniques and approaches introduced in this chapter are of crucial importance, as they are repeated over and over again in the analysis of other gate structures and more complex gate structures.

In **Chapter 6** this fundamental knowledge is extended to address the design of *simple and complex digital CMOS gates*, such as NOR and NAND structures. It is demonstrated that, depending upon the dominant design constraint (reliability, area, performance, or power), other

CMOS gate structures besides the complementary static gate can be attractive. The properties of a number of contemporary gate-logic families are analyzed and compared. Techniques to optimize the performance and power consumption of complex gates are introduced.

**Chapter 7** discusses how memory function can be accomplished using either positive feedback or charge storage. Besides analyzing the traditional bistable flip-flops, other sequential circuits such as the mono- and astable multivibrators are also introduced. All chapters prior to Chapter 7 deal exclusively with combinational circuits, that is circuits without a sense of the past history of the system. *Sequential logic circuits*, in contrast, can remember and store the past state.

All chapters preceding **Chapter 8** present a circuit-oriented approach towards digital design. The analysis and optimization process has been constrained to the individual gate. In this chapter, we take our approach one step further and analyze how gates can be connected together to form the building blocks of a system. The system-level part of the book starts, appropriately, with a discussion of *design methodologies*. Design automation is the only way to cope with the ever-increasing complexity of digital designs. In Chapter 8, the prominent ways of producing large designs in a limited time are discussed. The chapter spends considerable time on the different implementation methodologies available to today's designer. Custom versus semi-custom, hardwired versus fixed, regular array versus ad-hoc are some of the issues put forward.

**Chapter 9** revisits the impact of *interconnect wiring* on the functionality and performance of a digital gate. A wire introduces parasitic capacitive, resistive, and inductive effects, which are becoming ever more important with the scaling of the technology. Approaches to minimize the impact of these interconnect parasitics on performance, power dissipation and circuit reliability are introduced. The chapter also addresses some important issues such as supply-voltage distribution, and input/output circuitry.

In **Chapter 10** details how that in order to operate sequential circuits correctly, a strict ordering of the switching events has to be imposed. Without these *timing* constraints, wrong data might be written into the memory cells. Most digital circuits use a synchronous, clocked approach to impose this ordering. In Chapter 10, the different approaches to digital circuit timing and clocking are discussed. The impact of important effects such as clock skew on the behavior of digital synchronous circuits is analyzed. The synchronous approach is contrasted with alternative techniques, such as self-timed circuits. The chapter concludes with a short introduction to synchronization and clock-generation circuits.

In **Chapter 11**, the design of a variety of complex *arithmetic building blocks* such as adders, multipliers, and shifters, is discussed. This chapter is crucial because it demonstrates how the design techniques introduced in chapters 5 and 6 are extended to the next abstraction layer. The concept of the critical path is introduced and used extensively in the performance analysis and optimization. Higher-level performance models are derived. These help the designer to get a fundamental insight into the operation and quality of a design module, without having to resort to an in-depth and detailed analysis of the underlying circuitry.

**Chapter 12** discusses in depth the different memory classes and their implementation. Whenever large amounts of data storage are needed, the digital designer resorts to special circuit modules, called *memories*. Semiconductor memories achieve very high storage density by compromising on some of the fundamental properties of digital gates. Instrumental in the design of reliable and fast memories is the implementation of the peripheral circuitry, such as the decoders, sense amplifiers, drivers, and control circuitry, which are extensively covered. Finally, as the primary issue in memory design is to ensure that the device works consistently under all operating circumstances, the chapter concludes with a detailed discussion of memory reliability. This chapter as well as the previous one are optional for undergraduate courses.

## Acknowledgments

The authors would like to thank all those who contributed to the emergence, creation and correction of this manuscript. First of all, thanks to all the graduate students that helped over the years to bring the text to where it is today. Thanks also to the students of the eecs141 and eecs241 courses at Berkeley and the 6.374 course at MIT, who suffered through many of the experimental class offerings based on this book. The feedback from instructors, engineers, and students from all over the world has helped tremendously in focusing the directions of this new edition, and in fine-tuning the final text. The continuous stream of e-mails indicate to us that we are on the right track.

In particular, we would like to acknowledge the contributions of Mary-Jane Irwin, Vijay Narayanan, Eby Friedman, Fred Rosenberger, Wayne Burleson, Sekhar Borkar, Ivo Bolsens, Duane Boning, Olivier Franza, Lionel Kimerling, Josie Ammer, Mike Sheets, Tufan Karalar, Huifang Qin, Rhett Davis, Nathan Chan, Jeb Durant, Andrei Vladimirescu, Radu Zlatanovici, Yasuhisa Shimazaki, Fujio Ishihara, Dejan Markovic, Vladimir Stojanovic, SeongHwan Cho, James Kao, Travis Simpkins, Siva Narendra, James Goodman, Vadim Gutnik, Theodoros Konstantakopoulos, Rex Min, Vikas Mehrotra, and Paul-Peter Sotiriadis. Their help, input, and feedback are greatly appreciated. Obviously, we remain thankful to those who helped create and develop the first edition.

I am extremely grateful to the staff at Prentice Hall, who have been instrumental in turning a rough manuscript into an enjoyable book. First of all, I would like to acknowledge the help and constructive feedback of Tom Robbins, Publisher, Daniel Sandin, Production Editor, and David George, Managing Editor. A special word of thanks to Brenda Vanoni at Berkeley, for her invaluable help in the copy editing and the website creation process. The web expertise of Carol Sitea came in very handy as well.

I would like to highlight to role of computer aids in developing this manuscript. All drafts were completely developed on the FrameMaker publishing system (Adobe Systems). Graphs were mostly created using MATLAB. Microsoft Frontpage is the tool of choice for the web-page creation. For circuit simulations, we used HSPICE (Avant!). All layouts were generated using the Cadence physical design suite.

Finally, some words of gratitude to the people that had to endure the creation process of this book, Kathelijn, Karthiyayani, Krithivasan, and Rebecca. While the generation of a new edition brings substantially less pain than a first edition, we consistently underestimate what it takes, especially in light of the rest of our daily loads. They been a constant support, help and encouragement during the writing of this manuscript.

JAN M. RABAEY
ANANTHA CHANDRAKASAN
BORIVOJE NIKOLIĆ
*Berkeley, Calistoga, Cambridge*

# Contents


**Preface** **vii**

***Part 1*** **The Fabrics** **1**

***Chapter 1*** **Introduction** **3**

1.1 A Historical Perspective 4
1.2 Issues in Digital Integrated Circuit Design 6
1.3 Quality Metrics of a Digital Design 15
1.3.1 Cost of an Integrated Circuit 16
1.3.2 Functionality and Robustness 18
1.3.3 Performance 27
1.3.4 Power and Energy Consumption 30
1.4 Summary 31
1.5 To Probe Further 31
Reference Books 32
References 33

***Chapter 2*** **The Manufacturing Process** **35**

2.1 Introduction 36
2.2 Manufacturing CMOS Integrated Circuits 36
2.2.1 The Silicon Wafer 37
2.2.2 Photolithography 37
2.2.3 Some Recurring Process Steps 41
2.2.4 Simplified CMOS Process Flow 42
2.3 Design Rules—The Contract between Designer and Process Engineer 47
2.4 Packaging Integrated Circuits 51
2.4.1 Package Materials 52
2.4.2 Interconnect Levels 53
2.4.3 Thermal Considerations in Packaging 59
2.5 Perspective—Trends in Process Technology 61
2.5.1 Short-Term Developments 61
2.5.2 In the Longer Term 63
2.6 Summary 64




Dell Ex. 1025
Page 18

2.7 To Probe Further 64
References 64

*Design Methodology Insert A* **IC LAYOUT** **67**
A.1 To Probe Further 71
References 71

*Chapter 3* **The Devices** **73**
3.1 Introduction 74
3.2 The Diode 74
3.2.1 A First Glance at the Diode—The Depletion Region 75
3.2.2 Static Behavior 77
3.2.3 Dynamic, or Transient, Behavior 80
3.2.4 The Actual Diode—Secondary Effects 84
3.2.5 The SPICE Diode Model 85
3.3 The MOS(FET) Transistor 87
3.3.1 A First Glance at the Device 87
3.3.2 The MOS Transistor under Static Conditions 88
3.3.3 The Actual MOS Transistor—Some Secondary Effects 114
3.3.4 SPICE Models for the MOS Transistor 117
3.4 A Word on Process Variations 120
3.5 Perspective—Technology Scaling 122
3.6 Summary 128
3.7 To Probe Further 129
References 130

*Design Methodology Insert B* **Circuit Simulation** **131**
References 134

*Chapter 4* **The Wire** **135**
4.1 Introduction 136
4.2 A First Glance 136
4.3 Interconnect Parameters—Capacitance, Resistance, and Inductance 138
4.3.1 Capacitance 138
4.3.2 Resistance 144
4.3.3 Inductance 148
4.4 Electrical Wire Models 150
4.4.1 The Ideal Wire 151
4.4.2 The Lumped Model 151
4.4.3 The Lumped *RC* Model 152
4.4.4 The Distributed *rc* Line 156
4.4.5 The Transmission Line 159

4.5 SPICE Wire Models 170
4.5.1 Distributed *rc* Lines in SPICE 170
4.5.2 Transmission Line Models in SPICE 170
4.5.3 Perspective: A Look into the Future 171
4.6 Summary 174
4.7 To Probe Further 174
References 174

***Part 2*** **A Circuit Perspective** **177**

***Chapter 5*** **The CMOS Inverter** **179**
5.1 Introduction 180
5.2 The Static CMOS Inverter—An Intuitive Perspective 180
5.3 Evaluating the Robustness of the CMOS Inverter: The Static Behavior 184
5.3.1 Switching Threshold 185
5.3.2 Noise Margins 188
5.3.3 Robustness Revisited 191
5.4 Performance of CMOS Inverter: The Dynamic Behavior 193
5.4.1 Computing the Capacitances 194
5.4.2 Propagation Delay: First-Order Analysis 199
5.4.3 Propagation Delay from a Design Perspective 203
5.5 Power, Energy, and Energy Delay 213
5.5.1 Dynamic Power Consumption 214
5.5.2 Static Consumption 223
5.5.3 Putting It All Together 225
5.5.4 Analyzing Power Consumption Using SPICE 227
5.6 Perspective: Technology Scaling and its Impact on the Inverter Metrics 229
5.7 Summary 232
5.8 To Probe Further 233
References 233

***Chapter 6*** **Designing Combinational Logic Gates in CMOS** **235**
6.1 Introduction 236
6.2 Static CMOS Design 236
6.2.1 Complementary CMOS 237
6.2.2 Ratioed Logic 263
6.2.3 Pass-Transistor Logic 269
6.3 Dynamic CMOS Design 284
6.3.1 Dynamic Logic: Basic Principles 284
6.3.2 Speed and Power Dissipation of Dynamic Logic 287

6.3.3 Signal Integrity Issues in Dynamic Design 290
6.3.4 Cascading Dynamic Gates 295
6.4 Perspectives 303
6.4.1 How to Choose a Logic Style? 303
6.4.2 Designing Logic for Reduced Supply Voltages 303
6.5 Summary 306
6.6 To Probe Further 307
References 308

***Design Methodology Insert C*** **How to Simulate Complex Logic Circuits** **309**
C.1 Representing Digital Data as a Continuous Entity 310
C.2 Representing Data as a Discrete Entity 310
C.3 Using Higher-Level Data Models 315
References 317

***Design Methodology Insert D*** **Layout Techniques for Complex Gates** **319**

***Chapter 7*** **Designing Sequential Logic Circuits** **325**
7.1 Introduction 326
7.1.1 Timing Metrics for Sequential Circuits 327
7.1.2 Classification of Memory Elements 328
7.2 Static Latches and Registers 330
7.2.1 The Bistability Principle 330
7.2.2 Multiplexer-Based Latches 332
7.2.3 Master-Slave Edge-Triggered Register 333
7.2.4 Low-Voltage Static Latches 339
7.2.5 Static SR Flip-Flops—Writing Data by Pure Force 341
7.3 Dynamic Latches and Registers 344
7.3.1 Dynamic Transmission-Gate Edge-triggered Registers 344
7.3.2 $C^2$MOS—A Clock-Skew Insensitive Approach 346
7.3.3 True Single-Phase Clocked Register (TSPCR) 350
7.4 Alternative Register Styles* 354
7.4.1 Pulse Registers 354
7.4.2 Sense-Amplifier-Based Registers 356
7.5 Pipelining: An Approach to Optimize Sequential Circuits 358
7.5.1 Latch- versus Register-Based Pipelines 360
7.5.2 NORA-CMOS—A Logic Style for Pipelined Structures 361
7.6 Nonbistable Sequential Circuits 364
7.6.1 The Schmitt Trigger 364
7.6.2 Monostable Sequential Circuits 367
7.6.3 Astable Circuits 368
7.7 Perspective: Choosing a Clocking Strategy 370
7.8 Summary 371

7.9 To Probe Further 372
References 372

*Part 3* **A System Perspective** **375**

*Chapter 8* **Implementation Strategies for Digital ICS** **377**
8.1 Introduction 378
8.2 From Custom to Semicustom and Structured-Array Design Approaches 382
8.3 Custom Circuit Design 383
8.4 Cell-Based Design Methodology 384
8.4.1 Standard Cell 385
8.4.2 Compiled Cells 390
8.4.3 Macrocells, Megacells and Intellectual Property 392
8.4.4 Semicustom Design Flow 396
8.5 Array-Based Implementation Approaches 399
8.5.1 Prediffused (or Mask-Programmable) Arrays 399
8.5.2 Prewired Arrays 404
8.6 Perspective—The Implementation Platform of the Future 420
8.7 Summary 423
8.8 To Probe Further 423
References 424

*Design Methodology Insert E* **Characterizing Logic and Sequential Cells** **427**
References 434

*Design Methodology Insert F* **Design Synthesis** **435**
References 443

*Chapter 9* **Coping with Interconnect** **445**
9.1 Introduction 446
9.2 Capacitive Parasitics 446
9.2.1 Capacitance and Reliability—Cross Talk 446
9.2.2 Capacitance and Performance in CMOS 449
9.3 Resistive Parasitics 460
9.3.1 Resistance and Reliability—Ohmic Voltage Drop 460
9.3.2 Electromigration 462
9.3.3 Resistance and Performance—*RC* Delay 464
9.4 Inductive Parasitics* 469
9.4.1 Inductance and Reliability— Voltage Drop 469
9.4.2 Inductance and Performance—Transmission-line Effects 475
9.5 Advanced Interconnect Techniques 480

9.5.1 Reduced-Swing Circuits 480
9.5.2 Current-Mode Transmission Techniques 486
9.6 Perspective: Networks-on-a-Chip 487
9.7 Summary 488
9.8 To Probe Further 489
References 489

*Chapter 10* **Timing Issues in Digital Circuits** **491**
10.1 Introduction 492
10.2 Timing Classification of Digital Systems 492
10.2.1 Synchronous Interconnect 492
10.2.2 Mesochronous interconnect 493
10.2.3 Plesiochronous Interconnect 493
10.2.4 Asynchronous Interconnect 494
10.3 Synchronous Design—An In-depth Perspective 495
10.3.1 Synchronous Timing Basics 495
10.3.2 Sources of Skew and Jitter 502
10.3.3 Clock-Distribution Techniques 508
10.3.4 Latch-Based Clocking* 516
10.4 Self-Timed Circuit Design* 519
10.4.1 Self-Timed Logic—An Asynchronous Technique 519
10.4.2 Completion-Signal Generation 522
10.4.3 Self-Timed Signaling 526
10.4.4 Practical Examples of Self-Timed Logic 531
10.5 Synchronizers and Arbiters* 534
10.5.1 Synchronizers—Concept and Implementation 534
10.5.2 Arbiters 538
10.6 Clock Synthesis and Synchronization Using a Phase-Locked Loop* 539
10.6.1 Basic Concept 540
10.6.2 Building Blocks of a PLL 542
10.7 Future Directions and Perspectives 546
10.7.1 Distributed Clocking Using DLLs 546
10.7.2 Optical Clock Distribution 548
10.7.3 Synchronous versus Asynchronous Design 549
10.8 Summary 550
10.9 To Probe Further 551
References 551

*Design Methodology Insert G* **Design Verification** **553**
References 557

*Chapter 11* **Designing Arithmetic Building Blocks** **559**
11.1 Introduction 560
11.2 Datapaths in Digital Processor Architectures 560
11.3 The Adder 561
11.3.1 The Binary Adder: Definitions 561
11.3.2 The Full Adder: Circuit Design Considerations 564
11.3.3 The Binary Adder: Logic Design Considerations 571
11.4 The Multiplier 586
11.4.1 The Multiplier: Definitions 586
11.4.2 Partial-Product Generation 587
11.4.3 Partial-Product Accumulation 589
11.4.4 Final Addition 593
11.4.5 Multiplier Summary 594
11.5 The Shifter 594
11.5.1 Barrel Shifter 595
11.5.2 Logarithmic Shifter 596
11.6 Other Arithmetic Operators 596
11.7 Power and Speed Trade-offs in Datapath Structures* 600
11.7.1 Design Time Power-Reduction Techniques 601
11.7.2 Run-Time Power Management 611
11.7.3 Reducing the Power in Standby (or Sleep) Mode 617
11.8 Perspective: Design as a Trade-off 618
11.9 Summary 619
11.10 To Probe Further 620
References 621

*Chapter 12* **Designing Memory and Array Structures** **623**
12.1 Introduction 624
12.1.1 Memory Classification 625
12.1.2 Memory Architectures and Building Blocks 627
12.2 The Memory Core 634
12.2.1 Read-Only Memories 634
12.2.2 Nonvolatile Read-Write Memories 647
12.2.3 Read-Write Memories (RAM) 657
12.2.4 Contents-Addressable or Associative Memory (CAM) 670
12.3 Memory Peripheral Circuitry* 672
12.3.1 The Address Decoders 672
12.3.2 Sense Amplifiers 679
12.3.3 Voltage References 686
12.3.4 Drivers/Buffers 689
12.3.5 Timing and Control 689

12.4 Memory Reliability and Yield* 693
12.4.1 Signal-to-Noise Ratio 693
12.4.2 Memory Yield 698
12.5 Power Dissipation in Memories* 701
12.5.1 Sources of Power Dissipation in Memories 701
12.5.2 Partitioning of the Memory 702
12.5.3 Addressing the Active Power Dissipation 702
12.5.4 Data-Retention Dissipation 704
12.5.5 Summary 707
12.6 Case Studies in Memory Design 707
12.6.1 The Programmable Logic Array (PLA) 707
12.6.2 A 4-Mbit SRAM 710
12.6.3 A 1-Gbit NAND Flash Memory 712
12.7 Perspective: Semiconductor Memory Trends and Evolutions 714
12.8 Summary 716
12.9 To Probe Further 717
References 718

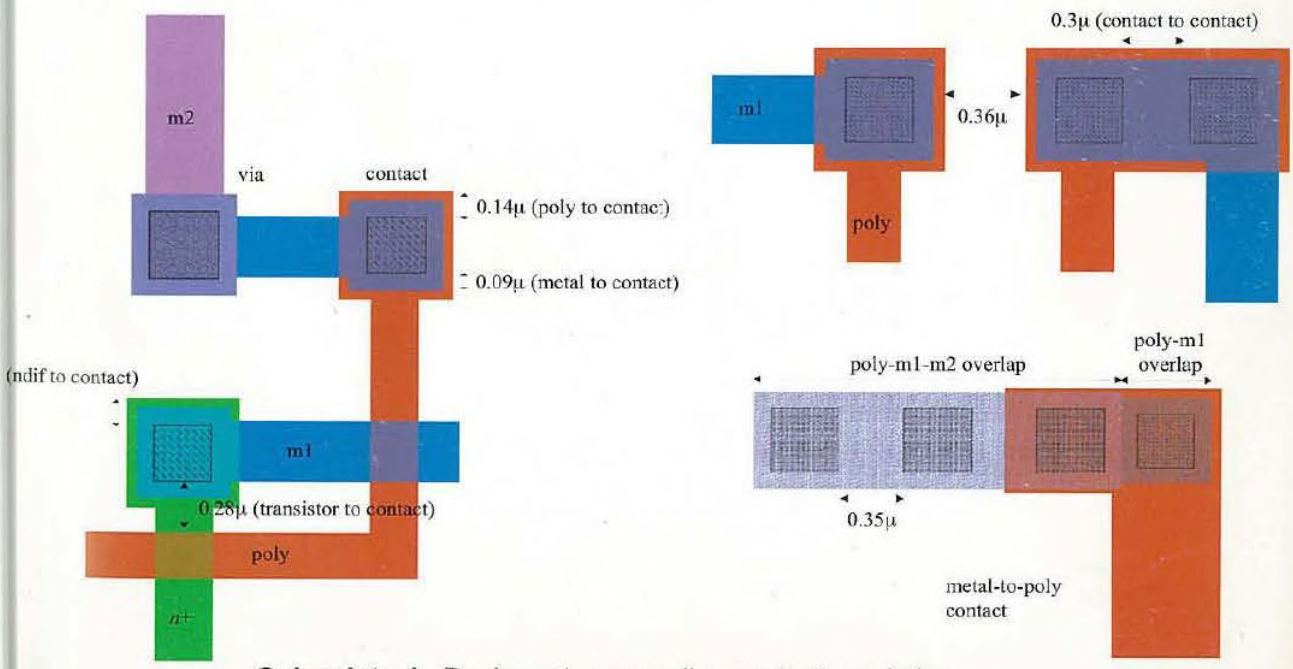
***Design Methodology Insert H*** **Validation and Test of Manufactured Circuits** **721**
H.1 Introduction 721
H.2 Test Procedure 722
H.3 Design for Testability 723
H.3.1 Issues in Design for Testability 723
H.3.2 Ad Hoc Testing 725
H.3.3 Scan-Based Test 726
H.3.4 Boundary-Scan Design 729
H.3.5 Built-in Self-Test (BIST) 730
H.4 Test-Pattern Generation 734
H.4.1 Fault Models 734
H.4.2 Automatic Test-Pattern Generation (ATPG) 736
H.4.3 Fault Simulation 737
H.5 To Probe Further 737
References 737

**Problem Solutions** **739**

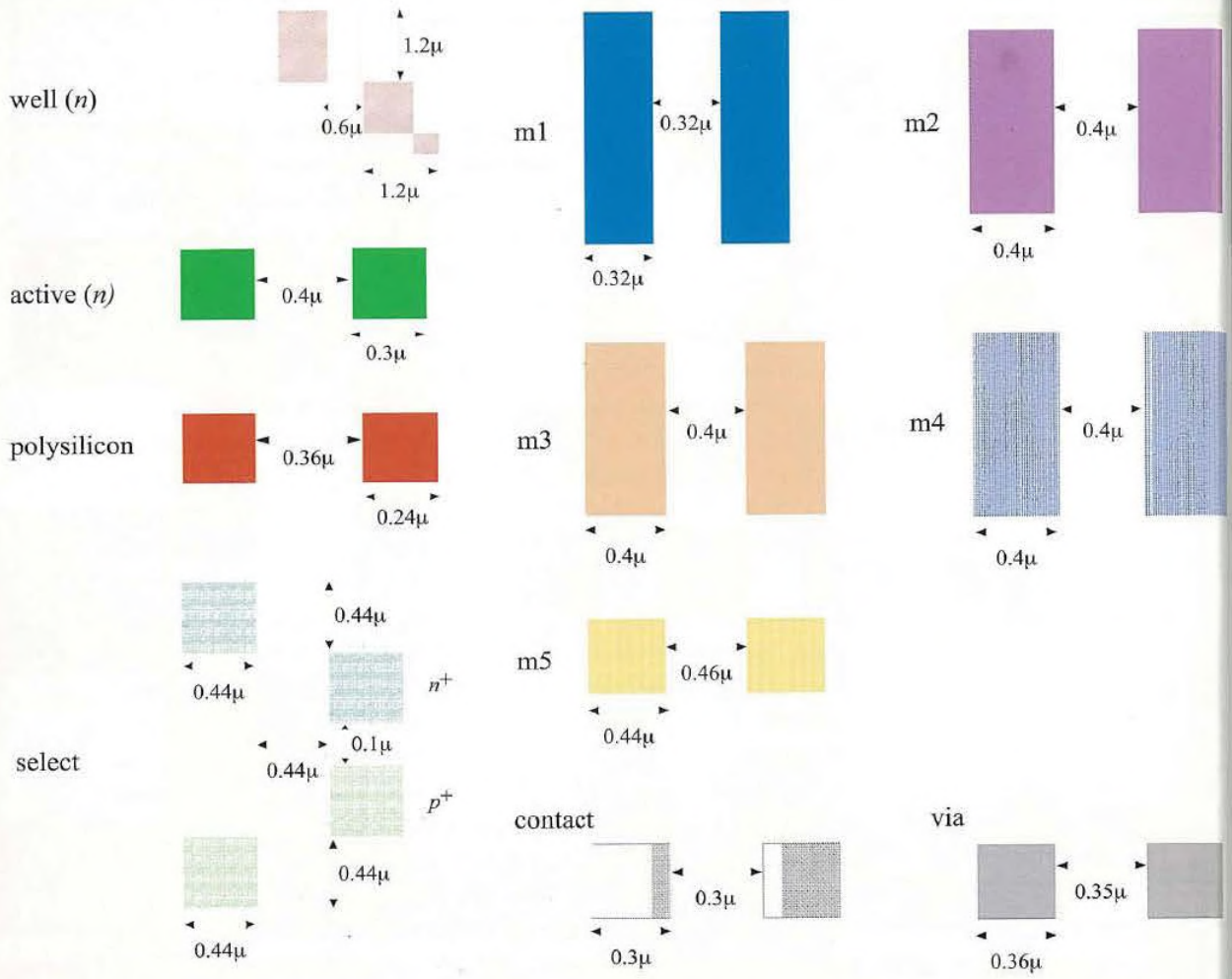
**Index** **745**

| Layer Description | Representation | | | | |
|---|---|---|---|---|---|
| metal | m1 | m2 | m3 | m4 | m5 |
| well | nw | | | | |
| polysilicon | poly | | | | |
| contacts & vias | ct | v12,v23,v34,v45 | nwc | pwc | |
| active area and FETs | ndif | pdif | nfet | pfet | |
| select | nplus | pplus | prb | | |

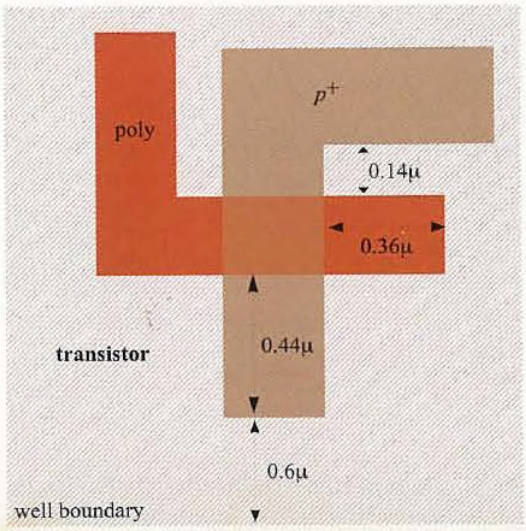
**Colorplate 1.** CMOS layers and representations (for vanilla 0.25 μm CMOS process)



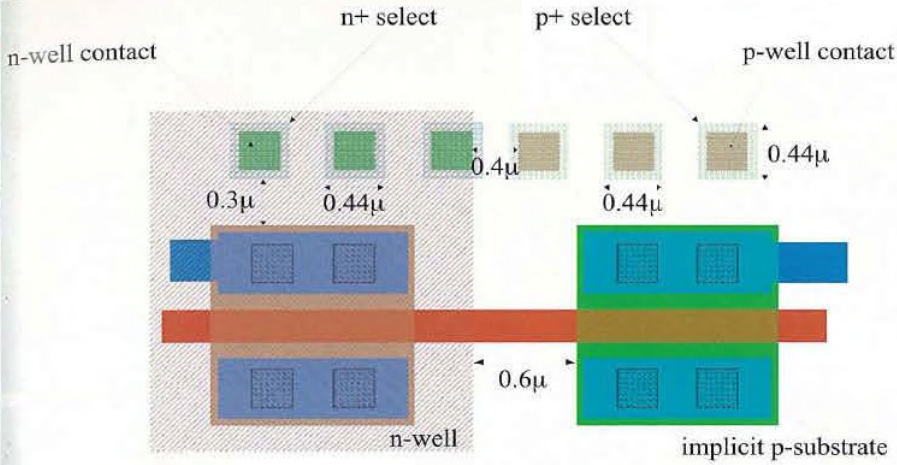

**Colorplate 4.** Design rules regarding contacts and vias. Overlapping layers are marked by merged colors.



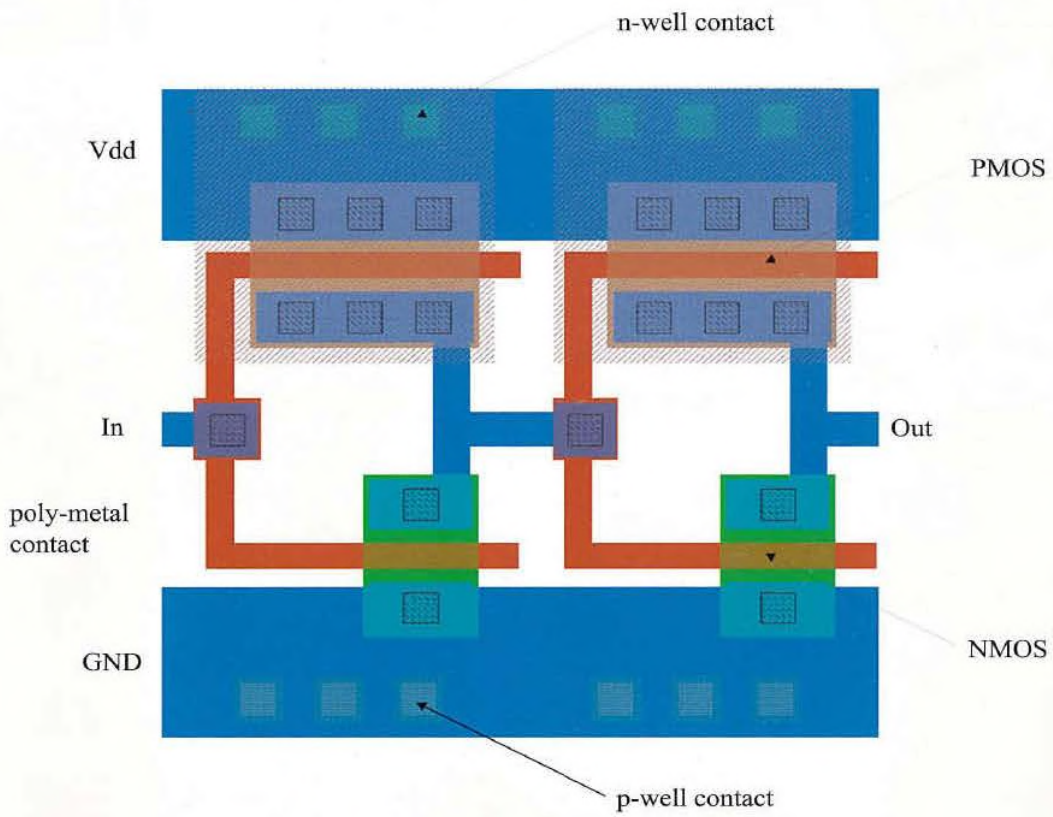

**Colorplate 2.** Intra-layer layout design rules, expressed as minimum dimensions and spacings.



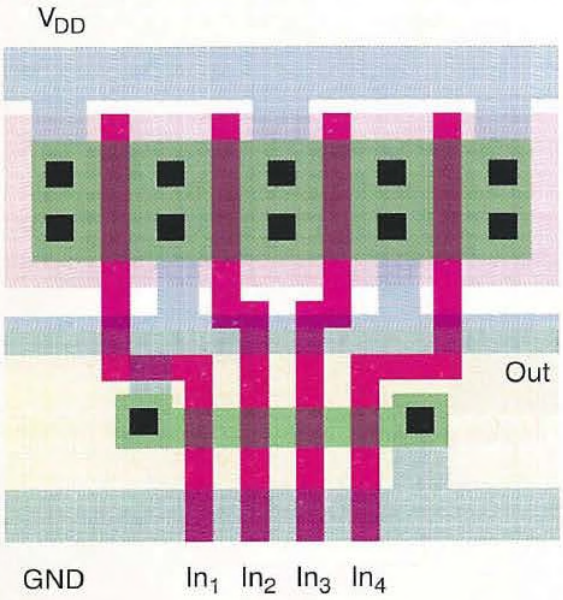

**Colorplate 3.** Design rules concerning transistor layout. The device shown is a PMOS transistor.



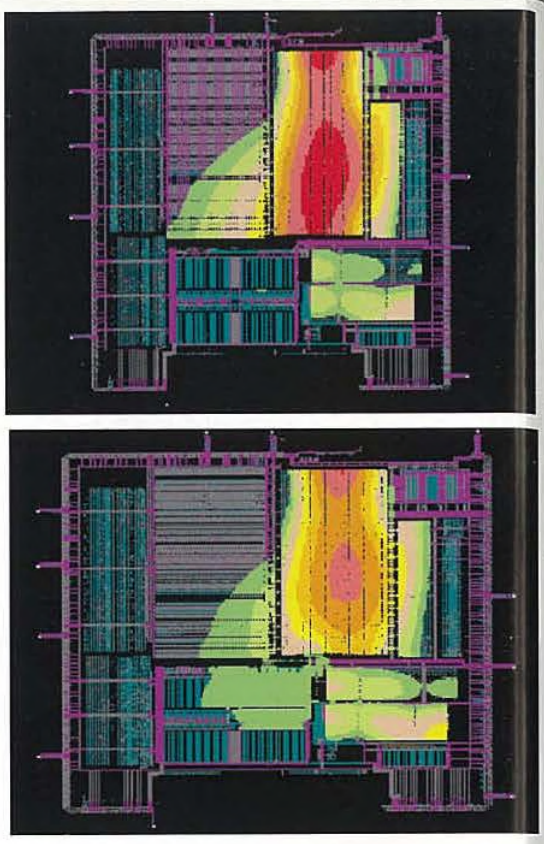

**Colorplate 5.** Design rules regarding well contacts and select layers.



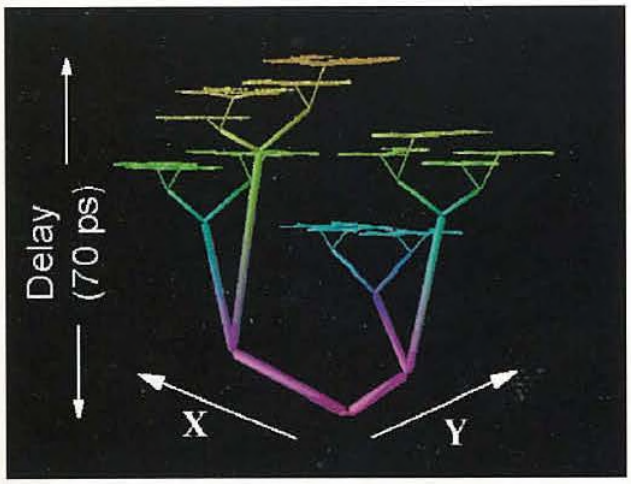

**Colorplate 6.** Layout of inverter in 0.25 μm CMOS technology.

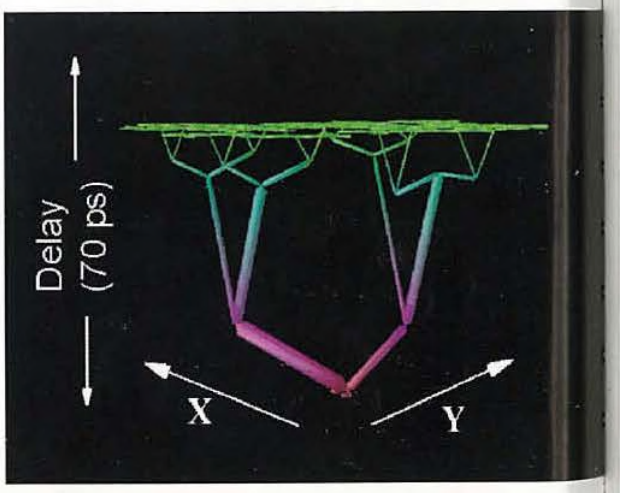



Colorplate 7. Four-input CMOS NAND.



Colorplate 8. Simulated IR voltage drop in power-distribution network.




Colorplate 9. Visualizations of clock delay in a tree network driving different loads.

CHAPTER

2

# The Manufacturing Process

*Overview of manufacturing process*

*Design rules*

*IC packaging*

*Future Trends in Integrated Circuit Technology*


2.1 Introduction

2.2 Manufacturing CMOS Integrated Circuits
- 2.2.1 The Silicon Wafer
- 2.2.2 Photolithography
- 2.2.3 Some Recurring Process Steps
- 2.2.4 Simplified CMOS Process Flow

2.3 Design Rules—Between the Designer and the Process Engineer

2.4 Packaging Integrated Circuits
- 2.4.1 Package Materials
- 2.4.2 Interconnect Levels
- 2.4.3 Thermal Considerations in Packaging

2.5 Perspective—Trends in Process Technology
- 2.5.1 Short-Term Developments
- 2.5.2 In the Longer Term

2.6 Summary

2.7 To Probe Further




Dell Ex. 1025
Page 30

## 2.1 Introduction

Most digital designers will never be confronted with the details of the manufacturing process that lay at the core of the semiconductor revolution. Still, some insight into the steps that lead to an operational silicon chip comes in quite handy in understanding the physical constraints imposed on a designer of an integrated circuit, as well as the impact of the fabrication process on issues such as cost.

In this chapter, we briefly describe the steps and techniques used in a modern integrated circuit manufacturing process. It is not our aim to present a detailed description of the fabrication technology, which easily deserves a complete course [Plummer00]. Rather, we aim at presenting the general outline of the flow and the interaction between the various steps. We learn that a set of *optical masks* forms the central interface between the intrinsics of the manufacturing process and the design that the user wants to see transferred to the silicon fabric. The masks define the patterns that, when transcribed onto the different layers of the semiconductor material, form the elements of the electronic devices and the interconnecting wires. As such, these patterns have to adhere to some constraints, in terms of minimum width and separation, if the resulting circuit is to be fully functional. This collection of constraints is called the *design rule set*, and acts as the contract between the circuit designer and the process engineer. If the designer adheres to these rules, he gets a guarantee that his circuit will be manufacturable. An overview of the common design rules encountered in modern CMOS processes is given, as well as a perspective on the *IC packaging* options. The package forms the interface between the circuit implemented on the silicon die and the outside world, and as such has a major impact on the performance, reliability, longevity, and cost of the integrated circuit.

## 2.2 Manufacturing CMOS Integrated Circuits

A simplified cross section of a typical CMOS inverter is shown in Figure 2-1. The CMOS process requires that both *n*-channel (NMOS) and *p*-channel (PMOS) transistors be built in the same silicon material. To accommodate both types of devices, special regions called *wells* must be created in which the semiconductor material is opposite to the type of the channel. A PMOS transistor has to be created in either an *n*-type substrate or an *n*-well, while an NMOS device resides in either a *p*-type substrate or a *p*-well. The cross section shown in Figure 2-1 features an

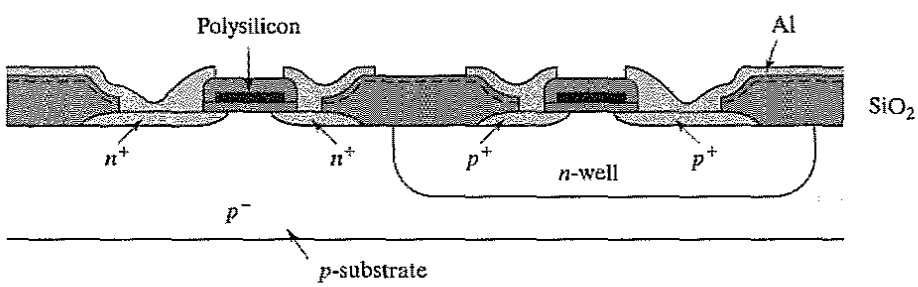



**Figure 2-1** Cross section of an *n*-well CMOS process.

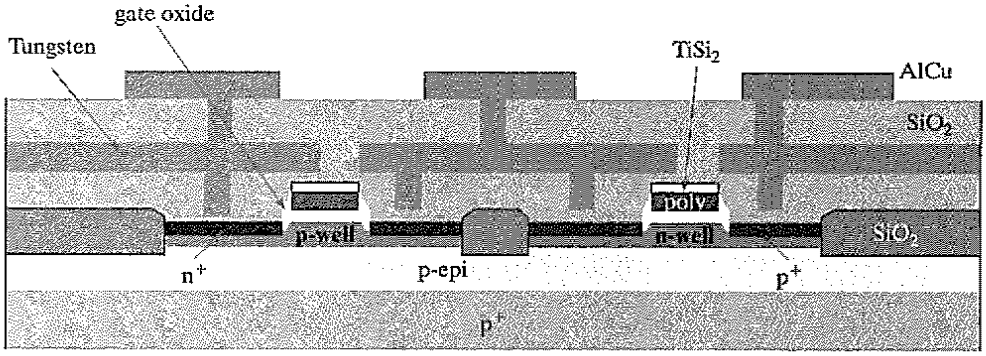



**Figure 2-2** Cross section of modern dual-well CMOS process.

*n*-well CMOS process, where the NMOS transistors are implemented in the *p*-doped substrate, and the PMOS devices are located in the *n*-well. Modern processes are increasingly using a *dual-well* approach that uses both *n*- and *p*-wells, grown on top of an epitaxial layer, as shown in Figure 2-2.

The CMOS process requires a large number of steps, each of which consists of a sequence of basic operations. A number of these steps and/or operations are executed very repetitively in the course of the manufacturing process. Rather than immediately delving into a description of the overall process flow, we first discuss the starting material followed by a detailed perspective on some of the most frequently recurring operations.

### 2.2.1 The Silicon Wafer

The base material for the manufacturing process comes in the form of a single-crystalline, lightly doped *wafer*. These wafers have typical diameters between 4 and 12 inches (10 and 30 cm, respectively) and a thickness of, at most 1 mm. They are obtained by cutting a single-crystal ingot into thin slices (see Figure 2-3). A starting wafer of the $p^-$-type might be doped around the levels of $2 \times 10^{21}$ impurities/m$^3$. Often, the surface of the wafer is doped more heavily, and a single crystal *epitaxial layer* of the opposite type is grown over the surface before the wafers are handed to the processing company. One important metric is the defect density of the base material. High defect densities lead to a larger fraction of nonfunctional circuits, and consequently an increase in cost of the final product.

### 2.2.2 Photolithography

In each processing step, a certain area on the chip is masked out using the appropriate optical mask so that a desired processing step can be selectively applied to the remaining regions. The processing step can be any of a wide range of tasks, including oxidation, etching, metal and polysilicon deposition, and ion implantation. The technique to accomplish this selective masking, called *photolithography*, is applied throughout the manufacturing process. Figure 2-4 gives

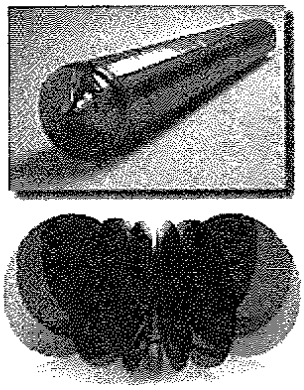


**Figure 2-3** Single-crystal ingot and sliced wafers (from [Fullman99]).

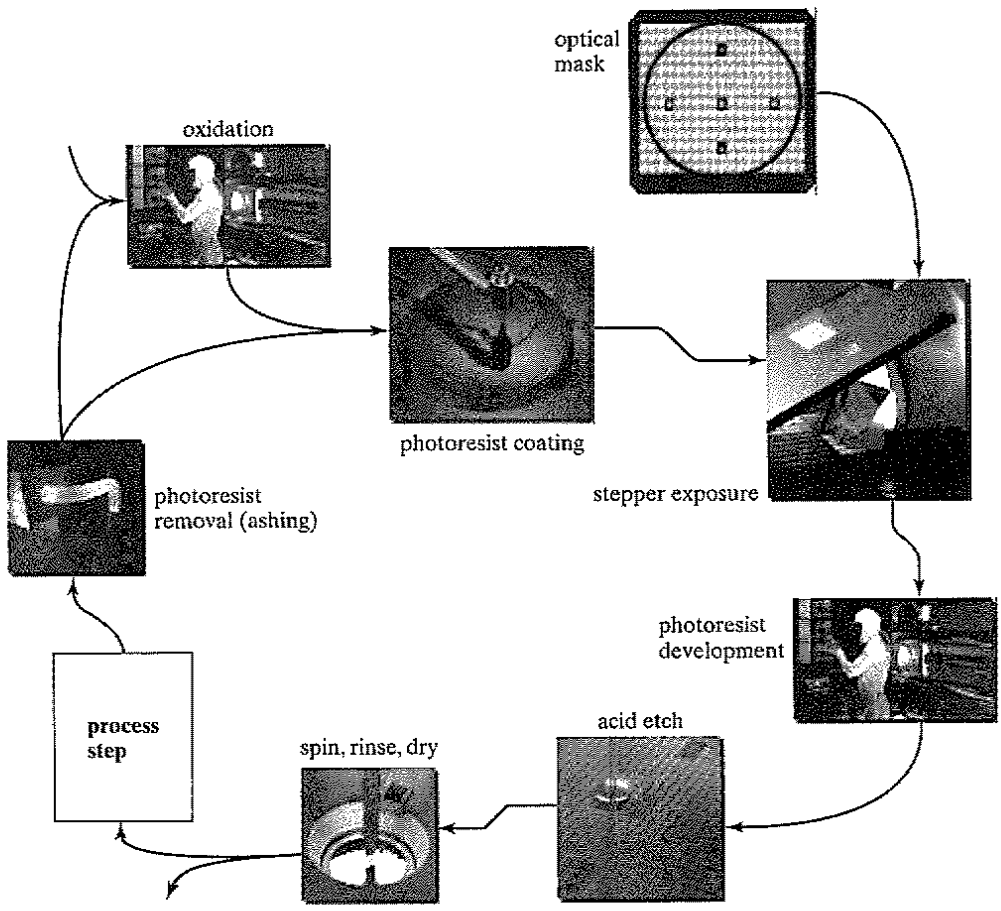



**Figure 2-4** Typical operations in a single photolithographic cycle (from [Fullman99]).

a graphical overview of the different operations involved in a typical photolithographic process. The following steps can be identified:

1. *Oxidation layering*—this optional step deposits a thin layer of $SiO_2$ over the complete wafer by exposing it to a mixture of high-purity oxygen and hydrogen at approximately 1000°C. The oxide is used as an insulation layer and also forms transistor gates.
2. *Photoresist coating*—a light-sensitive polymer (similar to latex) is evenly applied to a thickness of approximately 1 μm by spinning the wafer. This material is originally soluble in an organic solvent, but has the property that the polymers cross-link when exposed to light, making the affected regions insoluble. A photoresist of this type is called *negative*. A positive photoresist has the opposite properties; originally insoluble, but soluble after exposure. By using both positive and negative resists, a single mask can sometimes be used for two steps, making complementary regions available for processing. Since the cost of a mask is increasing quite rapidly with the scaling of technology, reducing the number of masks surely is a high priority.
3. *Stepper exposure*—a glass mask (or reticle) containing the patterns that we want to transfer to the silicon is brought in close proximity to the wafer. The mask is opaque in the regions that we want to process, and transparent in the others (assuming a negative photoresist). The glass mask can be thought of as the negative of one layer of the microcircuit. The combination of mask and wafer is now exposed to ultraviolet light. Where the mask is transparent, the photoresist becomes insoluble.
4. *Photoresist development and bake*—the wafers are developed in either an acid or base solution to remove the nonexposed areas of photoresist. Once the exposed photoresist is removed, the wafer is "soft baked" at a low temperature to harden the remaining photoresist.
5. *Acid etching*—material is selectively removed from areas of the wafer that are not covered by photoresist. This is accomplished through the use of many different types of acid, base and caustic solutions as a function of the material that is to be removed. Much of the work with chemicals takes place at large wet benches where special solutions are prepared for specific tasks. Because of the dangerous nature of some of these solvents, safety and environmental impact is a primary concern.
6. *Spin, rinse, and dry*—a special tool (called SRD) cleans the wafer with deionized water and dries it with nitrogen. The microscopic scale of modern semiconductor devices means that even the smallest particle of dust or dirt can destroy the circuitry. To prevent this from happening, the processing steps are performed in ultraclean rooms where the number of dust particles per cubic foot of air ranges between 1 and 10. Automatic wafer handling and robotics are used whenever possible. This explains why the cost of a state-of-the-art fabrication facility easily reaches multiple billions of dollars. Even then, the wafers must be constantly cleaned to avoid contamination and to remove the leftover of the previous process steps.

7. *Various process steps*—the exposed area can now be subjected to a wide range of process steps, such as ion implantation, plasma etching, or metal deposition. These are the subjects of the subsequent section.
8. *Photoresist removal (or ashing)*—a high-temperature plasma is used to selectively remove the remaining photoresist without damaging device layers.

In Figure 2-5, we illustrate the use of the photolithographic process for one specific example, the patterning of a layer of $SiO_2$. The sequence of process steps shown in the figure patterns exactly one layer of the semiconductor material and may seem very complex. Yet, the reader has to bear in mind that the same sequence patterns the layer of **the complete surface of the wafer.** Hence, it is a very parallel process, transferring hundreds of millions of patterns to the semiconductor surface simultaneously. The concurrent and scalable nature of the optolithographical process is what makes the cheap manufacturing of complex semiconductor circuits possible, and lies at the core of the economic success of the semiconductor industry.

The continued scaling of the minimum feature sizes in integrated circuits puts an enormous burden on the developer of semiconductor manufacturing equipment. This is especially true for the optolithographical process. The dimensions of the features to be transcribed surpass the wavelengths of the optical light sources, so that achieving the necessary resolution and accuracy

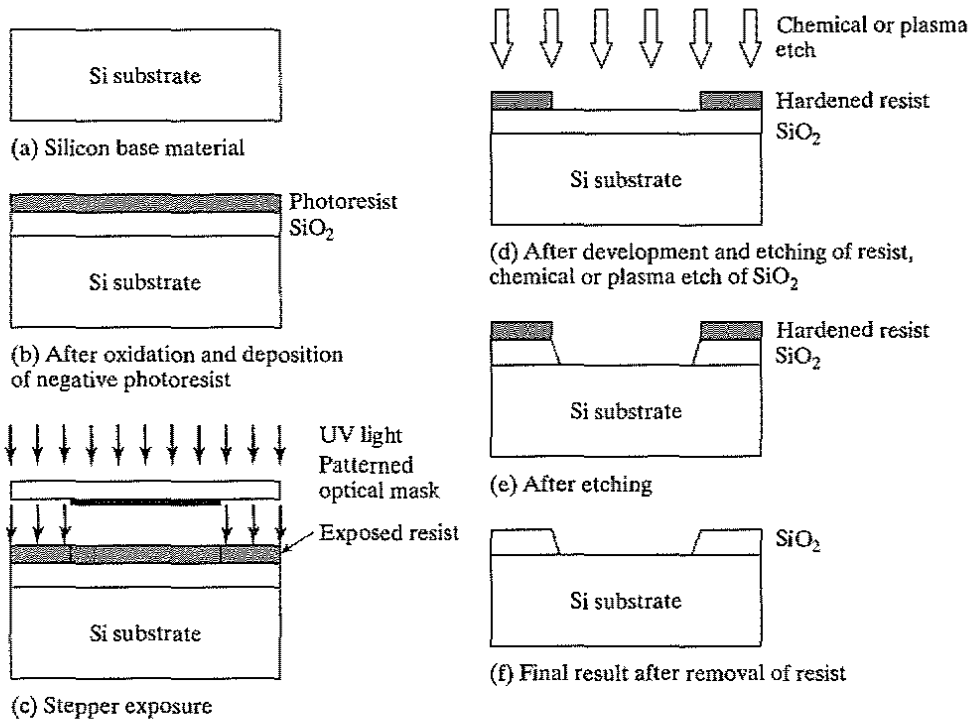



**Figure 2-5** Process steps for patterning of $SiO_2$.

becomes more and more difficult. So far, electrical engineering has extended the lifetime of this process at least until the 100 nm (or 0.1 μm) process generation. Techniques such as *optical mask correction* (OPC) prewarp the drawn patterns to account for the diffraction phenomena, encountered when printing close to the wavelength of the available optical source. This adds substantially to the cost of mask making. In the foreseeable future, other solutions that offer a finer resolution, such as extreme ultraviolet (EUV), X ray, or electron beam, may be needed. These techniques, while fully functional, are currently less attractive from an economic viewpoint.

### 2.2.3 Some Recurring Process Steps

**Diffusion and Ion Implantation**

Many steps of the integrated circuit manufacturing process require a change in the dopant concentration of some parts of the material. Examples include the creation of the source and drain regions, well and substrate contacts, the doping of the polysilicon, and the adjustments of the device threshold. Two approaches exist for introducing these dopants—diffusion and ion implantation. In both techniques, the area to be doped is exposed, while the rest of the wafer is coated with a layer of buffer material, typically $SiO_2$.

In *diffusion implantation*, the wafers are placed in a quartz tube embedded in a heated furnace. A gas containing the dopant is introduced in the tube. The high temperatures of the furnace, typically 900 to 1100 °C, cause the dopants to diffuse into the exposed surface both vertically and horizontally. The final dopant concentration is the greatest at the surface and decreases in a gaussian profile deeper in the material.

In *ion implantation*, dopants are introduced as ions into the material. The ion implantation system directs and sweeps a beam of purified ions over the semiconductor surface. The acceleration of the ions determines how deep they will penetrate the material, while the beam current and the exposure time determine the dosage. The ion implantation method allows for an independent control of depth and dosage. This is the reason that ion implantation has largely displaced diffusion in modern semiconductor manufacturing.

Ion implantation has some unfortunate side effects, however, the most important one being lattice damage. Nuclear collisions during the high energy implantation cause the displacement of substrate atoms, leading to material defects. This problem is largely resolved by applying a subsequent *annealing* step, in which the wafer is heated to around 1000°C for 15 to 30 minutes, and then allowed to cool slowly. The heating step thermally vibrates the atoms, which allows the bonds to reform.

**Deposition**

Any CMOS process requires the repetitive deposition of layers of a material over the complete wafer, to either act as buffers for a processing step, or as insulating or conducting layers. We have already discussed the oxidation process, which allows a layer of $SiO_2$ to be grown. Other materials require different techniques. For instance, silicon nitride ($Si_3N_4$) is used as a sacrificial buffer material during the formation of the field oxide and the introduction of the stopper

implants. This silicon nitride is deposited everywhere using a process called *chemical vapor deposition* or CVD. This process is based on a gas-phase reaction, with energy supplied by heat at around 850°C.

Polysilicon, on the other hand, is deposited using a chemical deposition process, which flows silane gas over the heated wafer coated with $SiO_2$ at a temperature of approximately 650°C. The resulting reaction produces a noncrystalline or amorphous material called *polysilicon*. To increase the conductivity of the material, the deposition has to be followed by an implantation step.

The Aluminum interconnect layers typically are deployed using a process known as *sputtering*. The aluminum is evaporated in a vacuum, with the heat for the evaporation delivered by electron-beam or ion-beam bombarding. Other metallic interconnect materials such as Copper require different deposition techniques.

**Etching**

Once a material has been deposited, etching is used selectively to form patterns such as wires and contact holes. We already discussed the *wet etching* process, which makes use of acid or basic solutions. Hydrofluoric acid buffered with ammonium fluoride typically is used to etch $SiO_2$, for example.

In recent years, *dry* or *plasma etching* has advanced substantially. A wafer is placed into the etch tool's processing chamber and given a negative electrical charge. The chamber is heated to 100°C and brought to a vacuum level of 7.5 Pa, then filled with a positively charged plasma (usually a mix of nitrogen, chlorine, and boron trichloride). The opposing electrical charges cause the rapidly moving plasma molecules to align themselves in a vertical direction, forming a microscopic chemical and physical "sandblasting" action which removes the exposed material. Plasma etching has the advantage of offering a well-defined directionality to the etching action, creating patterns with sharp vertical contours.

**Planarization**

To reliably deposit a layer of material onto the semiconductor surface, it is essential that the surface be approximately flat. If special steps were not taken, this would definitely present problems in modern CMOS processes, where multiple patterned metal interconnect layers are superimposed onto each other. Therefore, a *chemical–mechanical planarization* (CMP) step is included before the deposition of an extra metal layer on top of the insulating $SiO_2$ layer. This process uses a slurry compound—a liquid carrier with a suspended abrasive component such as aluminum oxide or silica—to microscopically plane a device layer and to reduce the step heights.

### 2.2.4 Simplified CMOS Process Flow

The gross outline of a potential CMOS process flow is given in Figure 2-6. The process starts with the definition of the *active regions*—these are the regions where transistors will be constructed. All other areas of the die will be covered with a thick layer of silicon dioxide ($SiO_2$)

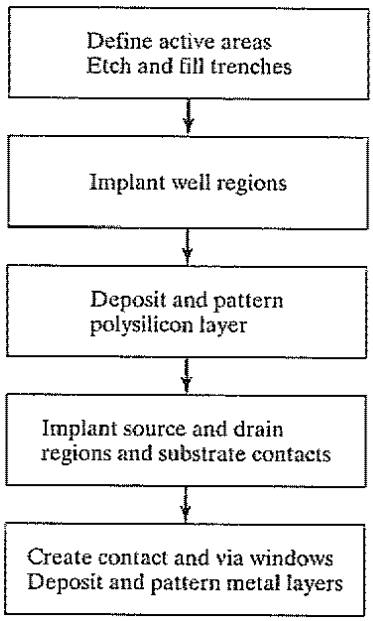



**Figure 2-6** Simplified process sequence for the manufacturing of a n-dual-well CMOS circuit.

called the *field oxide*. This oxide acts as the insulator between neighboring devices, and it is either grown (as in the process of Figure 2-1) or deposited in etched trenches (Figure 2-2)—hence, the name *trench insulation*. Further insulation is provided by the addition of a reverse-biased *np*-diode, formed by adding an extra $p^+$ region called the *channel-stop implant* (or *field implant*) underneath the field oxide. Next, lightly doped *p*- and *n*-wells are formed through ion implantation. To construct an NMOS transistor in a *p*-well, heavily doped *n*-type *source* and *drain* regions are implanted (or diffused) into the lightly doped *p*-type substrate. A thin layer of $SiO_2$ called the *gate oxide* separates the region between the source and drain, and is itself covered by conductive polycrystalline silicon (or *polysilicon*, for short). The conductive material forms the *gate* of the transistor. PMOS transistors are constructed in an *n*-well in a similar fashion (just reverse *n*'s and *p*'s). Multiple insulated layers of metallic (most often Aluminum) wires are deposited on top of these devices to provide for the necessary interconnections between the transistors.

A more detailed breakdown of the flow into individual process steps and their impact on the semiconductor material is shown graphically in Figure 2-7. While most of the operations should be self-explanatory in light of the previous descriptions, some comments on individual operations are worthwhile. The process starts with a *p*-substrate surfaced with a lightly doped *p*-epitaxial layer (a). A thin layer of $SiO_2$ is then deposited, which will serve as the gate oxide for the transistors, followed by a deposition of a thicker sacrificial silicon nitride layer (b). A plasma etching step using the complementary of the active area mask creates the trenches used for insulating the devices (c). After providing the channel stop implant, the trenches are filled

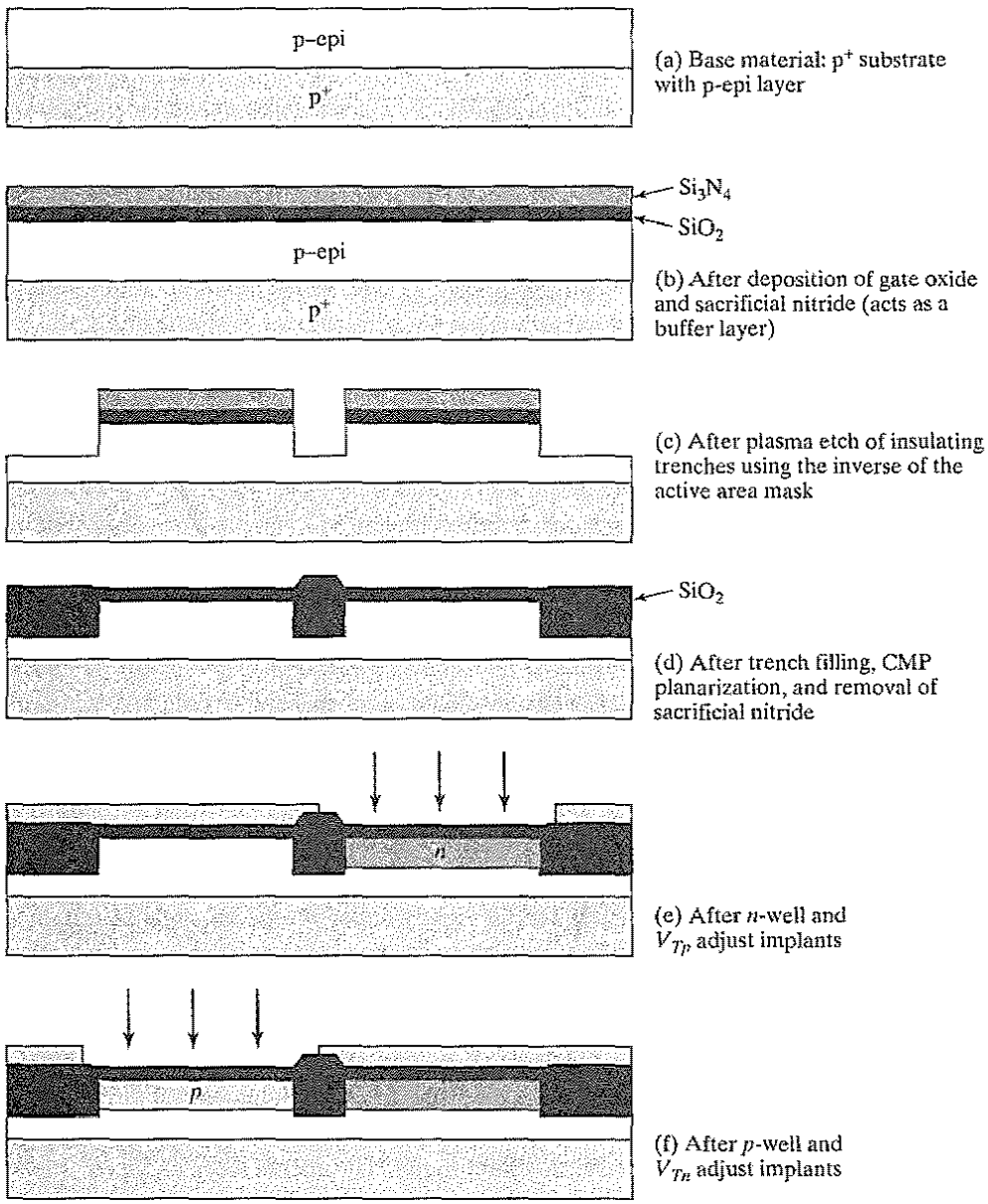



**Figure 2-7** Process flow for the fabrication of an NMOS and a PMOS transistor in a dual-well CMOS process. Be aware that the drawings are stylized for understanding and that the aspects ratios are not proportioned to reality.

with $SiO_2$ followed by a number of steps to provide a flat surface (including inverse active pattern oxide etching, and chemical–mechanical planarization). At that point, the sacrificial nitride is removed (d). The $n$-well mask is used to expose only the $n$-well areas (the rest of the wafer is covered by a thick buffer material), after which an implant-annealing sequence is applied to

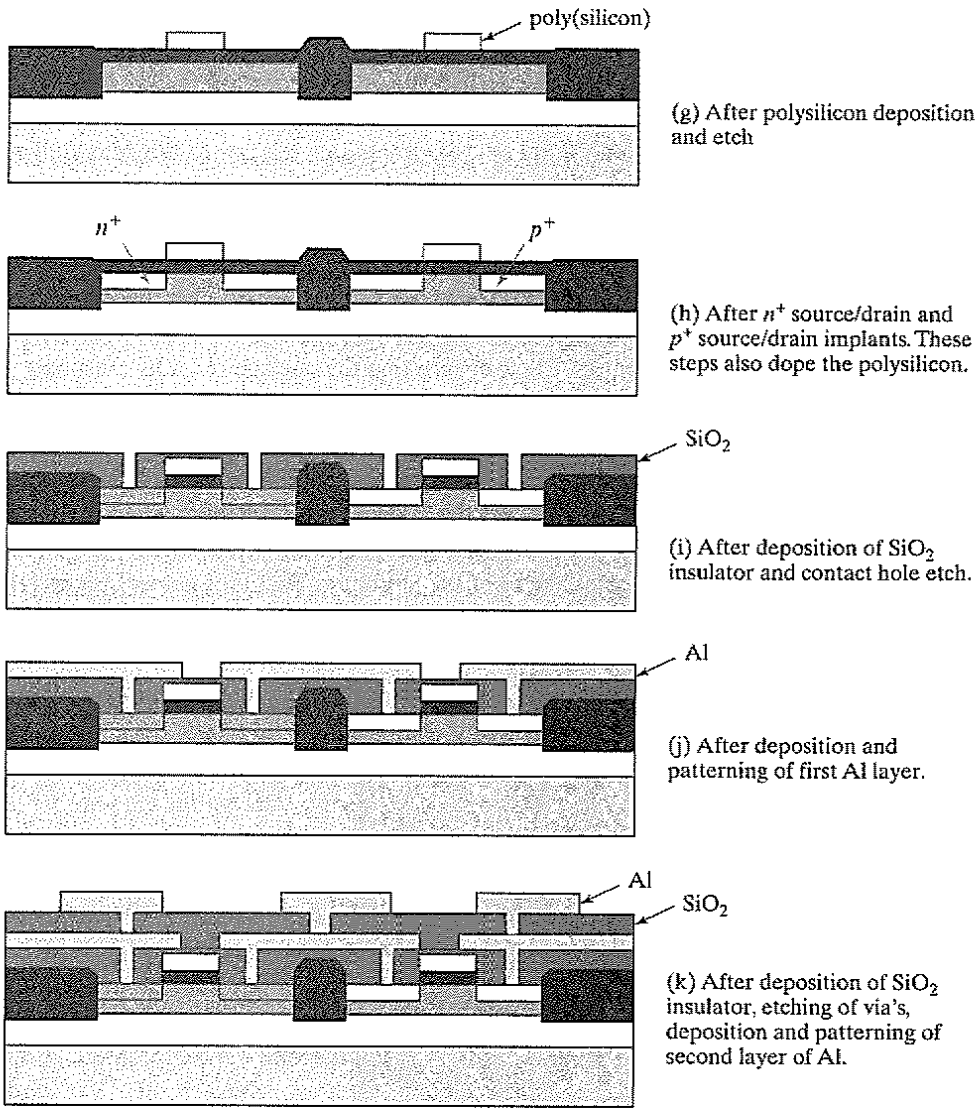



**Figure 2-7** *(cont.)*

adjust the well-doping. This is followed by a second implant step to adjust the threshold voltages of the PMOS transistors. This implant only impacts the doping in the area just below the gate oxide (e). Similar operations (using other dopants) are performed to create the *p*-wells, and to adjust the thresholds of the NMOS transistors (f). A thin layer of polysilicon is chemically deposited and patterned with the aid of the polysilicon mask. Polysilicon is used both as gate electrode material for the transistors and as an interconnect medium (g). Consecutive ion implantations are used to dope the source and drain regions of the PMOS ($p^+$) and NMOS ($n^+$) transistors, respectively (h), after which the thin gate oxide not covered by the polysilicon is

etched away.[1] The same implants also are used to dope the polysilicon on the surface, reducing its resistivity. Undoped polysilicon has a very high resistivity. Note that the polysilicon gate, which is patterned before the doping, actually defines the precise location of the channel region, and thus the location of the source and drain regions. This procedure, called the *self-aligned process*, allows for a very precise positioning of the two regions relative to the gate. Self-alignment is instrumental in reducing parasitic capacitances in the transistor. The process continues with the deposition of the metallic interconnect layers. These consist of a repetition of the following steps (i–k): deposition of the insulating material (most often $SiO_2$), etching of the contact or via holes, deposition of the metal (most often aluminum and copper, although tungsten often is used for the lower layers), and patterning of the metal. Intermediate planarization steps, using *chemical–mechanical polishing* or CMP, ensure that the surface remains reasonably flat, even in the presence of multiple interconnect layers. After the last level of metal is deposited, a final passivation or *overglass* is deposited for protection. This layer would be CVD $SiO_2$, although often an additional layer of nitride is deposited because it is more impervious to moisture. The final processing step etches openings to the pads used for bonding.

A cross section of the final artifact is shown in Figure 2-8. Observe how the transistors occupy only a small fraction of the total height of the structure. The interconnect layers take up the majority of the vertical dimension.

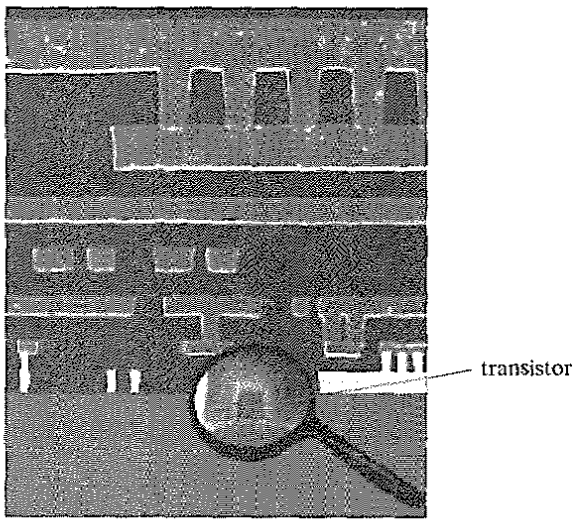



**Figure 2-8** Cross section of state-of-the-art CMOS process.

[1]Most modern processes also include extra implants for the creation of the lightly doped drain regions (LDD), and the creation of gate spacers at this point. We have omitted these for the sake of simplicity.

## 2.3 Design Rules—Between the Designer and the Process Engineer

As processes become more complex, requiring the designer to understand the intricacies of the fabrication process and interpret the relations between the different masks is a sure road to trouble. The goal of defining a set of design rules is to allow for a ready translation of a circuit concept into an actual geometry in silicon. The design rules act as the interface or even the contract between the circuit designer and the process engineer.

Circuit designers generally want tighter, smaller designs, which lead to higher performance and higher circuit density. The process engineer, on the other hand, wants a reproducible and high-yield process. Consequently, design rules are a compromise that attempts to satisfy both sides.

The design rules provide a set of guidelines for constructing the various masks needed in the patterning process. They consist of minimum-width and minimum-spacing constraints and requirements between objects on the same or different layers.

The fundamental unity in the definition of a set of design rules is the *minimum line width*. It stands for the minimum mask dimension that can be safely transferred to the semiconductor material. In general, the minimum line width is set by the resolution of the patterning process, which is most commonly based on optical lithography. More advanced approaches use electron-beam EUV, or X-ray sources, all of which offer a finer resolution, but currently they are less attractive from an economical standpoint.

Even for the same minimum dimension, design rules tend to differ from company to company, and from process to process. This makes porting an existing design between different processes a time-consuming task. One approach to address this issue is to use advanced CAD techniques, which allow for migration between compatible processes. Another approach is to use *scalable design rules*. The latter approach, made popular by Mead and Conway [Mead80], defines all rules as a function of a single parameter, most often called $\lambda$. The rules are chosen so that a design is easily ported over a cross section of industrial processes. Scaling of the minimum dimension is accomplished by simply changing the value of $\lambda$. This results in a *linear scaling* of all dimensions. For a given process, $\lambda$ is set to a specific value, and all design dimensions are consequently translated into absolute numbers. Typically, the minimum line width of a process is set to $2\lambda$. For instance, for a 0.25 μm process (i.e., a process with a minimum line width of 0.25 μm), $\lambda$ equals 0.125 μm.

This approach, while attractive, suffers from two disadvantages:

1. Linear scaling is possible only over a limited range of dimensions (for instance, between 0.25 μm and 0.18 μm). When scaling over larger ranges, the relations between the different layers tend to vary in a nonlinear way that cannot be adequately covered by the linear scaling rules.
2. Scalable design rules are conservative: They represent a cross section over different technologies, and they must represent the worst case rules for the whole set. This results in overdimensioned and less dense designs.

For these and other reasons, scalable design rules normally are avoided by industry.[2] As circuit density is a prime goal in industrial designs, most semiconductor companies tend to use *micron rules*, which express the design rules in absolute dimensions and therefore can exploit the features of a given process to a maximum degree. Scaling and porting designs between technologies under these rules is more demanding and has to be performed either manually or using advanced CAD tools.

For this book, we have selected a "vanilla" 0.25 μm CMOS process as our preferred implementation medium. The rest of this section is devoted to a short introduction and overview of the design rules of this process, which fall in the micron-rules class. A complete design-rule set consists of the following entities: a set of layers, relations between objects on the same layer, and relations between objects on different layers. We discuss each of them in sequence.

### Layer Representation

The layer concept translates the intractable set of masks currently used in CMOS into a simple set of conceptual layout levels that are easier to visualize by the circuit designer. From a designer's viewpoint, all CMOS designs are based on the following entities:

- *Substrates* and/or *wells*, which are *p*-type (for NMOS devices) and *n*-type (for PMOS)
- *Diffusion regions* ($n^+$ and $p^+$), which define the areas where transistors can be formed. These regions are often called the *active areas*. Diffusions of an inverse type are needed to implement contacts to the wells or to the substrate. These are called *select regions*.
- One or more *polysilicon* layers, which are used to form the gate electrodes of the transistors (but serve as interconnect layers as well).
- A number of *metal interconnect* layers.
- *Contact and via* layers, which provide interlayer connections.

A layout consists of a combination of polygons, each of which is attached to a certain layer. The functionality of the circuit is determined by the choice of the layers, as well as the interplay between objects on different layers. For example, an MOS transistor is formed by the cross section of the diffusion layer and the polysilicon layer. An interconnection between two metal layers is formed by a cross section between the two metal layers and an additional contact layer. To visualize these relations, each layer is assigned a standard color (or stipple pattern for a black-and-white representation). The different layers used in our CMOS process are represented in Colorplate 1 (color insert).

### Intralayer Constraints

A first set of rules defines the minimum dimensions of objects on each layer, as well as the minimum spacings between objects on the same layer. All distances are expressed in μm. These constraints are presented in pictorial fashion in Colorplate 2.

[2] While not entirely accurate, lambda rules are still useful to estimate the impact of a technology scale on the area of a design.

### Interlayer Constraints

Interlayer rules tend to be more complex. Because multiple layers are involved, it is harder to visualize their meaning or functionality. Understanding layout requires the capability of translating the two-dimensional picture of the layout drawing into the three-dimensional reality of the actual device. This takes some practice.

We present these rules in a set of separate groupings:

1. *Transistor Rules* (Colorplate 3). A transistor is formed by the overlap of the active and the polysilicon layers. From the intralayer design rules, it is already clear that the minimum length of a transistor equals 0.24 μm (the minimum width of polysilicon), while its width is at least 0.3 μm (the minimum width of diffusion). Extra rules include the spacing between the active area and the well boundary, the gate overlap of the active area, and the active overlap of the gate.
2. *Contact and Via Rules* (Colorplates 2 and 4). A contact (which forms an interconnection between metal and active or polysilicon) or a via (which connects two metal layers) is formed by overlapping the two interconnecting layers and providing a contact hole, filled with metal, between the two. In our process, the minimum size of the contact hole is 0.3 μm, while the polysilicon and diffusion layers have to extend at least 0.14 μm beyond the area of the contact hole. This sets the minimum area of a contact to 0.44 μm × 0.44 μm. This is larger than the dimensions of a minimum-size transistor! Excessive changes between interconnect layers in routing should therefore be avoided. The figure, furthermore, points out the minimum spacings between contact and via holes, as well as their relationship with the surrounding layers.
3. *Well and Substrate Contacts* (Colorplate 5). For robust digital circuit design, it is important for the well and substrate regions to be adequately connected to the supply voltages. Failing to do so results in a resistive path between the substrate contact of the transistors and the supply rails, and can lead to possibly devastating parasitic effects, such as latchup. It is therefore advisable to provide numerous substrate (well) contacts spread over the complete region. To establish an ohmic contact between a supply rail, implemented in metal1, and a $p$-type material, a $p^+$ diffusion region must be provided. This is enabled by the *select* layer, which reverses the type of diffusion. A number of rules regarding the use of the *select layer* are illustrated in Colorplate 5.

Consider an $n$-well process, which implements the PMOS transistors into an $n$-type well diffused in a $p$-type material. The nominal diffusion is $p^+$. To invert the polarity of the diffusion, an $n$-select layer is provided that helps to establish the $n^+$ diffusions for the well contacts in the $n$-region, as well as the $n^+$ source and drain regions for the NMOS transistors in the substrate.

### Verifying the Layout

Ensuring that none of the design rules are violated is a fundamental requirement of the design process. Failing to do so will almost surely lead to a nonfunctional design. Doing so for a complex design that can contain millions of transistors is no simple task either, especially given the complexity of some design-rule sets. While design teams in the past used to spend numerous

hours staring at room-size layout plots, most of this work is now done by computers. Computer-aided *Design-Rule Checking* (called *DRC*) is an integral part of the design cycle for virtually every chip produced today. A number of layout tools even perform *on-line DRC* and check the design in the background during the time of conception.

**Example 2.1 Layout Example**

An example of a complete layout containing an inverter is shown in Figure 2-9. To help the visualization process, a vertical cross section of the process along the design center is included, as well as a circuit schematic.

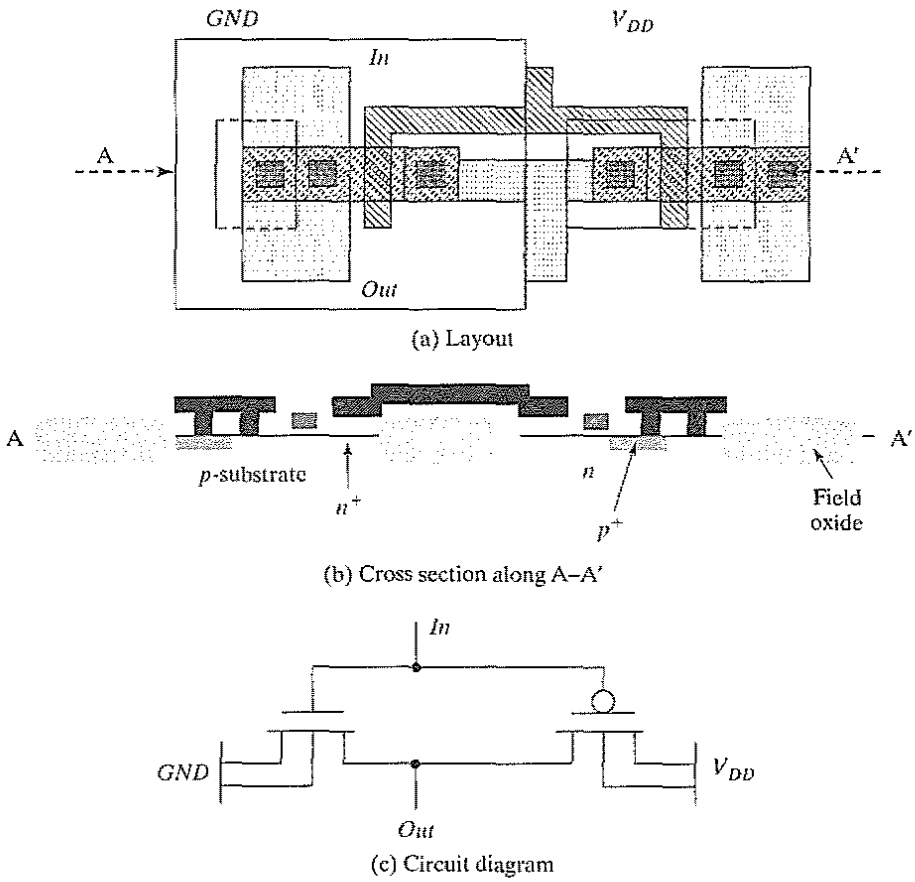



**Figure 2-9** A detailed layout example, including vertical process cross section and circuit diagram.

It is left as an exercise for the reader to determine the sizes of both the NMOS and the PMOS transistors.

## 2.4 Packaging Integrated Circuits

The IC package plays a fundamental role in the operation and performance of a component. Besides providing a means of bringing signal and supply wires in and out of the silicon die, it also removes the heat generated by the circuit and provides mechanical support. Finally, it also protects the die against environmental conditions such as humidity.

In addition, the packaging technology has a major impact on the performance and power dissipation of a microprocessor or signal processor. This influence is getting more pronounced as time progresses due to the reduction in internal signal delays and on-chip capacitance resulting from technology scaling. Currently, up to 50% of the delay of a high-performance computer is due to packaging delays, and this number is expected to rise. The search for higher performance packages with fewer inductive or capacitive parasitics has accelerated in recent years. The increasing complexity of what can be integrated on a single die also translates into a need for ever more input/output pins, as the number of connections going off-chip tends to be roughly proportional to the complexity of the circuitry on the chip. This relationship was first observed by E. Rent of IBM (published in [Landman71]), who translated it into an empirical formula that, appropriately, is called *Rent's rule*. This formula relates the number of input/output pins to the complexity of the circuit, as measured by the number of gates. It is written as

$$P = K \times G^{\beta} \tag{2.1}$$

where $K$ is the average number of I/Os per gate, $G$ the number of gates, $\beta$ the Rent exponent, and $P$ the number of I/O pins to the chip. $\beta$ varies between 0.1 and 0.7. Its value depends strongly upon the application area, architecture, and organization of the circuit, as demonstrated in Table 2-1. Clearly, microprocessors display a very different input/output behavior compared to memories.

The observed rate of pin-count increase for integrated circuits varies from 8% to 11% per year, and it has been projected that packages with more than 2000 pins will be required by 2010. For all these reasons, traditional dual-in-line, through-hole mounted packages have been replaced by other approaches, such as surface-mount, ball-grid array, and multichip module

**Table 2-1** Rent's constant for various classes of systems ([Bakoglu90])

| Application | $\beta$ | $K$ |
|---|---|---|
| Static memory | 0.12 | 6 |
| Microprocessor | 0.45 | 0.82 |
| Gate array | 0.5 | 1.9 |
| High-speed computer (chip) | 0.63 | 1.4 |
| High-speed computer (board) | 0.25 | 82 |

techniques. It is useful for the circuit designer to be aware of the available options and their pros and cons.

Due to its multifunctionality, a good package must comply with a large variety of requirements:

- **Electrical requirements**—Pins should exhibit low capacitance (both interwire and to the substrate), resistance, and inductance. A large characteristic impedance should be tuned to optimize transmission line behavior. Observe that intrinsic integrated-circuit impedances are high.
- **Mechanical and thermal properties**—The heat removal rate should be as high as possible. Mechanical reliability requires a good matching between the thermal properties of the die and the chip carrier. Long-term reliability requires a strong connection from die to package, as well as from package to board.
- **Low Cost**—Cost is one of the more important properties to consider in any project. For example, while ceramics have a superior performance over plastic packages, they are also substantially more expensive. Increasing the heat removal capacity of a package also tends to raise the package cost. The least expensive plastic packaging can dissipate up to 1 W. Slightly more expensive, but still of somewhat low quality, plastic packages can dissipate up to 2 W. Higher dissipation requires more expensive ceramic packaging. Chips dissipating over 20 W require special heat sink attachments. Even more extreme techniques such as fans and blowers, liquid cooling hardware, or heat pipes are needed for higher dissipation levels.

  Packing density is a major factor in reducing board cost. The increasing pin count either requires an increase in the package size or a reduction in the pitch between the pins. Both have a profound effect on the packaging economics.

Packages can be classified in many different ways: by their main material, the number of interconnection levels, and the means used to remove heat. In this brief section, we provide only sketches of each of those issues.

### 2.4.1 Package Materials

The most common materials used for the package body are ceramic and polymers (plastics). The latter have the advantage of being substantially cheaper, but they suffer from inferior thermal properties. For example, the ceramic $Al_2O_3$ (alumina) conducts heat better than $SiO_2$ and the polyimide plastic by factors of 30 and 100, respectively. Furthermore, its thermal expansion coefficient is substantially closer to the typical interconnect metals. The disadvantage of alumina and other ceramics is their high dielectric constant, which results in large interconnect capacitances.

### 2.4.2 Interconnect Levels

The traditional packaging approach uses a two-level interconnection strategy. The die is first attached to an individual chip carrier or substrate. The package body contains an internal cavity where the chip is mounted. These cavities provide ample room for many connections to the chip leads (or pins). The leads compose the second interconnect level and connect the chip to the global interconnect medium, which normally is a PC board. Complex systems contain even more interconnect levels, since boards are connected together using backplanes or ribbon cables. The first two layers of the interconnect hierarchy are illustrated in the drawing of Figure 2-10. The sections that follow provide a brief overview of the interconnect techniques used at levels one and two of the interconnect hierarchy, and a brief discussion of some more advanced packaging approaches.

**Interconnect Level 1—Die-to-Package Substrate**

For a long time, *wire bonding* was the technique of choice to provide an electrical connection between die and package. In this approach, the backside of the die is attached to the substrate using glue with a good thermal conductance. Next, the chip pads are individually connected to the lead frame with aluminum or gold wires. The wire-bonding machine used for this purpose operates much like a sewing machine. An example of wire bonding is shown in Figure 2-11. Although the wire-bonding process is automated to a large degree, it has some major disadvantages:

1. Wires must be attached serially, one after the other. This leads to longer manufacturing times with increasing pin counts.
2. Larger pin counts make it substantially more challenging to find bonding patterns that avoid shorts between the wires.
3. The exact value of the parasitics is hard to predict because of the manufacturing approach and irregular outlay.

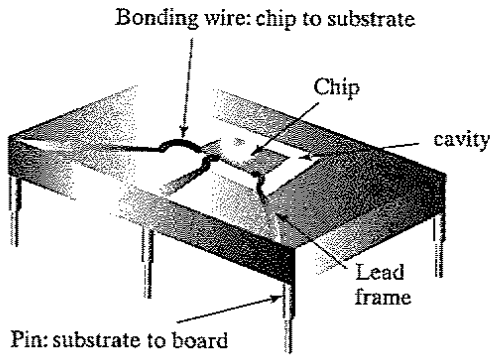



**Figure 2-10** Interconnect hierarchy in traditional IC packaging.

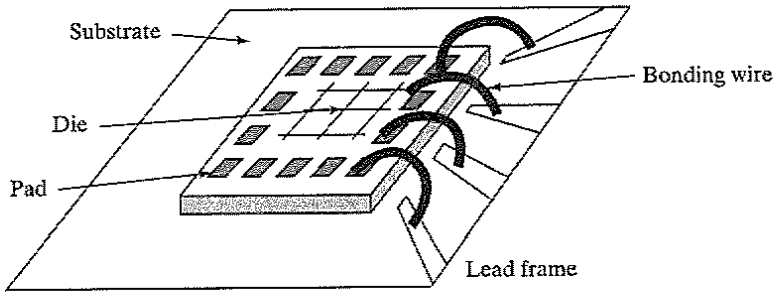



**Figure 2-11** Wire bonding.

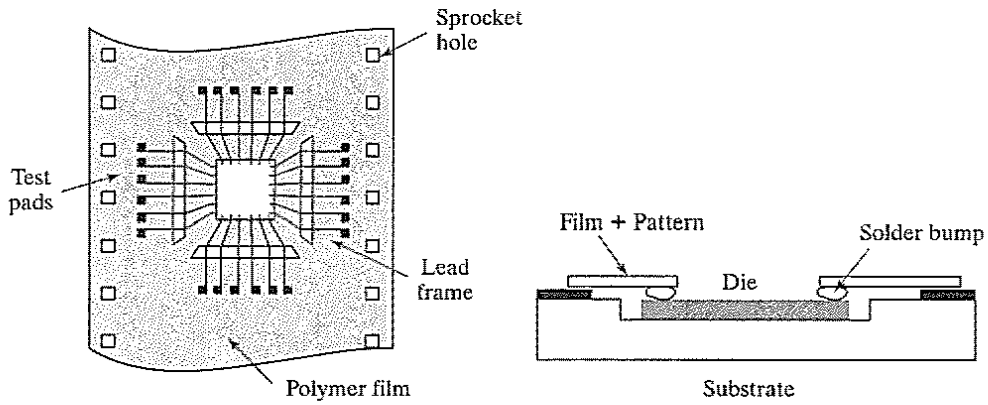
Bonding wires have inferior electrical properties, such as a high individual inductance (5 nH or more) and mutual inductance with neighboring signals. The inductance of a bonding wire is typically about 1 nH/mm, while the inductance per package pin ranges between 7 and 40 nH per pin, depending on the type of package as well as the positioning of the pin on the package boundary [Steidel83]. Typical values of the parasitic inductances and capacitances for a number of commonly used packages are summarized in Table 2-2.

New attachment techniques are being explored as a result of these deficiencies. In one approach, called *Tape Automated Bonding* (or TAB), the die is attached to a metal lead frame that is printed on a polymer film, typically polyimide (see Figure 2-12a). The connection between chip pads and polymer film wires is made using solder bumps (Figure 2-12b). The tape can then be connected to the package body using a number of techniques. One possible approach is to use pressure connectors.

The advantage of the TAB process is that it is highly automated. The sprockets in the film are used for automatic transport. All connections are made simultaneously. The printed approach helps to reduce the wiring pitch, which results in higher lead counts. Elimination of the long bonding wires improves the electrical performance. For instance, for a two-conductor layer, 48 mm

**Table 2-2** Typical capacitance and inductance values of package and bonding styles (from [Steidel83], [Franzon93], and [Harper00]).

| Package Type | Capacitance (pF) | Inductance (nH) |
|---|---|---|
| 68-pin plastic DIP | 4 | 35 |
| 68-pin ceramic DIP | 7 | 20 |
| 300 pin Ball Grid Array | 1–5 | 2–15 |
| Wire bond | 0.5–1 | 1–2 |
| Solder bump | 0.1–0.5 | 0.01–0.1 |




(a) Polymer tape with imprinted wiring pattern (b) Die attachment using solder bumps

**Figure 2-12** Tape-automated bonding (TAB).

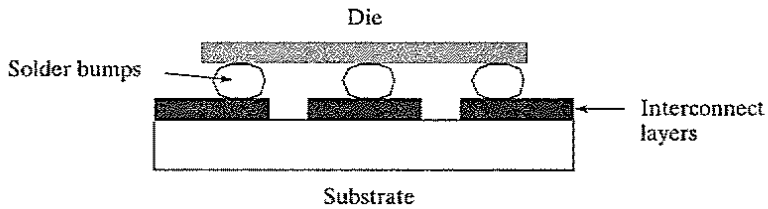


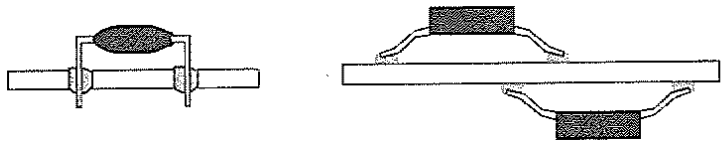

**Figure 2-13** Flip-chip bonding.

TAB Circuit, the following electrical parameters hold: $L \approx 0.3$–$0.5$ nH, $C \approx 0.2$–$0.3$ pF, and $R \approx 50$–$200\ \Omega$ [Doane93, p. 420].

Another approach is to flip the die upside down and attach it directly to the substrate using solder bumps. This technique, called *flip-chip* mounting, has the advantage of a superior electrical performance (see Figure 2-13.) Instead of making all the I/O connections on the die boundary, pads can be placed at any position on the chip. This can help address the power- and clock-distribution problems, since the interconnect materials on the substrate (e.g., Cu or Au) typically are of better quality than the Al on the chip.

### Interconnect Level 2—Package Substrate to Board

When connecting the package to the PC board, *through-hole mounting* has been the packaging style of choice. A PC board is manufactured by stacking layers of copper and insulating epoxy glass. In the through-hole mounting approach, holes are drilled through the board and plated with copper. The package pins are inserted and electrical connection is made with solder (see Figure 2-14a). The favored package in this class was the *dual-in-line* package or DIP, as in Figure 2-15-2. The packaging density of the DIP degrades rapidly when the number of pins exceeds 64. This problem can be alleviated by using the *pin-grid-array* (PGA) package that has



(a) Through-hole mounting (b) Surface mount

**Figure 2-14** Board-mounting approaches.

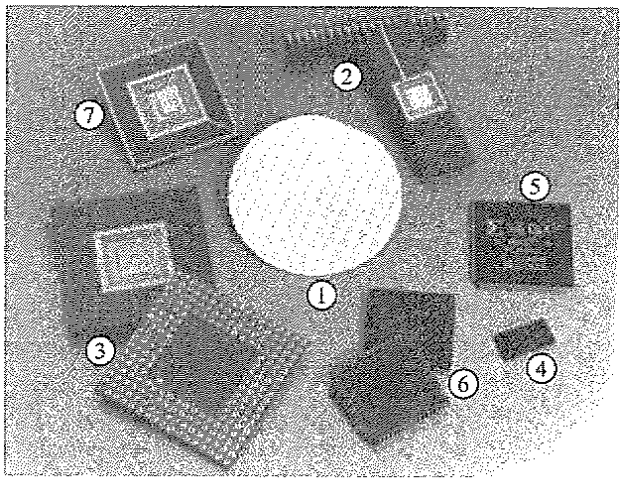
leads on the entire bottom surface instead of only on the periphery (Figure 2-15-3). PGAs can extend to large pin counts (over 400 pins are possible).

The through-hole mounting approach offers a mechanically reliable and sturdy connection. However, this comes at the expense of packaging density. For mechanical reasons, a minimum pitch of 2.54 mm between the through holes is required. Even under those circumstances, PGAs with large numbers of pins tend to substantially weaken the board. In addition, through holes limit the board packing density by blocking lines that might otherwise have been routed below them, which results in longer interconnections. PGAs with large pin counts therefore require extra routing layers to connect to the multitudes of pins. Finally, while the parasitic capacitance and inductance of the PGA are slightly lower than that of the DIP, their values are still substantial.

Many of the shortcomings of the through-hole mounting approach are solved by using the *surface-mount* technique. A chip is attached to the surface of the board with a solder connection without requiring any through holes (Figure 2-14b). Packing density is increased for the following reasons: (1) through holes are eliminated, which provides more wiring space; (2) the lead pitch is reduced; and (3) chips can be mounted on both sides of the board. In addition, the elimination of the through holes improves the mechanical strength of the board. On the negative side, the on-the-surface connection makes the chip-board connection weaker. Not only is it cumbersome to mount a component on a board, but also more expensive equipment is needed, since a simple soldering iron will no longer suffice. Finally, testing of the board is more complex, because the package pins are no longer accessible at the backside of the board. Signal probing becomes difficult or almost impossible.

A variety of surface-mount packages are currently in use with different pitch and pin-count parameters. Three of these packages are shown in Figure 2-15: the *small-outline package* with gull wings, the *plastic leaded package* (PLCC) with J-shaped leads, and the *leadless chip carrier*. An overview of the most important parameters for a number of packages is given in Table 2-3.

Even surface-mount packaging is unable to satisfy the quest for ever higher pin counts. This is worsened by the demand for power connections: today's high performance chips, operating at low supply voltages, require as many power and ground pins as signal I/Os! When more than 300 I/O connections are needed, solder balls replace pins as the preferred interconnect medium between package and board. An example of such a packaging approach, called ceramic




1 Bare die
2 DIP
3 PGA
4 Small-outline IC
5 Quad flat pack
6 PLCC
7 Leadless carrier

**Figure 2-15** An overview of commonly used package types.

**Table 2-3** Parameters of various types of chip carriers.

| Package Type | Lead Spacing (Typical) | Lead Count (Maximum) |
|---|---|---|
| Dual in line | 2.54 mm | 64 |
| Pin grid array | 2.54 mm | > 300 |
| Small-outline IC | 1.27 mm | 28 |
| Leaded chip carrier (PLCC) | 1.27 mm | 124 |
| Leadless chip carrier | 0.75 mm | 124 |

*ball grid array* (BGA), is shown in Figure 2-16. Solder bumps are used to connect both the die to the package substrate, and the package to the board. The area array interconnect of the BGA provides constant input/output density regardless of the number of total package I/O pins. A minimum pitch between solder balls of as low as 0.8 mm can be obtained, and packages with multiple thousands of I/O signals are feasible.

### Multichip Modules—Die-to-Board

The deep hierarchy of interconnect levels in the package is becoming unacceptable in today's complex designs due to their higher levels of integration, large signal counts, and increased performance requirements. The trend, therefore, is toward reducing the number of levels. For the time being, attention is focused on the elimination of the first level in the packaging hierarchy. Removing one layer in the packaging hierarchy by mounting the die directly on the

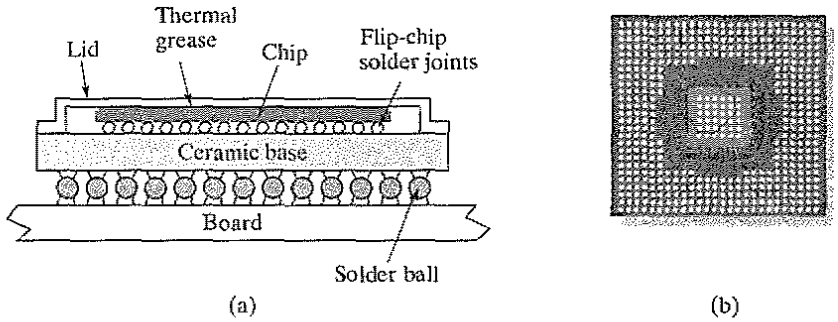



**Figure 2-16** Ball grid array packaging; (a) cross section, (b) photo of package bottom.

wiring backplanes—board or substrate—offers a substantial benefit when performance or density is a major issue. This packaging approach is called the multichip module technique (or MCM), and results in a substantial increase in packing density, as well as improved performance overall.

A number of the previously mentioned die-mounting techniques can be adapted to mount dies directly on the substrate, including wire bonding, TAB, and flip-chip, although the latter two are preferable. The substrate itself can vary over a wide range of materials, depending upon the required mechanical, electrical, thermal, and economical requirements. Materials of choice are epoxy substrates (similar to PC boards), metal, ceramics, and silicon. Silicon has the advantage of presenting a perfect match in mechanical and thermal properties with respect to the die material.

The main advantages of the MCM approach are the increased packaging density and performance. An example of an MCM module implemented using a silicon substrate (commonly dubbed *silicon on silicon*) is shown in Figure 2-17. The module, which implements an avionics processor module and is fabricated by Rockwell International, contains 53 ICs and 40 discrete devices on a 2.2″ × 2.2″ substrate with aluminum polyimide interconnect. The interconnect wires are only an order of magnitude wider than what is typical for on-chip wires, since similar patterning approaches are used. The module itself has 180 I/O pins. Performance is improved by the elimination of the chip-carrier layer with its assorted parasitics, and through a reduction of the global wiring lengths on the die, a result of the increased packaging density. For instance, a solder bump has an assorted capacitance and inductance of only 0.1 pF and 0.01 nH, respectively. The MCM technology can also reduce power consumption significantly, since large output drivers—and associated dissipation—become superfluous due to the reduced load capacitance of the output pads. The dynamic power associated with the switching of the large load capacitances is simultaneously reduced.

While MCM technology offers some clear benefits, its main disadvantage is economic. This technology requires some advanced manufacturing steps that make the process expensive. The approach was until recently only justifiable when either dense housing or extreme performance is essential. In recent years, the economics have been shifting, and advanced multichip

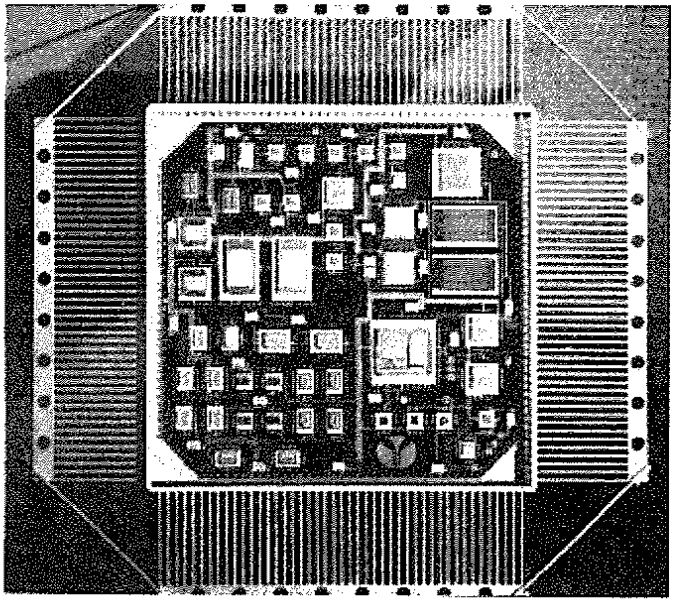


**Figure 2-17** Avionics processor module. *Courtesy of Rockwell Collins, Inc.*

packaging approaches have made inroad in several low-cost high-density applications as well. This trend is called the *system-in-a-package* (SIP) strategy.

### 2.4.3 Thermal Considerations in Packaging

As the power consumption of integrated circuits rises, it becomes increasingly important to efficiently remove the heat generated by the chips. A large number of failure mechanisms in ICs are accentuated by increased temperatures. Examples are leakage in reverse-biased diodes, electromigration, and hot-electron trapping. To prevent failure, the temperature of the die must be kept within certain ranges. The supported temperature range for commercial devices during operation equals 0° to 70°C. Military parts are more demanding and require a temperature range varying from –55° to 125°C.

The cooling effectiveness of a package depends on the thermal conduction (resistance) of the package material, which consists of the package substrate and body, the package composition, and the effectiveness of the heat transfer between package and cooling medium. Standard packaging approaches use still or circulating air as the cooling medium. The transfer efficiency can be improved by adding finned metal heat sinks to the package. More expensive packaging approaches, such as those used in mainframes or supercomputers, force air, liquids, or inert gases through tiny ducts in the package to achieve even greater cooling efficiencies.

Given the thermal resistance $\theta$ of the package, expressed in °C/W, we can derive the chip temperature by using the *heat flow equation*

$$\Delta T = T_{chip} - T_{env} = \theta Q, \tag{2.2}$$

with $T_{chip}$ and $T_{env}$ the chip and environment temperatures, respectively. $Q$ represents the heat flow (in Watt). Observe how closely the heat flow equation resembles Ohm's Law. The heat flow and temperature differential are the equivalents of current and voltage difference, respectively. Thermal modeling of a chip, its package, and its environment is a complex task. We refer the reader to [Lau98 – Chapter 3] for a more detailed discussion on the topic.

**Example 2.2 Thermal Conduction of Package**

As an example, a 40-pin DIP has a thermal resistance of 38°C/W and 25°C/W for natural and forced convection of air. This means that a DIP can dissipate 2 watts (3 watts) of power with natural (forced) air convection, and still keep the temperature difference between the die and the environment below 75°C. For comparison, the thermal resistance of a ceramic PGA ranges from 15° to 30°C/W.

Since packaging approaches with decreased thermal resistance are prohibitively expensive, keeping the power dissipation of an integrated circuit within bounds is an economic necessity. The increasing integration levels and circuit performance make this task nontrivial. An interesting relationship in this context has been derived by Nagata [Nagata92]. It provides a bound on the integration complexity and performance as a function of the thermal parameters. We write

$$\frac{N_G}{t_p} \leq \frac{\Delta T}{\theta E} \tag{2.3}$$

where $N_G$ is the number of gates on the chip, $t_p$ the propagation delay, $\Delta T$ the maximum temperature difference between chip and environment, $\theta$ the thermal resistance between them, and $E$ the switching energy of each gate.

**Example 2.3 Thermal Bounds on Integration**

For $\Delta T = 100°C$, $\theta = 2.5°C/W$ and $E = 0.1$ pJ, this results in $N_G/t_p \leq 4 \times 10^5$ (gates/nsec). In other words, the maximum number of gates on a chip, when all gates are operating simultaneously, must be less than 400,000 if the switching speed of each gate is 1 nsec. This is equivalent to a power dissipation of 40 W.

Fortunately, not all gates are operating simultaneously in real systems. The maximum number of gates can be substantially larger, based on the activity in the circuit. For example, it has been experimentally derived that the ratio between the average switching period and the propagation delay ranges from 20 to 200 in mini- and large-scale computers [Masaki92].

Nevertheless, Eq. (2.3) demonstrates that heat dissipation and thermal concerns present an important limitation on circuit integration. Design approaches for low power that reduce either $E$ or the activity factor are rapidly gaining importance.

## 2.5 Perspective—Trends in Process Technology

Modern CMOS processes pretty much track the flow described in the previous sections, although a number of the steps might be reversed, a single well approach might be followed, a grown field oxide instead of the trench approach might be used, or extra steps such as LDD (*Lightly Doped Drain*) might be introduced. Also, it is quite common to cover the polysilicon interconnections as well as the drain and source regions with a *silicide* such as $TiSi_2$ to improve the conductivity (see Figure 2-2). This extra operation is inserted between steps $i$ and $j$ of our process. Some important modifications or improvements to the technology are currently under way or are on the horizon, and deserve some attention. Beyond these, we expect no dramatic changes from the described CMOS technology in the next decade.

### 2.5.1 Short-Term Developments

**Copper and Low-k Dielectrics**

A recurring theme throughout this book will be the increasing impact of interconnect on the overall design performance. Process engineers are continuously evaluating alternative options for the traditional Aluminum-conductor–$SiO_2$-insulator combination that has been the norm for the last several decades. In 1998, engineers at IBM introduced an approach that finally made the use of copper as an interconnect material in a CMOS process viable and economical [Geppert98]. Copper has a resistivity that is substantially lower than aluminum. It has the disadvantage of easy diffusion into silicon, which degrades the characteristics of the devices. Coating the copper with a buffer material such as titanium-nitride, preventing the diffusion, addresses this problem, but requires a special deposition process. The Dual Damascene process, introduced by IBM, (Figure 2-18) uses a metallization approach that fills trenches etched into the insulator, followed by a chemical–mechanical polishing step. This is in contrast with the traditional approach that first deposits a full metal layer, and removes the redundant material through etching.

In addition to the lower resistivity interconnections, insulator materials with a lower dielectric constant than $SiO_2$, and hence, lower capacitance have also found their way into the production process, starting with the 0.18-μm CMOS process generation.

**Silicon on Insulator**

Although it has been around a long time, there seems to be a good chance that Silicon-on-Insulator (SOI) CMOS might replace the traditional CMOS process, described in the previous sections (also known as the *bulk CMOS process*). The main difference lies in the start material: the SOI transistors are constructed in a very thin layer of silicon, deposited on top of a thick layer of insulating $SiO_2$ (see Figure 2-19). The primary advantages of the SOI process are reduced

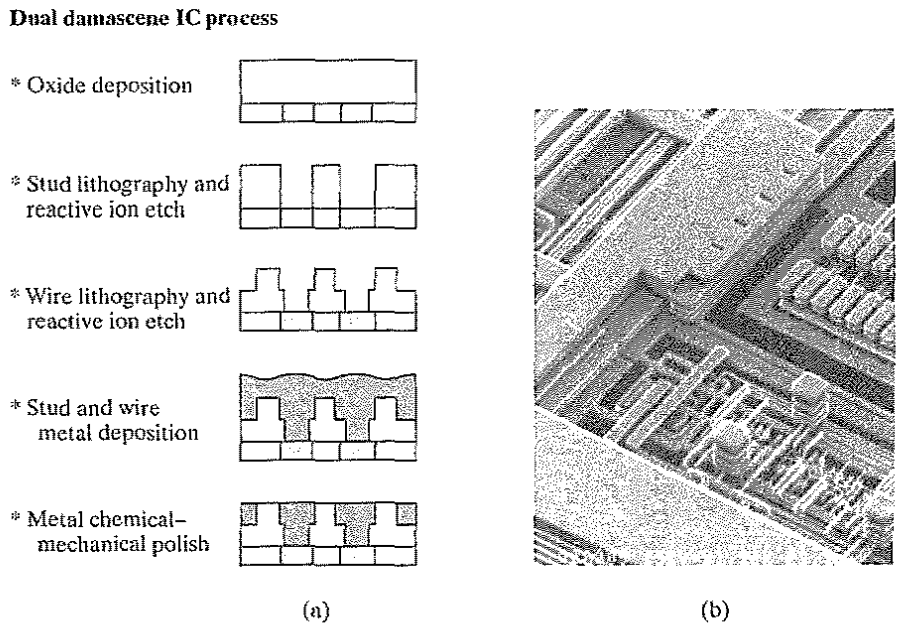



**Figure 2-18** The damascene process (from [Geppert98]): process steps (a), and microphotograph of interconnect after removal of insulator (b).

parasitics and better transistor on–off characteristics. It has, for example, been demonstrated by researchers at IBM, that the porting of a design from a bulk CMOS to an SOI process—leaving all other design and process parameters such as channel length and oxide thickness identical—yields a performance improvement of 22% [Allen99]. Preparing a high quality SOI substrate at an economical cost was long the main hindrance against a large-scale introduction of the process. This picture had changed by the end of the 1990s, and SOI is steadily moving into the mainstream.

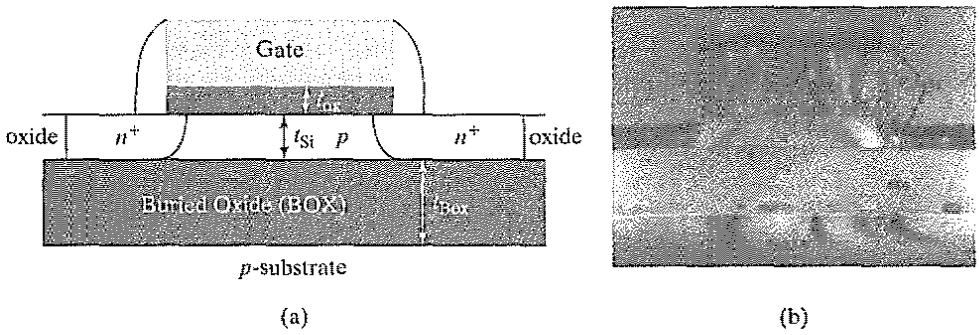



**Figure 2-19** Silicon-on-insulator process—schematic diagram (a) and SEM cross section (b). [Eaglesham99].

### 2.5.2 In the Longer Term

Extending the life of CMOS technology beyond the next decade, and going deeply below the 100 nm channel length region, however, will require redeveloping the process technology and the device structure. Already we are witnessing the emergence of a wide range of new devices (such as organic transistors, molecular switches, and quantum devices). While we cannot project what approaches will dominate in the next era, one interesting development is worth mentioning.

**Truly Three-Dimensional Integrated Circuits**

Getting signals in and out of the computation elements in a timely fashion is one of the main challenges presented by the continued increase in integration density. One way to address this problem is to introduce extra active layers, and to sandwich them between the metal interconnect layers, as shown in Figure 2-20. This enables us to position high density memory on top of the logic processors implemented in the bulk CMOS, reducing the distance between computation and storage, and thus also the delay [Souri00]. In addition, devices with different voltage, performance, or substrate material requirements can be placed in different layers. For instance, the top active layer can be reserved for the realization of optical transceivers, which may help to address the input/output requirements, or MEMS (Micro Electro-Mechanical Systems) devices providing sensoring functions or radio frequency (RF) interfaces.

While this approach may seem to be promising, a number of major challenges and hindrances have to be resolved to make it truly viable. How to remove the dissipated heat is one of the more compelling questions, ensuring yield is another. Researchers are demonstrating major progress on these issues, and 3D integration might well be on the horizon. Before the true solution arrives, we might have to rely on some intermediate approaches. One alternative, called 2.5D *integration*, is to bond two fully processed wafers, on which circuits are fabricated on the

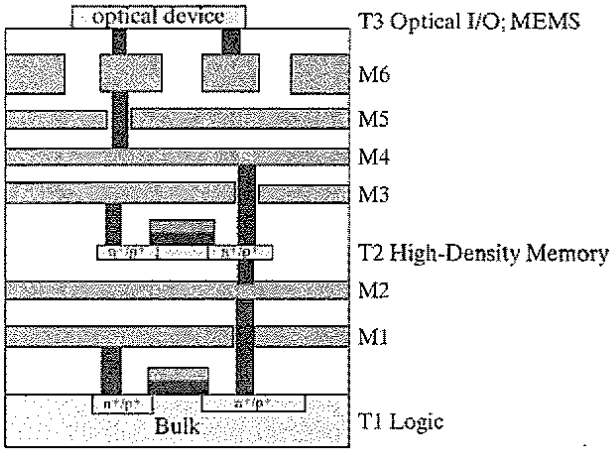



**Figure 2-20** Example of true 3D integration. Extra active layers (T*), implementing high-density memory and I/O, are sandwiched between the metal interconnect layers (M*).

surface such that the chips completely overlap. Vias are etched to electrically connect both chips after metallization. The advantages of this technology lie in the similar electrical properties of devices on all active levels and the independence of processing temperature since all chips can be fabricated separately and later bonded. The major limitation of this technique is its lack of precision (best case alignment: +/– 2 μm), which restricts the interchip communication to global metal lines.

One picture that strongly emerges from these futuristic devices is that the line between chip, substrate, package, and board is blurring. Designers of these *systems on a die* or *systems in a package* will have to consider all these aspects simultaneously.

## 2.6 Summary

This chapter has presented a bird's-eye view of the manufacturing and packaging process of CMOS integrated circuits:

- The manufacturing process of integrated circuits requires many steps, each of which consists of a sequence of basic operations. A number of these steps and/or operations, such as photolithograpical exposure and development, material deposition, and etching, are executed very repetitively in the course of the manufacturing process.
- The *optical masks* forms the central interface between the intrinsics of the manufacturing process and the design that the user wants to see transferred to the silicon fabric.
- The *design-rules set* defines the constraints in terms of minimum width and separation that the IC design has to adhere to if the resulting circuit is to be fully functional. These design rules act as the contract between the circuit designer and the process engineer.
- The *package* forms the interface between the circuit implemented on the silicon die and the outside world, and as such has a major impact on the performance, reliability, longevity, and cost of the integrated circuit.

## 2.7 To Probe Further

Many books on semiconductor manufacturing have been published in the last few decades. An excellent overview of the state of the-art in CMOS manufacturing is *Silicon VLSI Technology* by J. Plummer, M. Deal, and P. Griffin [Plummer00]. A visual overview of the different steps in the manufacturing process can be found on the Web at [Fullman99]. Other sources for information are the IEEE Transactions on Electron Devices, and the Technical Digest of the IEDM conference. A number of great compendia are available for up-to-date and in-depth information about electronic packaging. [Doane93], [Harper00], and [Lau98] are good examples of such.

## References

[Allen99] D. Allen, et al., "A 0.2 μm 1.8 V SOI 550 MHz PowerPC Microprocessor with Copper Interconnects," *Proceedings IEEE ISSCC Conference*, vol. XLII, pp. 438–439, February 1999.

[Bakoglu90] H. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, Addison-Wesley, 1990.

[Doane93] D. Doane, ed., *Multichip Module Technologies and Alternatives*, Van Nostrand-Reinhold, 1993.

[Englesham 99] D. Englesham, "0.18 μm CMOS and Beyond," *Proceedings 1999 Design Automation Conference*, pp. 703–708, June 1999.

[Franzon93] P. Franzon, "Electrical Design of Digital Multichip Modules," in [Doane93], pp 525–568, 1993.

[Fullman99] Fullman Kinetics, "The Semiconductor Manufacturing Process," *http://www.fullman-kinetics.com/semiconductors/semiconductors.html*, 1999.

[Geppert98] L. Geppert, "Technology—1998 Analysis and Forecast," *IEEE Spectrum*, vol. 35, no. 1, p. 23, January 1998.

[Harper00] C. Harper, Ed., *Electronic Packaging and Interconnection Handbook*, McGraw-Hill, 2000.

[Landman71] B. Landman and R. Russo, "On a Pin versus Block Relationship for Partitions of Logic Graphs," *IEEE Trans. on Computers*, vol. C-20, pp. 1469–1479, December 1971.

[Lau98] J. Lau et al., *Electronic Packaging—Design, Materials, Process, and Reliability*, McGraw-Hill, 1998.

[Masaki92] A. Masaki, "Deep-Submicron CMOS Warms Up to High-Speed Logic," *Circuits and Devices Magazine*, November 1992.

[Mead80] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

[Nagata92] M. Nagata, "Limitations, Innovations, and Challenges of Circuits and Devices into a Half Micrometer and Beyond," *IEEE Journal of Solid State Circuits*, vol. 27, no. 4, pp. 465–472, April 1992.

[Plummer00] J. Plummer, M. Deal, and P. Griffin, *Silicon VLSI Technology*, Prentice Hall, 2000.

[Steidel83] C. Steidel, *Assembly Techniques and Packaging*, in [Sze83], pp. 551–598, 1983.

[Souri00] S. J. Souri, K. Banerjee, A. Mehrotra and K. C. Saraswat, "*Multiple Si Layer ICs: Motivation, Performance Analysis, and Design Implications*," Proceedings 37th Design Automation Conference, pp. 213–220, June 2000.

DESIGN METHODOLOGY INSERT

# A

# IC LAYOUT

*Creating a manufacturable layout*

*Verifying the layout*

The increasing complexity of the integrated circuit has made the role of design-automation tools indispensable, and raises the abstractions the designer is working with to ever higher levels. Yet, when performance or design density is of primary importance, the designer has no other choice than to return to handcrafting the circuit topology and physical design. The labor-intensive nature of this approach, called *custom design*, translates into a high cost and a long time to market. Therefore, it can only be justified economically under the following conditions:

- The custom block can be reused many times, as a library cell, for instance.
- The cost can be amortized over a large volume. Microprocessors and semiconductor memories are examples of applications in this class.
- Cost is not among the prime design criteria.[1] Examples include space applications and scientific instrumentation.

With continuous progress in the design-automation arena, the share of custom design reduces from year to year. Even in high-performance microprocessors, large portions are

[1] This is becoming increasingly rare.



Dell Ex. 1025
Page 62

designed automatically using semicustom design approaches. Only the most performance-critical modules—such as the integer and floating-point execution units—are handcrafted.

Although the amount of design automation in the custom design process is minimal, some design tools have proven to be indispensable. Together with circuit simulators, these programs form the core of every design-automation environment, and they are the first tools an aspiring circuit designer will encounter.

### Layout Editor

The layout editor is the premier working tool of the designer and exists primarily for the generation of a physical representation of a design, given a circuit topology. Virtually every design-automation vendor offers an entry in this field. The most well known is the MAGIC tool developed at the University of California at Berkeley [Ousterhout84], which has been widely distributed. Even though MAGIC did not withstand the evolution of software technology and user interface, some of its offspring did. Throughout this book, we will be using a layout tool called **max**, a MAGIC descendant developed by a company called MicroMagic [mmi00]. A typical **max** display is shown in Figure A-1 and illustrates the basic function of the layout editor—placing polygons on

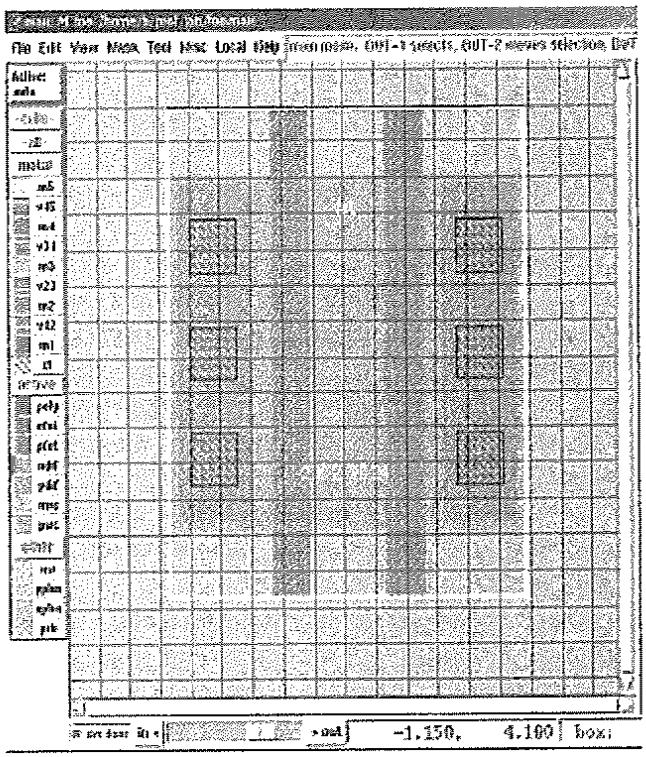



**Figure A-1** View of a **max** display window. It plots the layout of two stacked NMOS transistor. The menu on the left side allows for the selection of the layer a particular poligon will be placed on.

different mask layers so that a functional physical design is obtained (scathingly called *polygon pushing*).

### Symbolic Layout

Since physical design occupies a major fraction of the design time for a new cell or component, techniques to expedite this process have been in continual demand. The *symbolic-layout* approach has gained popularity over the years. In this design methodology, the designer only draws a shorthand notation for the layout structure. This notation indicates only the *relative* positioning of the various design components (transistors, contacts, wires). The *absolute* coordinates of these elements are determined automatically by the editor using a *compactor* [Hsueh79, Weste93]. The compactor translates the design rules into a set of constraints on the component positions, and solves a constrained optimization problem that attempts to minimize the area or another cost function.

An example of a symbolic notation for a circuit topology called a *sticks diagram* is shown in Figure A-2. The different layout entities are dimensionless, since only positioning is important. The advantage of this approach is that the designer does not have to worry about design rules, because the compactor ensures that the final layout is physically correct. Thus, she can avoid cumbersome polygon manipulations. Another plus of the symbolic approach is that cells can adjust themselves automatically to the environment. For example, automatic pitch matching of cells is an attractive feature in module generators. Consider the case of Figure A-3 (from [Croes88]), in which the original cells have different heights, and the terminal positions do not match. Connecting the cells would require extra wiring. The symbolic approach allows the cells to adjust themselves and connect without any overhead.

The disadvantage of the symbolic approach is that the outcome of the compaction phase often is unpredictable. The resulting layout can be less dense than what is obtained with the manual approach. This has prevented it from becoming a mainstream layout tool. Nonetheless, symbolic layout techniques have improved considerably over the years, and they have become very useful as a first-order drafting tool for new cells. More important, they form the solid underpining of the automatic cell-generation techniques, described later, in Chapter 8.

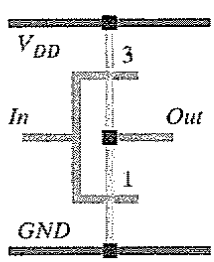



**Figure A-2** Sticks representation of CMOS inverter. The numbers represent the (*Width/Length*)-ratios of the transistors.

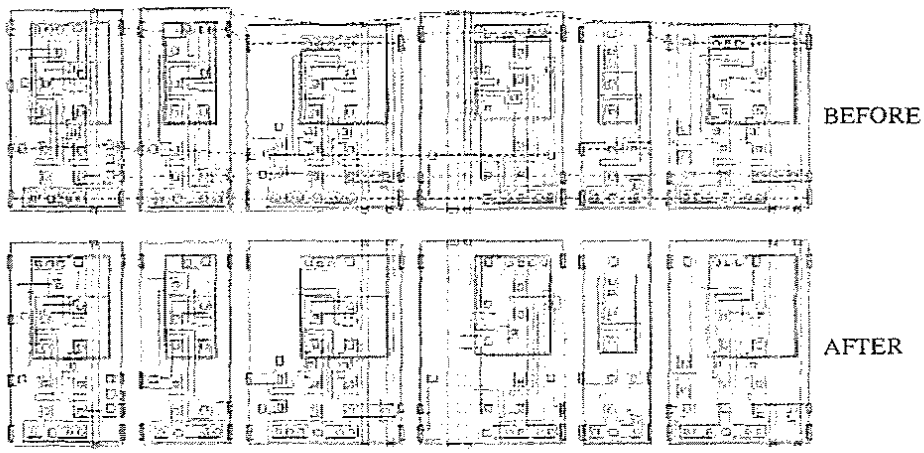



**Figure A-3** Automatic pitch matching of data path cells based on symbolic layout.

### Design-Rule Checking

Design rules were introduced in Chapter 2 as a set of layout restrictions that ensure the manufactured design will operate as desired with no short or open circuits. A prime requirement of the physical layout of a design is that it adhere to these rules. This can be verified with the aid of a *design-rule checker (DRC)*, which uses as inputs the physical layout of a design and a description of the design rules presented in the form of a *technology file*. Since a complex circuit can contain millions of polygons that must be checked against each other, efficiency is the most important property of a good DRC tool. The verification of a large chip can take hours or days of computation time. One way of expediting the process is to preserve the design hierarchy at the physical level. For example, if a cell is used multiple times in a design, it should be checked only once. Besides speeding up the process, the use of hierarchy can make error messages more informative by retaining knowledge of the circuit structure.

DRC tools come in two formats: (1) The *on-line DRC* runs concurrent with the layout editor and flags design violations during the cell layout. For instance, **max** has a built-in design-rule checking facility. An example of on-line DRC is shown in Figure A-4. (2) *Batch DRC* is used as a postdesign verifier; it is run on a complete chip prior to shipping the mask descriptions to the manufacturer.

### Circuit Extraction

Another important tool in the custom-design methodology is the circuit extractor, which derives a circuit schematic from a physical layout. By scanning the various layers and their interactions, the extractor reconstructs the transistor network, including the sizes of the devices and the interconnections. The schematic produced can be used to verify that the artwork implements the intended function. Furthermore, the resulting circuit diagram contains precise information on the parasitics, such as the diffusion and wiring capacitances and resistances. This allows for a

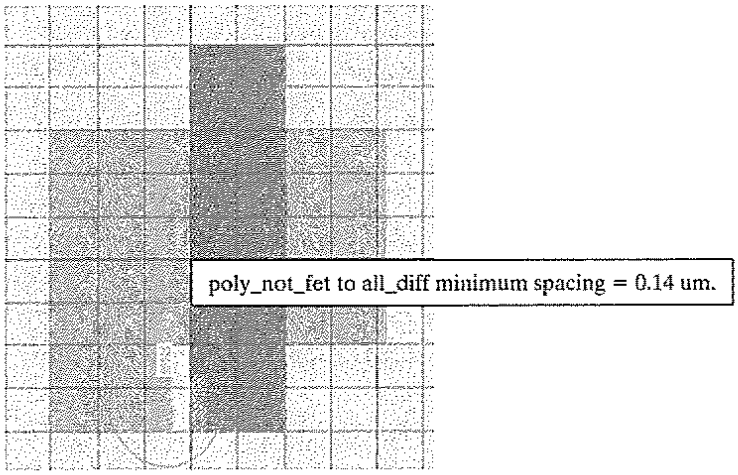



**Figure A-4** On-line design rule checking. The white dots indicate a design rule violation. The violated rule can be obtained with a simple mouse click.

more accurate simulation and analysis. The complexity of the extraction depends greatly upon the desired information. Most extractors extract the transistor network and the capacitances of the interconnect with respect to *GND* or other network nodes. Extraction of the wiring resistances already comes at a greater cost, yet it has become a necessity for virtually all high-performance circuits. Clever algorithms have helped to reduce the complexity of the resulting circuit diagrams. For very high-speed circuits, extraction of the inductance would be desirable as well. Unfortunately, this requires a three-dimensional analysis and is only feasible for small-sized circuits at present.

## A.1 To Probe Further

More detailed information regarding the MAGIC and **max** layout editors can be found on the web site of this book. In-depth textbooks on layout generation and verification have been published, and can be of great help to the novice designer. To mention a few of them, [Clein00], [Uyemura95], and [Wolf94] offer some comprehensive and well-illustrated treatment and discussion.

## References

[Clein00] D. Clein, *CMOS IC Layout—Concepts, Methodologies, and Tools*, Newnes, 2000.

[Croes88] K. Croes, H. De Man, and P. Six, "CAMELEON: A Process-Tolerant Symbolic Layout System," *Journal of Solid State Circuits*, vol. 23 no. 3, pp. 705–713, June 1988.

[Hsueh79] M. Hsueh and D. Pederson, "Computer-Aided Layout of LSI Building Blocks," *Proceedings ISCAS Conf.*, pp. 474–477, Tokyo, 1979.

[mmi00] MicroMagic, Inc, *http://www.micromagic.com*.

[Ousterhout84] J. Ousterhout, G. Hamachi, R. Mayo, W. Scott, and G. Taylor, "Magic: A VLSI Layout System," *Proc. 21st Design Automation Conference*, pp. 152–159, 1984.

[Uyemura95] J. Uyemura, *Physical Design of CMOS Integrated Circuits Using L-EDIT*, PWS Publishing Company, 1995.

[Weste93] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design—A Systems Perspective*, Addison-Wesley, 1993.

[Wolf94] W. Wolf, *Modern VLSI Design—A Systems Approach*, Prentice Hall, 1994.

CHAPTER

5

# The CMOS Inverter

*Quantification of integrity, performance, and energy metrics of an inverter*
*Optimization of an inverter design*


5.1 Introduction
5.2 The Static CMOS Inverter—An Intuitive Perspective
5.3 Evaluating the Robustness of the CMOS Inverter—The Static Behavior
  5.3.1 Switching Threshold
  5.3.2 Noise Margins
  5.3.3 Robustness Revisited
5.4 Performance of CMOS Inverter: The Dynamic Behavior
  5.4.1 Computing the Capacitances
  5.4.2 Propagation Delay: First-Order Analysis
  5.4.3 Propagation Delay from a Design Perspective
5.5 Power, Energy, and Energy Delay
  5.5.1 Dynamic Power Consumption
  5.5.2 Static Consumption
  5.5.3 Putting It All Together
  5.5.4 Analyzing Power Consumption by Using SPICE
5.6 Perspective: Technology Scaling and its Impact on the Inverter Metrics
5.7 Summary
5.8 To Probe Further




Dell Ex. 1025
Page 68

## 5.1 Introduction

The inverter is truly the nucleus of all digital designs. Once its operation and properties are clearly understood, designing more intricate structures such as logic gates, adders, multipliers, and microprocessors is greatly simplified. The electrical behavior of these complex circuits can be almost completely derived by extrapolating the results obtained for inverters. The analysis of inverters can be extended to explain the behavior of more complex gates such as NAND, NOR, or XOR, which in turn form the building blocks for modules such as multipliers and processors.

In this chapter, we focus on a single incarnation of the inverter gate—the static CMOS inverter. This is certainly the most popular inverter at present, and therefore deserves special attention. We analyze the gate with respect to the different design metrics that were outlined in Chapter 1:

- *cost*, expressed by the complexity and area
- *integrity and robustness*, expressed by the static (or steady-state) behavior
- *performance*, determined by the dynamic (or transient) response
- *energy efficiency*, set by the energy and power consumption

Using this analysis, we develop a model of the gate and identify its design parameters. We develop methods to choose the parameter values so that the resulting design meets the desired specifications. While each of these parameters can easily be quantified for a given technology, we also discuss how they are affected by *scaling of the technology*.

While the chapter focuses uniquely on the CMOS inverter, in the next chapter, we see that the same methodology also applies to other gate topologies.

## 5.2 The Static CMOS Inverter—An Intuitive Perspective

Figure 5-1 shows the circuit diagram of a static CMOS inverter. Its operation is readily understood with the aid of the simple switch model of the MOS transistor that we introduced in Chapter 3 (see Figure 3-26). The transistor is nothing more than a switch with an infinite off-

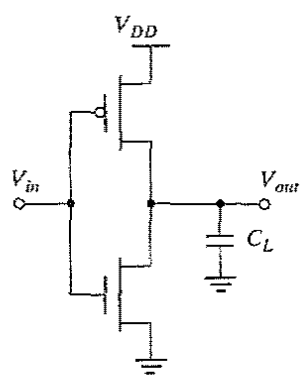



**Figure 5-1** Static CMOS inverter. $V_{DD}$ stands for the supply voltage.

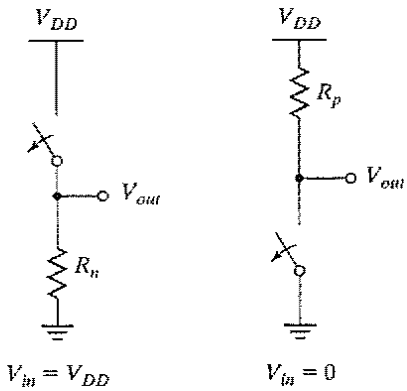



**Figure 5-2** Switch models of CMOS inverter.

resistance (for $|V_{GS}| < |V_T|$) and a finite on-resistance (for $|V_{GS}| > |V_T|$). This leads to the following interpretation of the inverter. When $V_{in}$ is high and equal to $V_{DD}$, the NMOS transistor is on and the PMOS is off. This yields the equivalent circuit of Figure 5-2a. A direct path exists between $V_{out}$ and the ground node, resulting in a steady-state value of 0 V. On the other hand, when the input voltage is low (0 V), NMOS and PMOS transistors are off and on, respectively. The equivalent circuit of Figure 5-2b shows that a path exists between $V_{DD}$ and $V_{out}$, yielding a high output voltage. The gate clearly functions as an inverter.

A number of other important properties of static CMOS can be derived from this switch level view:

- The high and low output levels equal $V_{DD}$ and *GND*, respectively; in other words, the voltage swing is equal to the supply voltage. This results in high noise margins.
- The logic levels are not dependent upon the relative device sizes, so that the transistors can be minimum size. Gates with this property are called *ratioless*. This is in contrast with *ratioed logic*, where logic levels are determined by the relative dimensions of the composing transistors.
- In steady state, there always exists a path with finite resistance between the output and either $V_{DD}$ or *GND*. A well-designed CMOS inverter, therefore, has a *low output impedance*, which makes it less sensitive to noise and disturbances. Typical values of the output resistance are in kΩ range.
- The *input resistance* of the CMOS inverter is extremely high, as the gate of an MOS transistor is a virtually perfect insulator and draws no dc input current. Since the input node of the inverter only connects to transistor gates, the steady-state input current is nearly zero. A single inverter can theoretically drive an infinite number of gates (or have an infinite fan-out) and still be functionally operational; however, increasing the fan-out also increases the propagation delay, as will become clear shortly. Although fan-out does not have any effect on the steady-state behavior, it degrades the transient response.

- No direct path exists between the supply and ground rails under steady-state operating conditions (i.e., when the input and outputs remain constant). The absence of current flow (ignoring leakage currents) means that the gate does not consume any static power.

**SIDELINE:** The preceding observation, while seemingly obvious, is of crucial importance, and is one of the primary reasons CMOS is the digital technology of choice at present. The situation was very different in the 1970s and early 1980s. All early microprocessors—such as the Intel 4004—were implemented in a pure NMOS technology. The lack of complementary devices (such as the NMOS and PMOS transistor) in such a technology makes the realization of inverters with zero static power nontrivial. The resulting static power consumption puts a firm upper bound on the number of gates that can be integrated on a single die; hence, the forced move to CMOS in the 1980s, when scaling of the technology allowed for higher integration densities.

The nature and the form of the voltage-transfer characteristic (VTC) can be graphically deduced by superimposing the current characteristics of the NMOS and the PMOS devices. Such a graphical construction is traditionally called *a load-line plot*. It requires that the *I-V* curves of the NMOS and PMOS devices are transformed onto a common coordinate set. We have selected the input voltage $V_{in}$, the output voltage $V_{out}$ and the NMOS drain current $I_{DN}$ as the variables of choice. The PMOS *I-V* relations can be translated into this variable space by the following relations (the subscripts $n$ and $p$ denote the NMOS and PMOS devices, respectively):

$$\begin{aligned} I_{DSp} &= -I_{DSn} \\ V_{GSn} &= V_{in}; \quad V_{GSp} = V_{in} - V_{DD} \\ V_{DSn} &= V_{out}; \quad V_{DSp} = V_{out} - V_{DD} \end{aligned} \tag{5.1}$$

The load-line curves of the PMOS device are obtained by a mirroring around the $x$-axis and a horizontal shift over $V_{DD}$. This procedure is outlined in Figure 5-3, where the subsequent steps to adjust the original PMOS *I-V* curves to the common coordinate set $V_{in}$, $V_{out}$, and $I_{Dn}$ are illustrated.

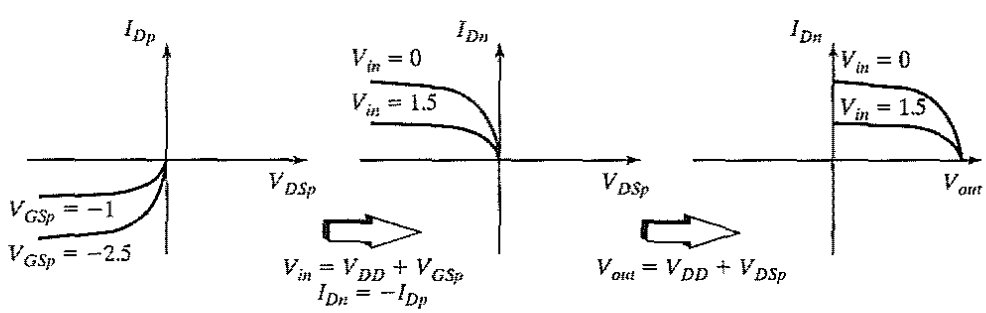



**Figure 5-3** Transforming PMOS *I-V* characteristic to a common coordinate set (assuming $V_{DD} = 2.5$ V).

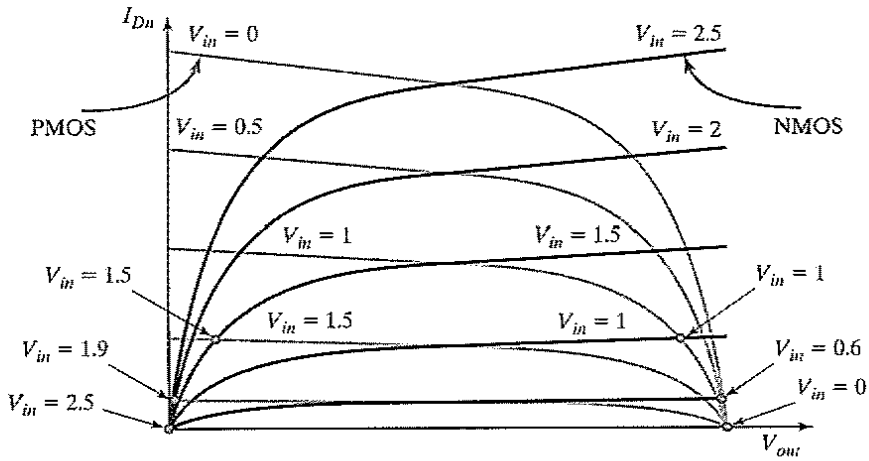



**Figure 5-4** Load curves for NMOS and PMOS transistors of the static CMOS inverter ($V_{DD} = 2.5$ V). The dots represent the dc operation points for various input voltages.

The resulting load lines are plotted in Figure 5-4. For a dc operating point to be valid, the currents through the NMOS and PMOS devices must be equal. Graphically, this means that the dc points must be located at the intersection of corresponding load lines. A number of those points (for $V_{in} = 0$, 0.5, 1, 1.5, 2, and 2.5 V) are marked on the graph. As can be seen, all operating points are located either at the high or low output levels. The VTC of the inverter thus exhibits a very narrow transition zone. This results from the high gain during the switching transient, when both NMOS and PMOS are simultaneously on and in saturation. In that operation region, a small change in the input voltage results in a large output variation. All these observations translate into the VTC shown in Figure 5-5.

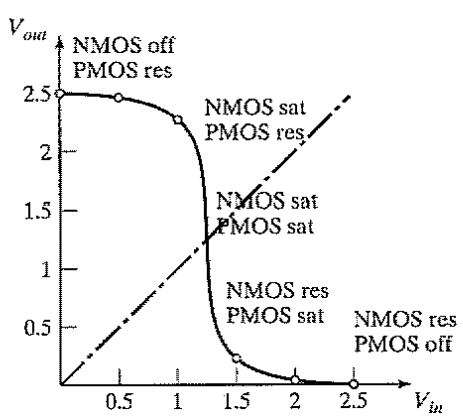



**Figure 5-5** VTC of static CMOS inverter, derived from Figure 5-4 ($V_{DD} = 2.5$ V). For each operation region, the modes of the transistors are annotated—off, res(istive), or sat(urated).

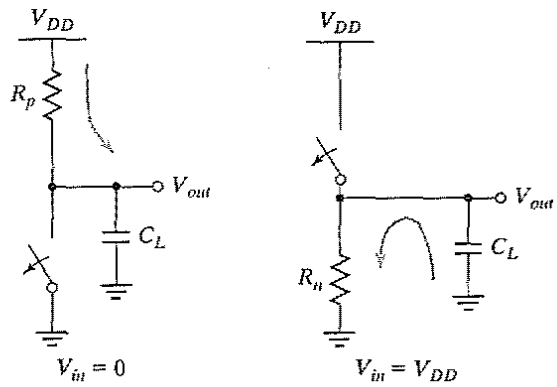



**Figure 5-6** Switch model of dynamic behavior of static CMOS inverter.

Before going into the analytical details of the operation of the CMOS inverter, a qualitative analysis of the transient behavior of the gate is appropriate. This response is dominated mainly by the output capacitance of the gate, $C_L$, which is composed of the drain diffusion capacitances of the NMOS and PMOS transistors, the capacitance of the connecting wires, and the input capacitance of the fan-out gates. Assuming temporarily that the transistors switch instantaneously, we can get an approximate idea of the transient response by using the simplified switch model again. Let us first consider the low-to-high transition (see Figure 5-6a). The gate response time is simply determined by the time it takes to charge the capacitor $C_L$ through the resistor $R_p$. In Example 4.5, we learned that the propagation delay of such a network is proportional to the time constant $R_pC_L$. **Hence, a fast gate is built either by keeping the output capacitance small or by decreasing the on-resistance of the transistor.** The latter is achieved by increasing the $W/L$ ratio of the device. Similar considerations are valid for the high-to-low transition (Figure 5-6b), which is dominated by the $R_nC_L$ time constant. The reader should be aware that the on-resistance of the NMOS and PMOS transistor is not constant; rather, it is a nonlinear function of the voltage across the transistor. This complicates the exact determination of the propagation delay. (An in-depth analysis of how to analyze and optimize the performance of the static CMOS inverter is offered in Section 5.4.)

## 5.3 Evaluating the Robustness of the CMOS Inverter—The Static Behavior

In the preceding qualitative discussion, the overall shape of the voltage-transfer characteristic of the static CMOS inverter was sketched, and the values of $V_{OH}$ and $V_{OL}$—which are evaluated to $V_{DD}$ and $GND$, respectively—were derived. It remains to determine the precise values of $V_M$, $V_{IH}$, and $V_{IL}$, as well as the noise margins.

### 5.3.1 Switching Threshold

The switching threshold $V_M$ is defined as the point where $V_{in} = V_{out}$. Its value can be obtained graphically from the intersection of the VTC with the line given by $V_{in} = V_{out}$ (see Figure 5-5). In this region, both PMOS and NMOS are always saturated, since $V_{DS} = V_{GS}$. An analytical expression for $V_M$ is obtained by equating the currents through the transistors. We solve for the case in which the supply voltage is high enough so that the devices can be assumed to be velocity-saturated (or $V_{DSAT} < V_M - V_T$). Furthermore, we ignore the channel length modulation effects. We have

$$k_n V_{DSATn}\left(V_M - V_{Tn} - \frac{V_{DSATn}}{2}\right) + k_p V_{DSATp}\left(V_M - V_{DD} - V_{Tp}\frac{V_{DSATp}}{2}\right) = 0 \tag{5.2}$$

Solving for $V_M$ yields

$$V_M = \frac{\left(V_{Tn} + \frac{V_{DSATn}}{2}\right) + r\left(V_{DD} + V_{Tp} + \frac{V_{DSATp}}{2}\right)}{1 + r} \quad \text{with} \quad r = \frac{k_p V_{DSATp}}{k_n V_{DSATn}} = \frac{\upsilon_{satp} W_p}{\upsilon_{satn} W_n} \tag{5.3}$$

assuming identical oxide thicknesses for PMOS and NMOS transistors. For large values of $V_{DD}$ (compared with threshold and saturation voltages), Eq. (5.3) can be simplified:

$$V_M \approx \frac{r V_{DD}}{1 + r} \tag{5.4}$$

Equation (5.4) states that the switching threshold is set by the ratio $r$, which compares the relative driving strengths of the PMOS and NMOS transistors. It is generally desirable for $V_M$ to be located around the middle of the available voltage swing (or at $V_{DD}/2$), since this results in comparable values for the low and high noise margins. This requires $r$ to be approximately 1, which is equivalent to sizing the PMOS device so that $(W/L)_p = (W/L)_n \times (V_{DSATn}k'_n)/(V_{DSATn}k'_p)$. To move $V_M$ upwards, a larger value of $r$ is required, which means making the PMOS wider. Increasing the strength of the NMOS, on the other hand, moves the switching threshold closer to GND.

From Eq. (5.2), we derive the required ratio of PMOS to NMOS transistor sizes such that the switching threshold is set to a desired value $V_M$:

$$\frac{(W/L)_p}{(W/L)_n} = \frac{k'_n V_{DSATn}(V_M - V_{Tn} - V_{DSATn}/2)}{k'_p V_{DSATp}(V_{DD} - V_M + V_{Tp} + V_{DSATp}/2)} \tag{5.5}$$

When using this expression, make sure that the assumption that both devices are velocity saturated still holds for the chosen operation point.

**Problem 5.1 Inverter Switching Threshold for Long-Channel Devices, or Low-Supply Voltages**

The preceding expressions were derived under the assumption that the transistors are velocity saturated. When the PMOS and NMOS are long-channel devices, or when the supply voltage is low,

velocity saturation does not occur ($V_M - V_T < V_{DSAT}$). Under these circumstances, the following equation holds for $V_M$:

$$V_M = \frac{V_{Tn} + r(V_{DD} + V_{Tp})}{1 + r} \quad \text{with } r = \sqrt{\frac{-k_p}{k_n}} \tag{5.6}$$

Derive this equation.

## Design Technique—Maximizing the Noise Margins

When designing static CMOS circuits, it is advisable to balance the driving strengths of the transistors by making the PMOS section wider than the NMOS section if maximizing the noise margins and obtaining symmetrical characteristics are desired. The required ratio is given by Eq. (5.5). ■

### Example 5.1 Switching Threshold of CMOS Inverter

We derive the sizes of PMOS and NMOS transistors such that the switching threshold of a CMOS inverter, implemented in our generic 0.25 μm CMOS process, is located in the middle between the supply rails. We use the process parameters presented in Example 3.7, and assume a supply voltage of 2.5 V. The minimum size device has a width-to-length ratio of 1.5. With the aid of Eq. (5.5), we find that

$$\frac{(W/L)_p}{(W/L)_n} = \frac{115 \times 10^{-6}}{30 \times 10^{-6}} \times \frac{0.63}{1.0} \times \frac{(1.25 - 0.43 - 0.63/2)}{(1.25 - 0.4 - 1.0/2)} = 3.5$$

Figure 5-7 plots the values of switching threshold as a function of the PMOS-to-NMOS ratio, as obtained by circuit simulation. The simulated PMOS-to-NMOS ratio of 3.4 for a 1.25-V switching threshold confirms the value predicted by Eq. (5.5).

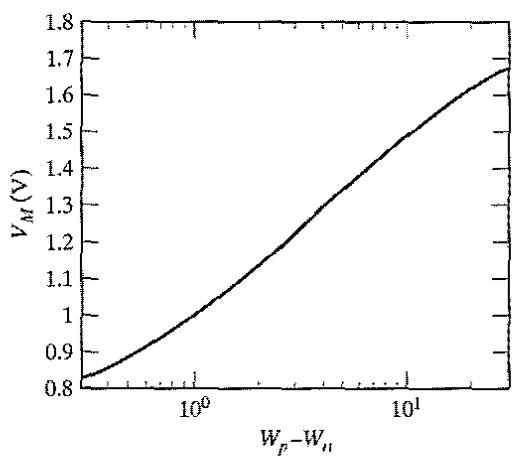
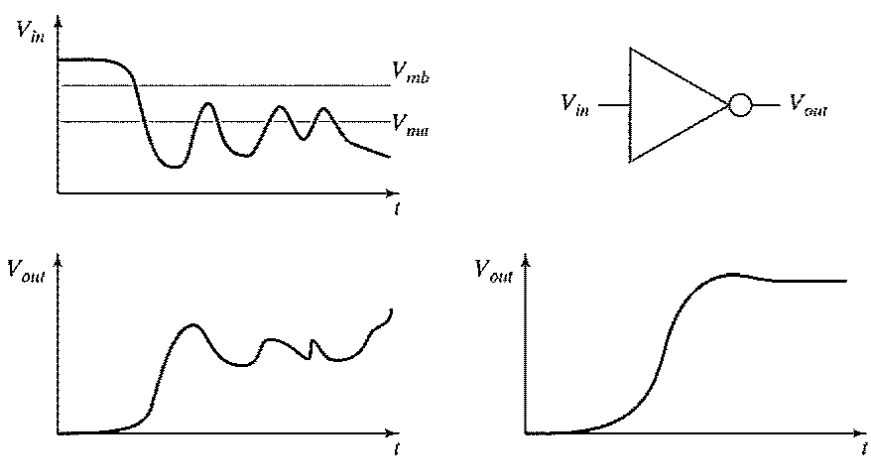



**Figure 5-7** Simulated inverter switching threshold versus PMOS-to-NMOS ratio (0.25-μm CMOS, $V_{DD} = 2.5$ V).

An analysis of the curve of Figure 5-7 leads to some interesting observations:

1. $V_M$ is relatively insensitive to variations in the device ratio. This means that small variations of the ratio (e.g., making it 3 or 2.5) do not disturb the transfer characteristic that much. It is therefore an accepted practice in industrial designs to set the width of the PMOS transistor to values smaller than those required for exact symmetry. For the preceding example, setting the ratio to 3, 2.5, and 2 yields switching thresholds of 1.22 V, 1.18 V, and 1.13 V, respectively.
2. The effect of changing the $W_p$-to-$W_n$ ratio is to shift the transient region of the VTC. Increasing the width of the PMOS or the NMOS moves $V_M$ toward $V_{DD}$ or $GND$, respectively. This property can be very useful, as asymmetrical transfer characteristics are actually desirable in some designs. This is demonstrated by the example of Figure 5-8. The incoming signal $V_{in}$ has a very noisy zero value. Passing this signal through a symmetrical inverter would lead to erroneous values (Figure 5-8a). This can be addressed by raising the threshold of the inverter, which results in a correct response (Figure 5-8b). Later in the text we will see other circuit instances in which inverters with asymmetrical switching thresholds are desirable. Changing the switching threshold by a considerable amount, however, is not easy, especially when the ratio of supply voltage to transistor threshold is relatively small (2.5/0.4 = 6, for our particular example). To move the threshold to 1.5 V requires a transistor ratio of 11, and further increases are prohibitively expensive. Observe that Figure 5-7 is plotted in a semilog format.




(a) Response of standard inverter

(b) Response of inverter with modified threshold

**Figure 5-8** Changing the inverter threshold can improve the circuit reliability.