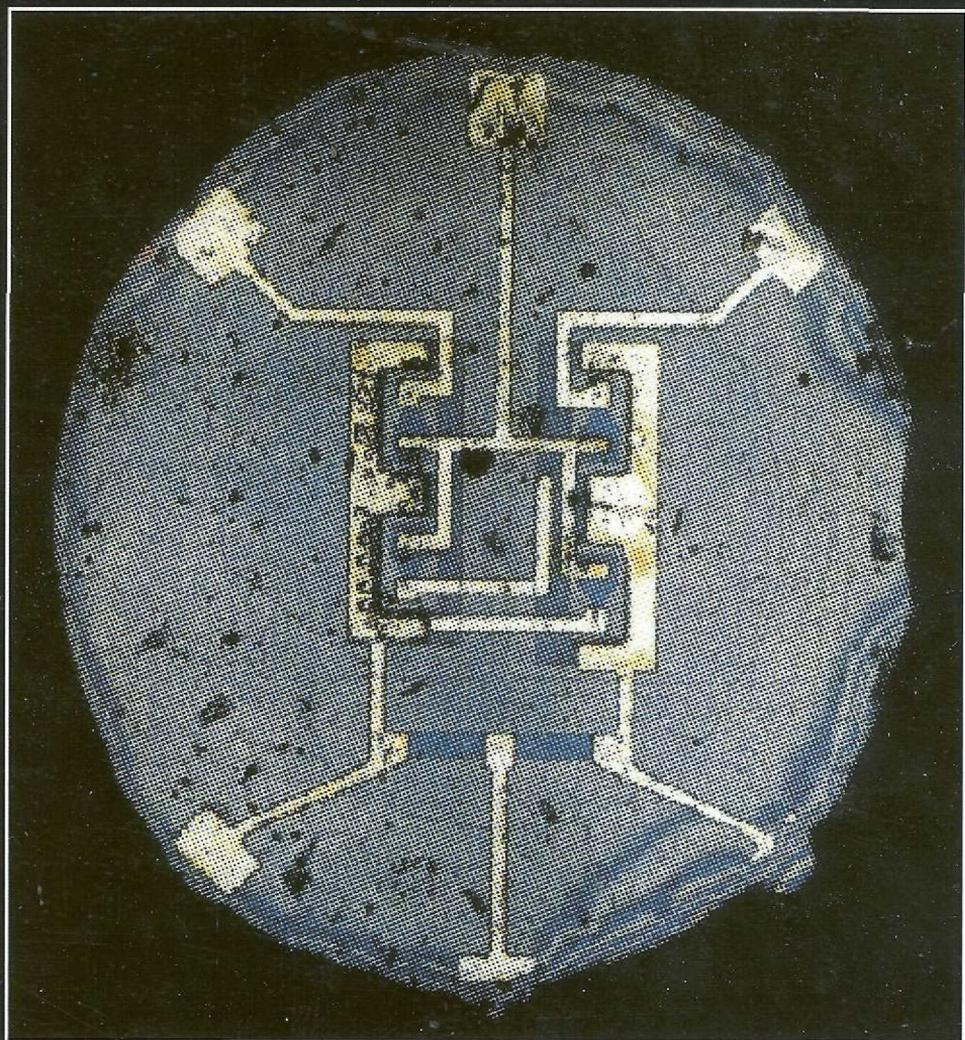


ULSI DEVICES

Edited by
C. Y. Chang S. M. Sze



ULSI DEVICES

*

ULSI DEVICES

Edited by

C. Y. Chang

National Chiao Tung University

S. M. Sze

National Chiao Tung University and
National Nano Device Laboratories



A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS, INC.

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper. ∞

Copyright © 2000 by John Wiley & Sons. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ @ WILEY.COM.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging-in-Publication Data:

ULSI devices / edited by C. Y. Chang, S. M. Sze

p. cm.

“A Wiley-Interscience publication.”

ISBN 0-471-24067-2 (alk. paper)

1. Integrated circuits—Ultra large scale integration.
2. Semiconductors. I. Chang, C. Y. II. Sze, S. M., 1936–
TK7874.76.U47 2000
621.39'5—dc21

99-29979

Printed in the United States of America.

10 9 8 7 6 5 4 3

CONTENTS

CONTRIBUTORS	ix
PREFACE	xi
1. Introduction	1
<i>C. Y. Chang and S. M. Sze</i>	
1.1 Semiconductor Industry	1
1.2 Milestones in ULSI-Related Devices	3
1.3 Technology Trends	9
1.4 Organization of the Book	13
References	15
PART I DEVICE FUNDAMENTALS	
2. Bipolar Transistor Fundamentals	19
<i>E. Kasper</i>	
2.1 Introduction	19
2.2 Structure and Operating Regimes	20
2.3 The Inner-Box-Shaped Transistor	25
2.4 Self-Adjusted Transistor Structures	51
2.5 Heterojunction Bipolar Transistor	60
2.6 Summary and Future Trends	66
References	67
Problems	71
3. MOSFET Fundamentals	73
<i>P. Wong</i>	
3.1 Introduction	73
3.2 MOSFET Device Physics: First-Order Models	74
3.3 Device Scaling	84
3.4 Short-Channel Effects	88
3.5 Transport Properties	102
3.6 Parasitic Effects	110
3.7 Evolution of MOSFET Device Structures	130

3.8 Summary and Future Trends	135
References	139
Problems	147
4. Device Miniaturization and Simulation	149
<i>S. Banerjee and B. Streetman</i>	
4.1 Introduction	149
4.2 Band Structure	150
4.3 Semiclassical Electron Dynamics	160
4.4 Equilibrium Statistics	163
4.5 Scattering Theory	170
4.6 Monte Carlo Simulations	182
4.7 Boltzmann Transport Equation	185
4.8 Summary and Future Trends	211
References	216
Problems	218
 PART II DEVICE BUILDING BLOCKS AND ADVANCED DEVICE STRUCTURES	
5. SOI and Three-Dimensional Structures	221
<i>J. P. Colinge</i>	
5.1 Introduction	221
5.2 SOI Substrates	222
5.3 The SOI MOSFET	230
5.4 3D and Novel SOI Devices	239
5.5 SOI Circuits	254
5.6 Summary and Future Trends	263
References	263
Problems	272
6. The Hot-Carrier Effect	275
<i>B. Doyle</i>	
6.1 Introduction	275
6.2 Damage Identification	278
6.3 Gate Voltage Dependence of Stress	281
6.4 Hot-Carrier Lifetime Estimation: AC and DC	294
6.5 Hot-Carrier Measurements	302
6.6 Structure Dependence	310
6.7 Process Dependence	314
6.8 Summary and Future Trends	322
References	324
Problems	332

7. DRAM and SRAM	333
<i>S. Shichijo</i>	
7.1 Introduction	333
7.2 DRAM Cell Structures	333
7.3 DRAM Operation Principle	340
7.4 DRAM Circuits	346
7.5 SRAM Memory Cell Structure	358
7.6 SRAM Operation Principle	362
7.7 SRAM Circuits	367
7.8 Summary and Future Trends	371
References	372
Problems	374
8. Nonvolatile Memory	377
<i>J. Caywood and G. Derbenwich</i>	
8.1 Introduction	377
8.2 Floating-Gate Memory	378
8.3 Floating-Gate Memory Arrays	393
8.4 Reliability of Floating-Gate Memories	421
8.5 Future Trends and Summary of Floating-Gate Memory	426
8.6 Silicon Nitride Memory	430
8.7 Ferroelectric Memory	448
References	463
Problems	470
PART III CIRCUIT BUILDING BLOCKS AND SYSTEM-IN-CHIP CONCEPT	
9. CMOS Digital and Analog Building Block Circuits for Mixed-Signal Applications	477
<i>D. Pehlke and M.-C. F. Chang</i>	
9.1 Introduction	477
9.2 CMOS for Digital Applications	479
9.3 CMOS Technology for Analog and RF Applications	492
9.4 CMOS Low-Noise Amplifiers	493
9.5 Mixers and Frequency Translation	508
9.6 CMOS Voltage-Controlled Oscillators	520
9.7 CMOS Power Amplifiers	535
9.8 Summary and Future Trends	540
References	541
Problems	543

10. High-Speed or Low-Voltage, Low-Power Operations	547
<i>I. C. Chen and W. Liu</i>	
10.1 Introduction	547
10.2 High-Speed Considerations for Digital Applications	548
10.3 Low Voltage/Low Power Considerations for Digital Applications	581
10.4 Cutoff and Maximum Oscillation Frequencies	585
10.5 Large-Signal Power and Efficiency	597
10.6 Noise Figure	607
10.7 Summary and Future Trends	616
References	618
Problems	627
11. System-on-Chip Concepts	631
<i>M. Pelgrom</i>	
11.1 Introduction	631
11.2 Embedded Modules on an IC	633
11.3 Technology for Mixed-Signal Circuits	639
11.4 Technology Limits	645
11.5 Analog Interfaces	651
11.6 Integration of Blocks	664
11.7 System-on-Chip Concept	676
11.8 Summary and Future Trends	679
References	681
Problems	683
Appendix A. List of Symbols	685
Appendix B. International System of Units (SI Units)	687
Appendix C. Unit Prefixes	689
Appendix D. Greek Alphabet	691
Appendix E. Physical Constants	693
Appendix F. Properties of Si at 300 K	695
INDEX	697

CONTRIBUTORS

Sanjay K. Banerjee, The University of Texas at Austin, Austin, Texas, USA

John Caywood, SubMicron Circuits, Inc., San Jose, California, USA

C. Y. Chang, National Chiao Tung University, Hsinchu, Taiwan, ROC

Mau-Chung Frank Chang, University of California at Los Angeles, Los Angeles, California, USA

I. C. Chen, Worldwide Semiconductor Manufacturing Corp., Hsinchu, Taiwan, ROC

Jean-Pierre Colinge, Department of Electrical and Computer Engineering, University of California, Davis, California, USA

Gary Derbenwich, Celis Semiconductor Corp., Colorado Springs, Colorado, USA

Brian Doyle, Intel Corporation, Santa Clara, California, USA

Erich Kasper, Institute of Semiconductor Engineering, University of Stuttgart, Stuttgart, Germany

W. Liu, Texas Instruments, Inc., Dallas, Texas, USA

David R. Pehlke, Rockwell Science Center, Thousand Oaks, California, USA; now at Ericsson, Research Triangle Park, North Carolina, USA

Marcel J. M. Pelgrom, Philips Research Laboratories, Eindhoven, The Netherlands

Hisashi (Sam) Shichijo, Texas Instruments, Inc., Dallas, Texas, USA

Ben G. Streetman, The University of Texas at Austin, Austin, Texas, USA

Simon M. Sze, National Chiao Tung University and National Nano Device Laboratories, Hsinchu, Taiwan, ROC

Hon-Sum Philip Wong, IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA

PREFACE

Beginning in 1990, the semiconductor industry has entered the ultra-large-scale integration (ULSI) era with 10 million or more devices per integrated circuit (IC) chip. We expect that this trend toward higher integration levels will continue at least to year 2020. By then, the most advanced IC chips will contain over 1 trillion (10^{12}) devices.

The most important ULSI device is, of course, the metal-oxide-semiconductor field-effect transistor (MOSFET), because of its advantages in device miniaturization, low power dissipation, and high yield compared to all other semiconductor devices. The MOSFET also serves as a basic component for many key device building blocks, including the complementary metal oxide semiconductor (CMOS), the dynamic random access memory (DRAM), and the static random access memory (SRAM). Another important ULSI device is the nonvolatile semiconductor memory (NVSM), which has the most desirable attributes for information storage compared to all other semiconductor memories. The third important ULSI device is the bipolar transistor. Although it is nearly completely replaced by the MOSFET in digital circuits, the bipolar transistor is often the preferred device for analog and mixed analog and digital circuits.

ULSI Devices gives a comprehensive and in-depth coverage of the physics and operational principles of the aforementioned devices and their building blocks. In addition, the book covers device simulation, reliability, digital and analog circuit building blocks, low-power and low-voltage operations, as well as system-on-chip for ULSI applications.

Each chapter starts with an introduction that provides a general discussion of a specific device, a device building block or a circuit building block. Subsequent sections present the physics and operational characteristics of these components, the evolution of device structures, and the ultimate limitation on device or circuit performances. The problem set at the ends of each chapter are an integral part of the development of the topic.

The book is intended as a textbook for senior undergraduate or first-year graduate students in applied physics and electrical and electronics engineering; it assumes that the reader has already acquired an introductory understanding of the physics of semiconductor devices. The book can also serve as a reference for those actively involved in IC design and process development.

In the course of writing this text, many people have assisted us and offered their support. First, we express our appreciation to the management of our academic and industrial institutions, without whose help this book could not have been written. We

have benefited from suggestions made by our reviewers: Dr. S. Cristoloveanu of ENSERG-LPCS, Prof. S. Datta of Purdue University, Prof. L. Der of the University of California, Los Angeles, Prof. K. Hess of the University of Illinois, Prof. C. M. Hu of the University of California, Berkeley, Dr. R. Koch of Mixed Signal Competence Center, Dr. C. C. Lu of Etron Technology, Dr. K. Mistry of Digital Equipment Corporation, Dr. K. K. Ng of Bell Laboratories, Lucent Technologies, Prof. Y. Omura of Kansai University, Dr. B. Prince of Memory Strategies International, Dr. F. J. Shone of Macronix International Company, Dr. J. Slotboom of Philips Research Laboratories, Prof. Y. Tsvividis of Columbia University, and Dr. H. J. Wann of IBM T. J. Watson Research Center.

We are further indebted to Mr. N. Erdos for technical editing of the entire manuscript. We also thank Ms. Y. C. Wang, Ms. M. H. Tai and Ms. W. L. Chen for handling the correspondence with our contributors and reviewers, and Mrs. T. W. Sze for preparing the Appendixes.

We wish to thank the Ministry of Education, ROC (Republic of China), the National Science Council, ROC, and the Spring Foundation of the National Chiao Tung University for their financial support. One of the editors (S. M. Sze) especially thanks the United Microelectronics Corporation (UMC), Taiwan, ROC, for the UMC Chair Professorship grant that provided the environment to work on this book.

Hsinchu, Taiwan
January 2000

C. Y. CHANG
S. M. SZE

Introduction

C. Y. CHANG

National Chiao Tung University
Hsinchu, Taiwan, ROC

S. M. SZE

National Chiao Tung University
National Nano Device Laboratories
Hsinchu, Taiwan, ROC

1.1 SEMICONDUCTOR INDUSTRY

The electronic industry is the largest industry in the world with global sales over 1 trillion dollars.¹⁻³ The foundation of the electronic industry is the semiconductor industry. Figure 1.1 shows the sales volumes of these two industries from 1980 to 1998 and projects the sales to year 2010. Also shown are the world gross national product (GNP), and the sales volumes of automobile and steel industries.²

We note that the electronic industry has surpassed the automobile industry in 1998. If the current trends continue, in year 2010 the sales volume of the electronic industry will reach \$2.7 trillion and will constitute about 10% of the world GNP. The semiconductor industry will grow at an even higher rate to surpass the steel industry in the early twenty-first century and to constitute 30% of the electronic industry in 2010 (i.e., about \$750 billion).

Figure 1.2 shows the composition of the world electronic industry.^{1,2} The largest segment is the computers and peripheral equipment (about 35% in 1998), followed by the telecommunication equipment (~25%), industrial electronics (~15%), consumer electronics (~15%), defense and space electronics (~5%), and transportation (~5%). By the year 2010, the composition will remain essentially the same; however, the computers and peripheral equipment segment will expand more rapidly to capture about 45% of the total market.

The market for the world semiconductor industry^{2,3} is shown in Figure 1.3. The largest segment is the microprocessor and microcontroller units (about

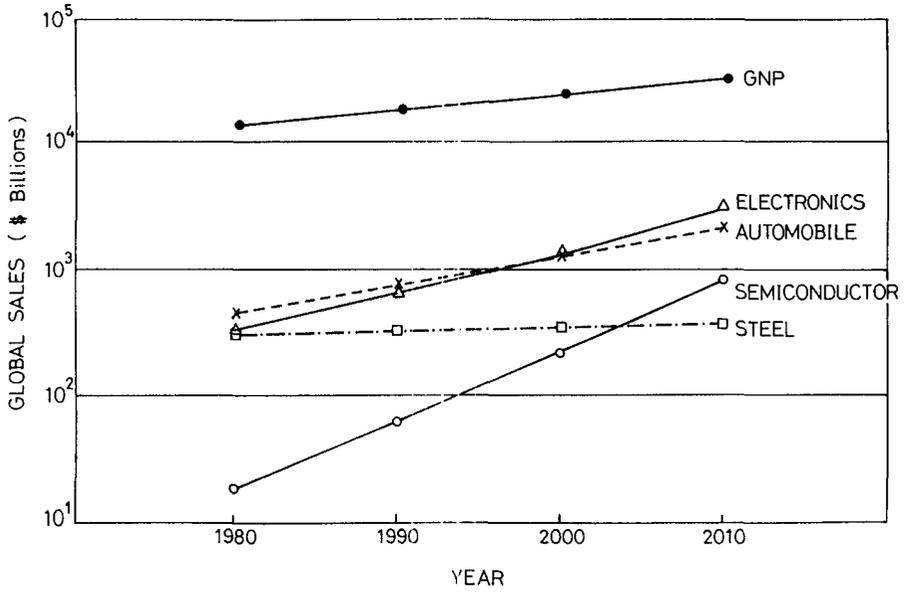


Figure 1.1 World gross national product and sales volumes of the electronics, automobile, semiconductor, and steel industries from 1980 to 1998 and projected to 2010. (After Refs. 1–3.)

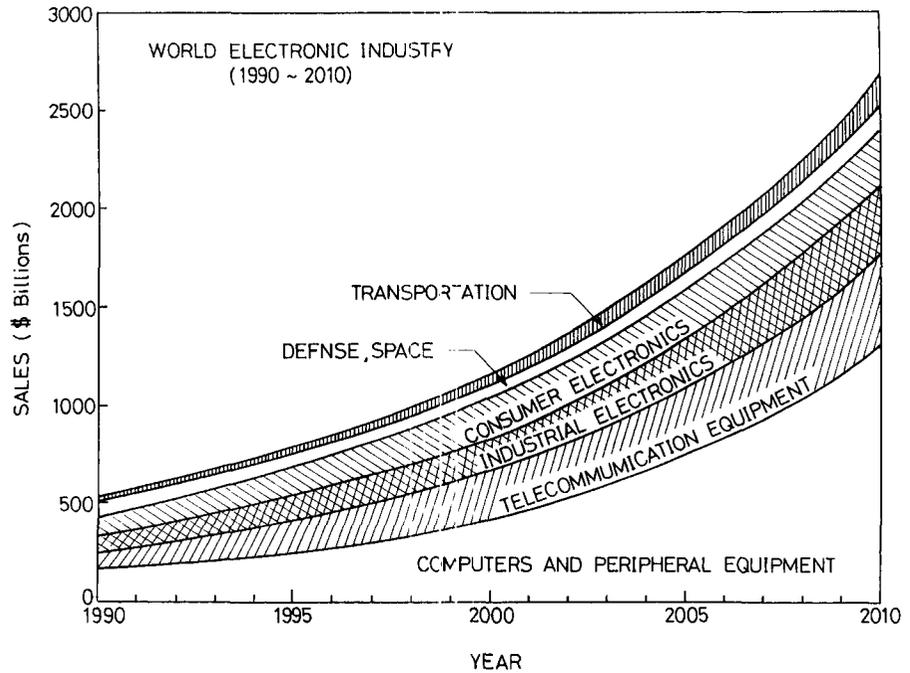


Figure 1.2 Composition of the world electronic industry. (After Refs. 2 and 3.)

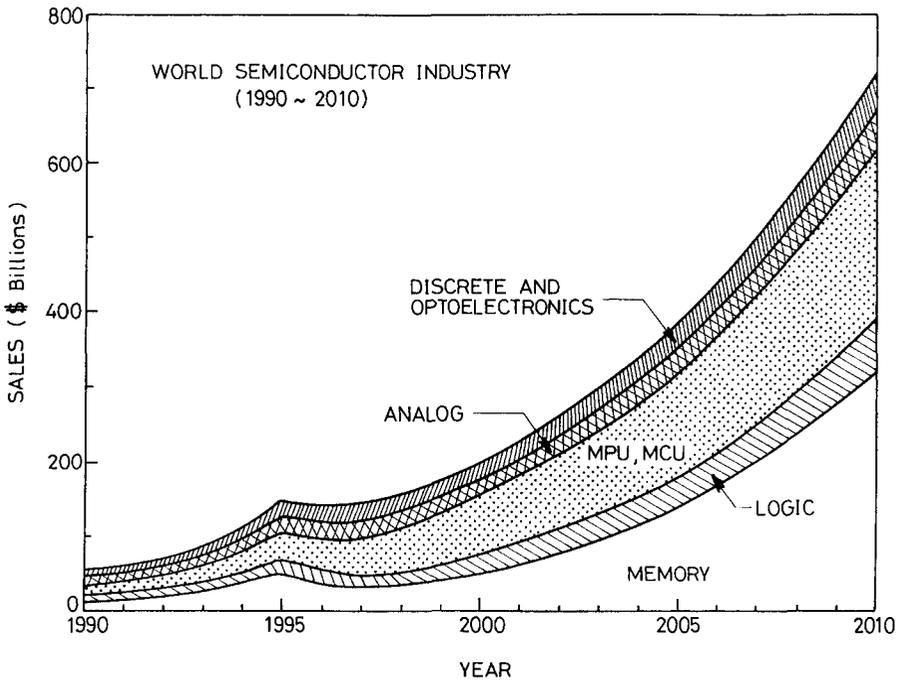


Figure 1.3 Composition of the world semiconductor industry. (After Refs. 2 and 3.)

35% in 1998), followed by memory (~25%), analog (~15%), discrete and optoelectronic (~15%), and logic (~10%). By the year 2010, memory may become the dominant segment and most of the memories will be integrated with microprocessors (and other logic circuits) to form systems-on-a-chip.

1.2 MILESTONES IN ULSI-RELATED DEVICES

The ULSI device is almost synonymous with the silicon MOSFET. This is mainly due to the MOSFET's intrinsic characteristics, which are uniquely suitable for highly complex ICs. In addition, many important device building blocks for ULSI such as CMOS, DRAM, and SRAM are also derived from MOSFET.

However, there are other important devices that can either enhance the MOSFET performance (e.g., bipolar transistor for BiCOMS) or have special features not obtainable from MOSFET (e.g., nonvolatile semiconductor memory for long-term information storage).

Table 1.1 lists some major ULSI related devices and device building blocks in chronological order. In 1947, the point-contact transistor was invented by Bardeen and Brattain.⁴ This was followed by Shockley's classic paper on bipolar transistor.⁵ Figure 1.4 shows the first transistor. The two point contacts at the bottom of the triangular quartz crystal were made from two stripes of gold foil

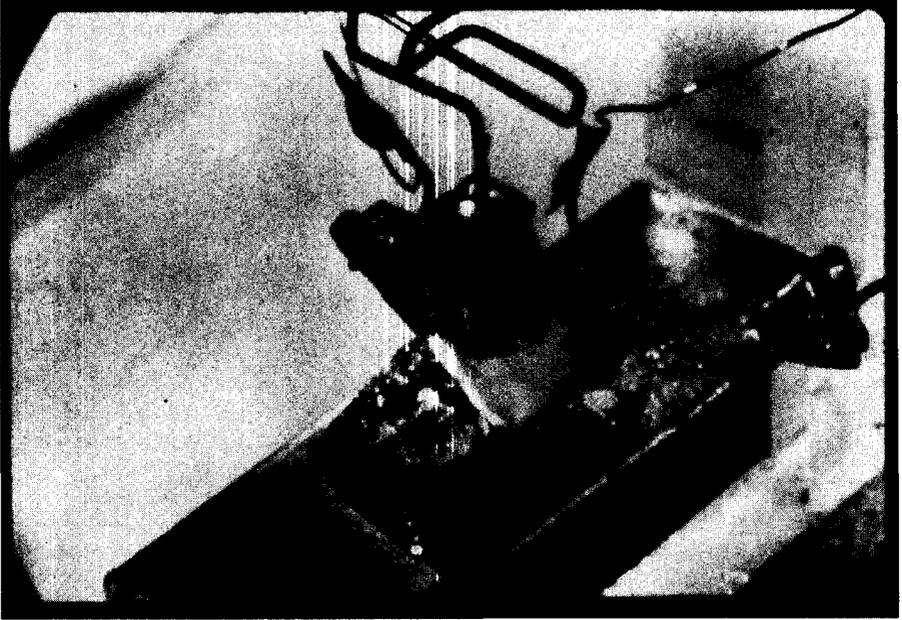


Figure 1.4 The first transistor. (After Bardeen and Brattain, Ref. 4.)

separated by about $50\ \mu\text{m}$ and pressed onto the germanium surface. With one gold contact forward-biased and the other reverse-biased, the transistor action was observed. The invention of the bipolar transistor had an unprecedented impact on the electronics industry, and 1947 marked the beginning of the modern electronics era. In 1957, Kroemer proposed the heterojunction bipolar transistor (HBT) to improve the emitter efficiency; this device is potentially one of the fastest semiconductor devices.⁶ At present, the bipolar transistor has been replaced by MOSFET in a digital circuit. However, the bipolar transistor and

TABLE 1.1 Milestones of ULSI-Related Devices

Year	Device or Device Building Block	Inventor(s) / Author(s)	Institution
1947	Bipolar transistor	Bardeen, Brattain, Shockley	Bell Labs
1957	Heterojunction bipolar transistor	Kroemer	RCA Labs
1958	Integrated circuit	Kilby	TI
1959	Monolithic IC	Noyce	Fairchild
1960	Enhancement-mode MOSFET	Kahng, Atalla	Bell Labs
1963	CMOS	Wanlass, Sah	Fairchild
1967	Nonvolatile semiconductor memory	Kahng, Sze	Bell Labs
1967	One-transistor DRAM	Dennard	IBM
1971	Microprocessor	Hoff et al.	Intel
1994	Room-temperature single-electron transistor	Yano et al.	Hitachi

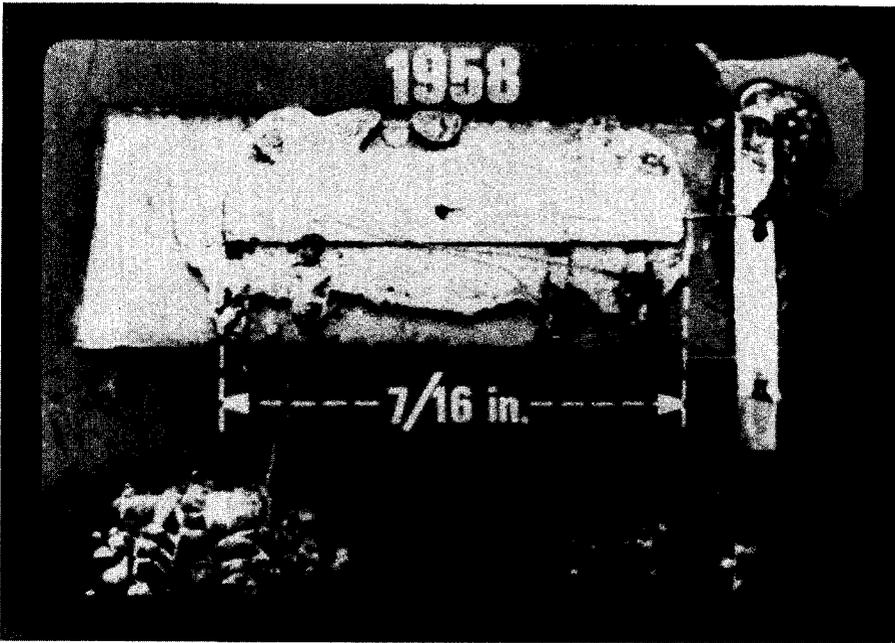


Figure 1.5 The first integrated circuit (After Kilby, Ref. 7.)

HBT are useful for analog and mixed analog and digital circuits. They can be combined with CMOS to form BiCMOS (bipolar CMOS) for high-speed/low-power applications.

In 1958, the first rudimentary integrated circuit (IC) was made by Kilby.⁷ It contained one bipolar transistor, three resistors, and one capacitor, all made in germanium and connected by wire bonding (see Fig. 1.5). In 1959, Noyce⁸ proposed the first monolithic IC by fabricating all devices in a single semiconductor substrate using oxide isolation and aluminum metallization (see Fig. 1.6). These inventions paved the way for the ever-expanding micro-electronic industry.

In 1960, Kahng and Atalla proposed and fabricated the first enhancement-mode MOSFET using a thermally oxidized silicon structure.⁹ Figure 1.7 shows the device, which has a channel length of $25\ \mu\text{m}$ and a gate oxide thickness of about $1000\ \text{\AA}$. The two keyholes are the source and drain contacts, and the top elongated area is the aluminum gate evaporated through a metal mask. Although present-day MOSFETs have been scaled down to deep submicrometer regime, the choice of silicon and thermally grown silicon dioxide used in the first enhancement-mode MOSFET remains the most important combination. The MOSFET and its related ICs now constitute over 90% of the semiconductor device market, and it is the most important device for ULSI applications.

As the complexity of the IC increases, we have moved from NMOS to the CMOS (complementary MOSFET) technology, which employs both NMOS and

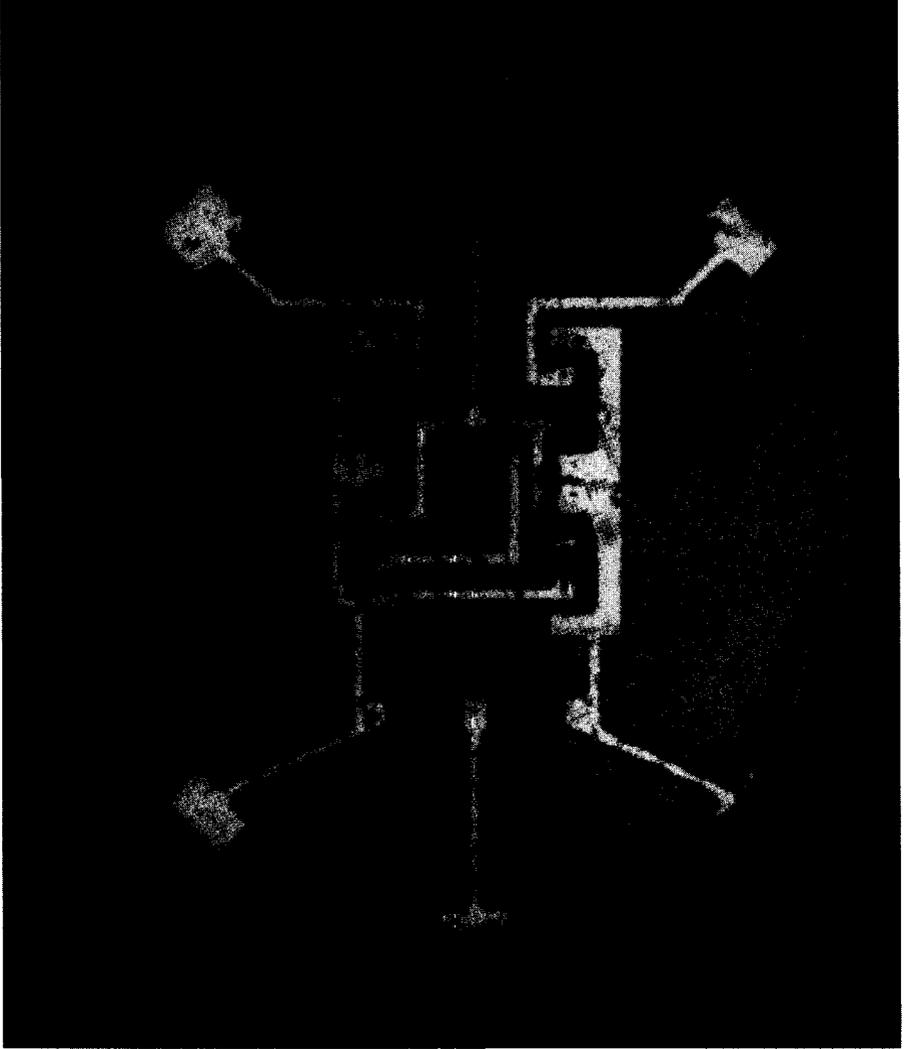


Figure 1.6 The first monolithic integrated circuit. (After Noyce, Ref. 8.)

PMOS to form the logic elements. The CMOS concept¹⁰ was proposed by Wanlass and Sah in 1963. The advantage of CMOS technology is that the logic elements draw significant current only during the transition from one state to another and draw very little current between transitions, allowing power consumption to be minimized. The CMOS technology is the dominant technology for ULSI.

In 1967, an important memory device was invented by Kahng and Sze.¹¹ This is the nonvolatile semiconductor memory (NVSM), which can retain its stored

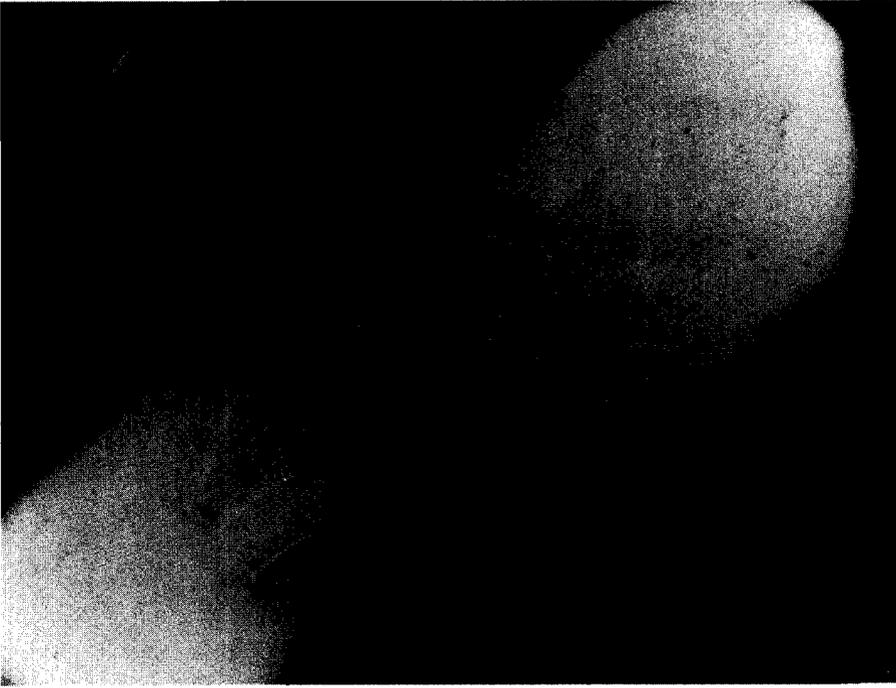


Figure 1.7 The first enhancement-mode MOSFET. (After Kahng and Atalla, Ref. 9.)

information when the power supply is switched off. A schematic diagram of the first NVSM is shown in Figure 1.8*a*. Although it looks like a conventional MOSFET, the major difference is the addition of the floating gate, in which semipermanent charge storage is possible. Because of its attributes of non-volatility, high density, low power consumption, and electrical rewritability, NVSM has become the dominant memory for portable electronic systems such as the cellular phone and the notebook computer.

A limiting case of the floating-gate memory is the single-electron memory cell (SEMC) shown in Figure 1.8*b*. By reducing the length of the floating gate to ultra small dimensions (e.g., 10 nm), we obtain the SEMC. Because of its small size the capacitance is also very small ($\sim 10^{-18}$ F, which is about a million times smaller than that of the device in Fig. 1.8*a*). At this dimension, when an electron tunnels into the floating gate, the tunneling of another electron is blocked due to quantum effect. The SEMC is an ultimate floating-gate memory cell, since we need at least one electron for information storage. The operation of a SEMC at room temperature was first demonstrated by Yano¹² et al. in 1994 (last item in Table 1.1). The SEMC can serve as a building block for multiterabit NVSM.

The one-transistor DRAM, which is a combination of a MOSFET and a charge-storage MOS capacitor, was proposed in 1967 by Dennard.¹³ A circuit representation and a cross-sectional view of the memory cell are shown in Figure

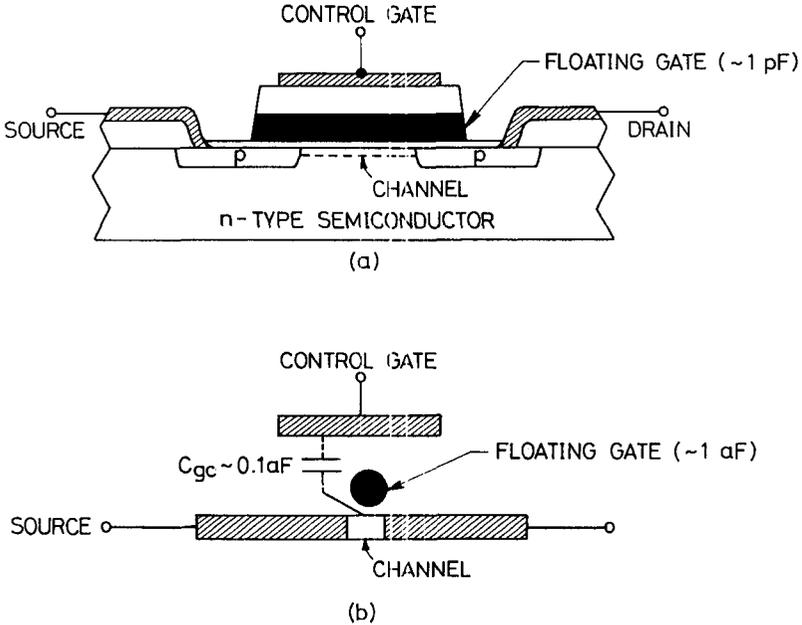


Figure 1.8 (a) The first nonvolatile semiconductor memory with a floating gate (after Kahng and Sze, Ref. 11); (b) a single-electron memory cell—the limiting case of the floating-gate memory (after Yano et al., Ref. 12.)

1.9. Although DRAM is volatile and consumes relatively high power, we expect that DRAM will continue to be the first choice for nonportable electronic systems in the foreseeable future.

In 1971, the first microprocessor was made by Hoff et al.¹⁴ They put the entire central processing unit (CPU) of a simple computer on one chip. It was a 4-bit microprocessor (Intel 4004) with a chip size of 3×4 mm, and contained 2300 MOSFETs. It was fabricated by a p-channel, silicon gate process using an 8- μ m design rule. This microprocessor performed as well as those in \$300,000 IBM computers of the early 1960s—each of which needed a CPU the size of a large desk. This was a major breakthrough for semiconductor industry. At the present time, the microprocessor and the microcontroller constitute the largest segment of the industry.

There are many other semiconductor devices that are potential candidates for ULSI applications. These devices include the ferroelectric nonvolatile memory proposed by Moll and Tarui¹⁵ in 1963, the MESFET by Mead¹⁶ in 1966, the resonant tunneling diode by Chang et al.¹⁷ in 1974, the modulation-doped field-effect transistor (MODFET) by Mimura et al.¹⁸ in 1980, and the charge-injection transistor (CHINT) by Luryi et al.¹⁹ in 1984, the induced-base transistor by Luryi et al.^{20,21} in 1985, and the modulation doped-base hot-electron transistor by Chang²² in 1992. The current status of these devices can be found in standard textbooks or reference books.²³ Their competitive performances are being

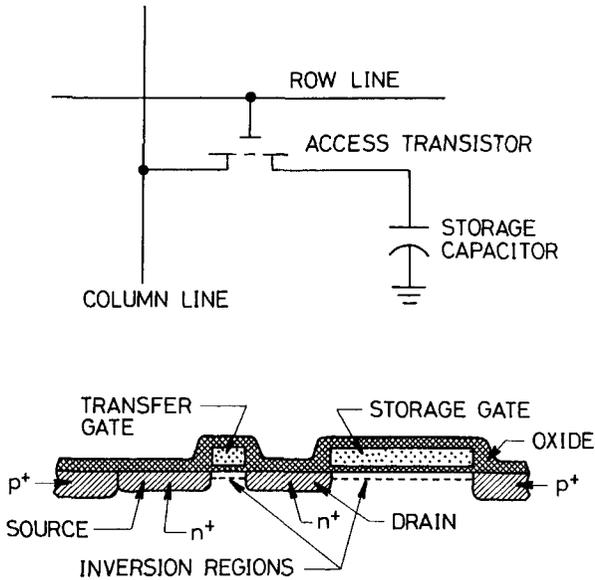


Figure 1.9 Circuit representation and cross-sectional view of the one-transistor DRAM. (After Dennard, Ref. 13.)

continuously evaluated against MOSFET for ULSI logic applications and against DRAM and NVSM for ULSI memory applications.

1.3 TECHNOLOGY TRENDS

The Semiconductor Industry Association (SIA) has recently issued their *Roadmap for Semiconductors*.²⁴ The *Roadmap* is based on the consensus of industrial experts and is shown in Table 1.2. We shall present a brief discussion on the future trends of ULSI technology with reference to the table.

First, *the design rule* will continue to follow the traditional rate of 40% reduction every 3 years. However, the industry has not yet decided which lithography alternatives (e.g., deep ultraviolet, projection electron beam, proximity X-ray, or extreme ultraviolet) will be adopted for future mass production. The design rule, of course, has a major impact on the gate length (usually slightly smaller than the design rule), the minimum contact size (slightly larger than the design rule), the oxide thickness, and the supply voltage. The recommended values are listed in Table 1.2.

Next, for *DRAM capacity*, Figure 1.10 shows the actual DRAM density versus the year of first production from 1979 to 1999. The density for 2002 to 2011 is based on the SIA *Roadmap*. It is interesting to note that before 1999 the DRAM density almost exactly followed the Moore's law,²⁵ that is, it increased by a factor of 4 every 3 years (or doubling every 18 months). However, in the next

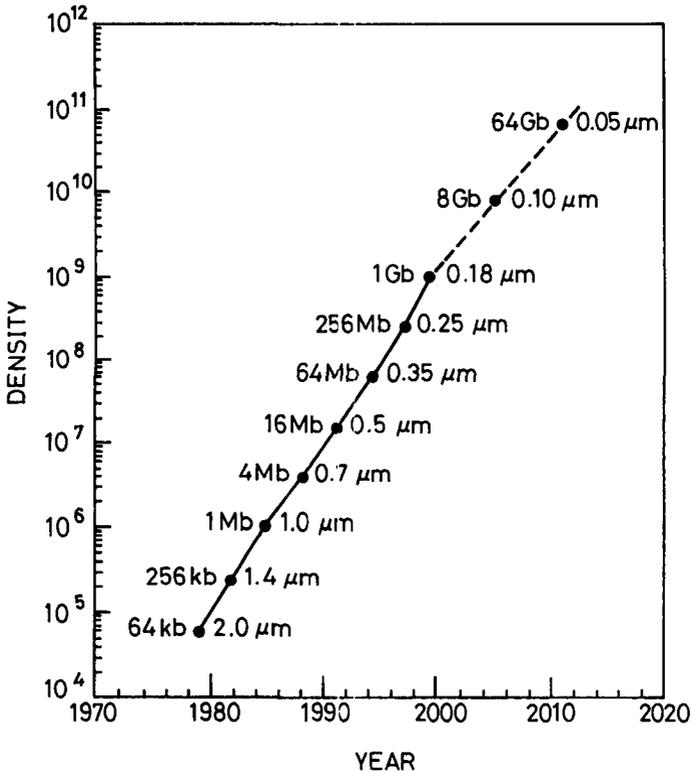


Figure 1.10 Exponential increase of DRAM density versus year based on the SIA *Roadmap*. (After SIA, Ref. 24 and Moore, Ref. 25.)

10 years, the density is expected to increase by a factor of 8 every 6 years or to double every 24 months to reach 64 Gb in year 2011.

Finally, for the *clock frequency*, the dominant ULSI technology is the CMOS technology.^{10,26} Figure 1.11 shows the delay time versus power dissipation for various devices.²⁷ We note that the CMOS indeed wins in circuit density, power dissipation, and system speed. For example, at a $0.1\text{-}\mu\text{m}$ design rule, a CMOS inverter can have a speed of 10 ps per gate, and a power dissipation of $1 \mu\text{W}$ per gate, corresponding to a power dissipation–delay time product of only 10^{-17} J . This is about 100 times smaller than a MODFET, and 1000 times smaller than a bipolar logic circuit based on the same design rule.

Figure 1.12 shows the exponential increase of the microprocessor computational power from 1980 to 1999 and projected to 2010. The computational power increases by a factor of 4 every 3 years. Currently, a Pentium personal computer has the computational power of a supercomputer, Cray 1, of the late 1960s. If the trends continue, we will reach 10^5 MIPS (million instructions per second) or 100 GIPS (billion instructions per second) in year 2010. With such dramatic

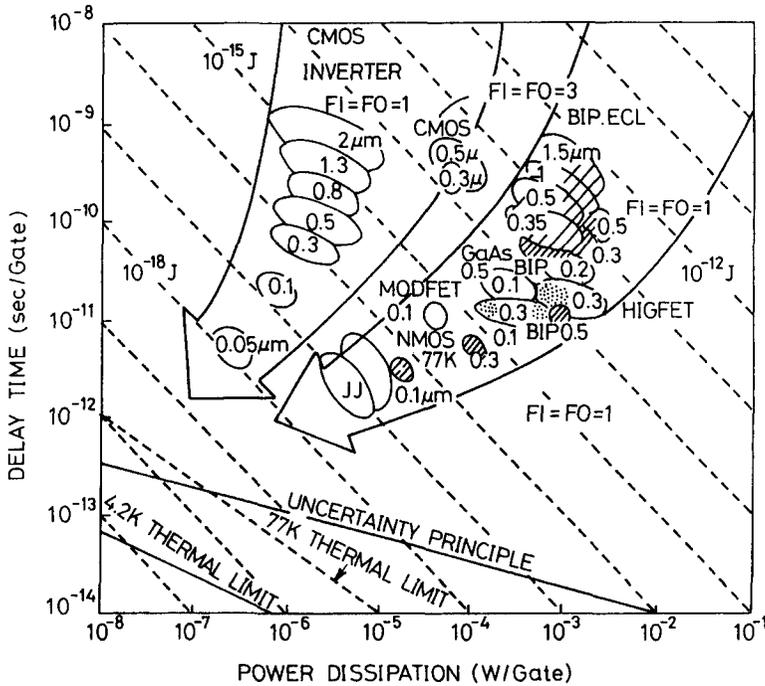


Figure 1.11 Delay time versus power dissipation for CMOS and other devices. (After Ref. 27.)

progress, we will soon have the computational power of a supercomputer Cray 2 on a single chip.

To increase the system speed, we also have to minimize the RC (resistance \times capacitance) delays due to parasitic resistance and capacitance. Low-dielectric-constant insulators should be used as interlayer materials. Copper is recommended as the metal material, and multilayer interconnection is recommended to minimize the total wiring length. These considerations are listed in Table 1.2.

As mentioned previously, a key advantage of CMOS technology is its low power dissipation. This is of paramount importance for full portability of electronic systems. It is anticipated that in the early twenty-first century, 50% of the electronic systems will be portable. While CMOS is the primary choice for logic design, the NVSM is the primary choice for information storage. This is because of NVSM's low power consumption compared to all other semiconductor memories.

The impact of NVSM on the semiconductor industry is shown in Figure 1.13, which illustrates the growth curves for different technology drivers.²⁸ At the beginning of the modern electronic era (1950–1970), the bipolar transistor was the technology driver. From 1970 to 1990, the DRAM and microprocessor based

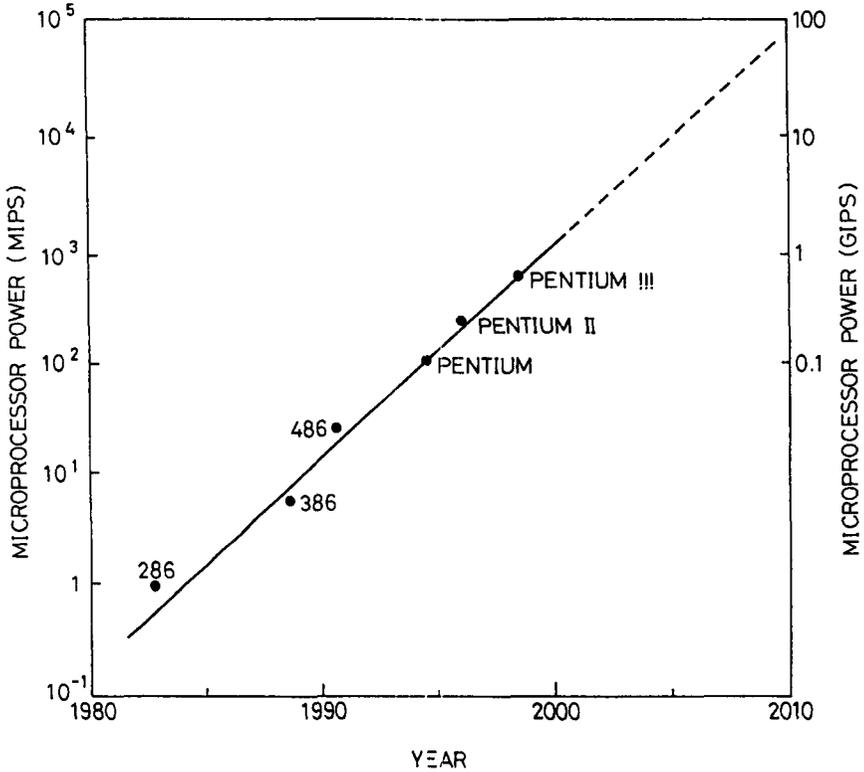


Figure 1.12 Exponential increase of the microprocessor computational power versus year. (After Ref. 25.)

TABLE 1.2 International Technology Roadmap for Semiconductors 1999

Year of the first Production	1997	1999	2002	2005	2008	2011
Design Rule (nm)	250	180	130	100	70	50
DRAM Capacity	256 M	1 G	—	8 G	—	64 G
Clock Frequency (MHz)	750	1200	1600	2000	2500	3000
Gate Length (nm)	200	140	100	65	45	30–32
Minimum Contact Size (nm)	280	200	150	130	100	70
Oxide Thickness (nm)	3–4	1.9–2.5	1.3–1.7	0.9–1.1	<1.0	<1.0
Supply Voltage (V)	1.8–2.5	1.5–1.8	1.2–1.5	0.9–1.2	0.6–0.9	0.5–0.6
Metal Material (s)	Al-Cu	Cu	Cu	Cu	Cu	Cu
Metal Layers	6	6–7	7–8	8–9	9	10

on CMOS technology were the technology drivers due to the rapid growth of personal computers and advanced electronic systems. Since 1990, NVSM has been the technology driver, mainly because of the rapid increase of portable electronic systems.²⁹

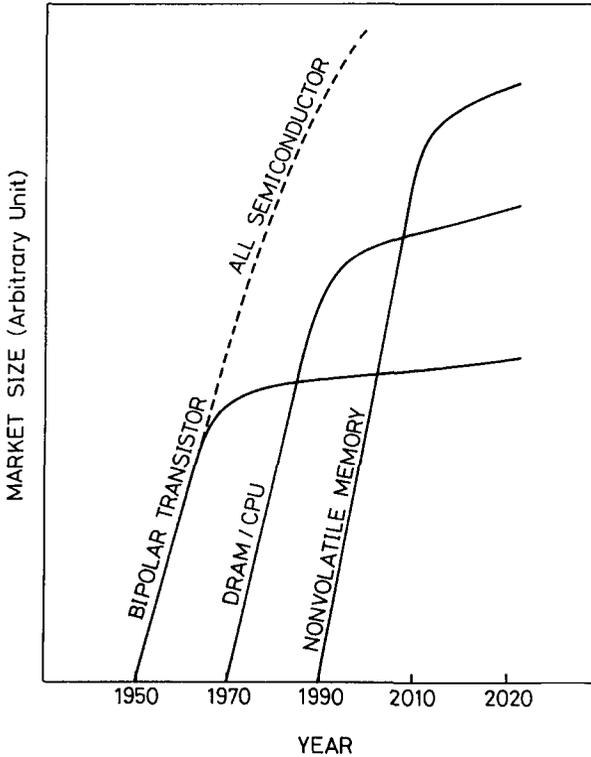


Figure 1.13 Growth curves for different technology drivers. (After Ref. 28.)

1.4 ORGANIZATION OF THE BOOK

The book is divided into three parts. Part 1, Chapters 2 through 4, considers device fundamentals. Chapter 2 treats recent technology trends of bipolar transistors with special emphasis on self-adjusted structures and silicon-based heterostructures. MOSFET physics, MOSFET structures, and limits of scaling are considered in Chapter 3. In addition, short-channel effects, transport properties, and various parasitic effects in MOSFET are discussed. Chapter 4 deals with various modeling and simulation techniques for ULSI devices. Both Monte Carlo simulation and the Boltzmann transport equation are considered for deep-submicrometer and nanometer devices.

Part 2, Chapters 5 through 8, deals with device building blocks and advanced device structures. The operational principles and recent performance of silicon-on-insulator (SOI) devices are presented in Chapter 5. SOI offers exceptional flexibility in device design and is particularly suited for low-voltage and low-power integrated circuits. Chapter 6 is concerned with the hot-carrier effect, and its consequences on device operations. We consider types of hot-carrier damage, methods of characterizing the damage, and device parameters that can improve

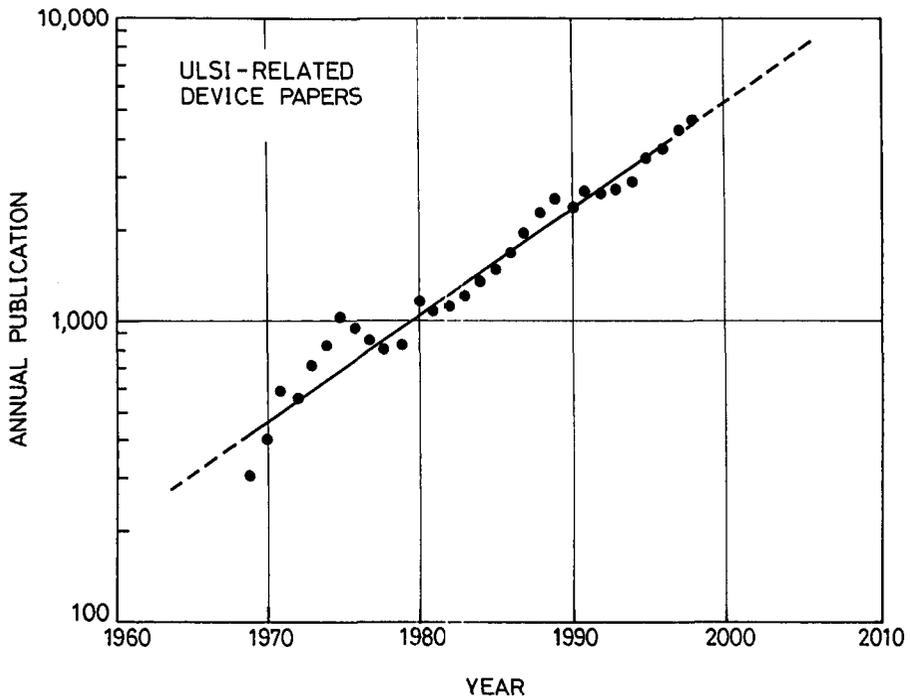


Figure 1.14 Publication of ULSI-related devices papers from 1969 to 1996 and projected to 2010. (After Ref. 30.)

hot-carrier hardness. Two of the most widely used semiconductor memories—DRAM and SRAM are considered in Chapter 7. The chapter includes memory-cell structures, operation principles, and related circuits. Chapter 8 focuses on reprogrammable nonvolatile memories including the floating-gate memory, silicon nitride memory, and ferroelectric memory.

Part 3, Chapters 9 through 11, treats circuit building blocks and the system-on-chip concept. The development of CMOS technology has expanded IC capability and density at an exponential rate. Chapter 9 considers CMOS digital and analog circuit building blocks, especially for modern, mixed-signal applications. In Chapter 10, CMOS design considerations for digital high-performance, low-voltage, low-power, and high-frequency operations are reviewed. The system-on-chip concept is presented in Chapter 11. The goal of system-on-chip is to integrate the complete functionality of regular systems into one single chip. In this chapter, we shall consider the embedded modules, mixed-signal technology, analog interface, and one-chip television IC.

At the present time, the ULSI-device field is moving at a rapid pace. The number of ULSI-related device publications has grown from about 300 papers in 1969 to over 4500 papers in 1998 with a 15-fold increase in 29 years (Fig. 1.14).³⁰ The annual growth rate of ULSI-related device papers is higher than the general semiconductor device field. In the early 1970s, ULSI-related device

papers constituted about 15% of the total semiconductor device literature. By year 2010, it is expected to increase to 30%! Note that many topics, such as the characteristics of nanoscale devices, the endurance of nonvolatile memory, and advanced systems-on-chip are still under intensive study. Their ultimate capabilities are not yet fully understood. The material presented in the book is intended to serve as a foundation. The references listed at the end of each chapter can supply more information.

REFERENCES

1. *1999 Electronic Market Data Book*, Electronic Industries Association, Washington, DC, 1999.
2. *1999 Semiconductor Industry Report*, Industrial Technology Research Institute, Hsinchu, 1999.
3. From Dataquest 1997.
4. J. Bardeen and W. H. Brattain, "The Transistor, a Semiconductor Triode," *Phys. Rev.* **71**, 230 (1948).
5. W. Shockley, "The Theory of p-n Junction in Semiconductors and p-n Junction Transistors," *Bell Syst. Tech. J.* **28**, 435 (1949).
6. H. Kroemer, "Theory of a Wide-Gap Emitter for Transistors," *Proc. IRE* **45**, 1535 (1957).
7. J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. Electron Devices* **ED-23**, 648 (1976).
8. R. N. Noyce. "Semiconductor Device-and-Lead Structure," U.S. Patent 2,981,877, filed July 30, 1959, granted April 25, 1961.
9. D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," IRE-AIEE Solid State Device Research Conf., Carnegie Institute of Technology, Pittsburgh, 1960.
10. F. M. Wanlass and C. T. Sah, "Nanowatt Logics Using Field-Effect Metal-Oxide Semiconductor Triodes," *Tech. Digest, IEEE Int. Solid-State Circuit Conf.*, 1963, p. 32.
11. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.* **46**, 1288 (1967).
12. K. Yano, T. Jshii, T. Hashimoto, T. Kabayashi, F. Murai, and K. Seki, "Room Temperature Single-Electron Memory," *IEEE Trans. Electron Devices* **41**, 1628 (1994).
13. R. H. Dennard, "Evolution of MOSFET Dynamic RAM—A Personal View," *IEEE Trans. Electron Devices* **ED-31**, 1549 (1984).
14. The inventors of the microprocessor are M. E. Hoff, F. Faggin, S. Mazor, and M. Shima. For a profile of M. E. Hoff, see R. Slater, *Portraits in Silicon*, MIT Press, Cambridge, MA, 1987, p. 175.
15. J. L. Moll and Y. Tarui, "A New Solid State Memory Resistor," *IEEE Trans. Electron Devices* **ED-10**, 338 (1963).
16. C. A. Mead, "Schottky Barrier Gate Field Effect Transistor," *Proc. IEEE* **54**, 307 (1966).
17. L. L. Chang, L. Esaki, and R. Tsu, "Resonant Tunneling in Semiconductor Double Barriers," *Appl. Phys. Lett.* **24**, 593 (1974).

18. T. Minura, S. Hiyamizu, T. Fujii, and K. Nanbu, "A New Field-Effect Transistor with Selectively Doped GaAs/n-Al_xGa_{1-x}As Heterojunction," *Jpn. J. Appl. Phys.* **19**, L 225 (1980).
19. S. Luryi, A. Kastalsky, A. G. Gossard, and R. H. Hended, "A Charge Injection Transistor Based on Real Space Hot-Electron Transfer," *IEEE Trans. Electron Devices* **ED-31**, 832 (1984).
20. S. Luryi, "An Induced Base Hot Electron Transistor," *IEEE Electron Devices Lett.* **EDL-6**, 178 (1985).
21. C. Y. Chang, W. C. Liu, M. S. Jame, Y. H. Wang, S. Luryi, and S. M. Sze, "Induced Base Transistor Fabricated by Molecular Beam Epitaxy," *IEEE Electron Devices Lett.* **EDL-7**, 491 (1986).
22. C. Y. Chang, "Structures of Modulation Doped Base Hot Electron Transistor," U.S. Patent 7,498,354, Dec. 15, 1992.
23. For example, see S. M. Sze, *Modern Semiconductor Device Physics*, Wiley-Interscience, New York, 1998.
24. *The National Technology Roadmap for Semiconductors* (NTRS), Semiconductor Industry Association (SIA), San Jose, CA, 1999.
25. G. Moore, "VLSI, What Does the Future Hold," *Electron. Aust.* **42**, 14 (1980).
26. C. Y. Chang and S. M. Sze, *ULSI Technology*, McGraw-Hill, New York, 1996.
27. A. W. Weider, "Si-Microelectronics: Perspectives, Risks, Opportunities, Challenges," in S. Luryi, J. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics*, Kluwer Academic Publishers, Boston, 1996.
28. F. Masuoka, "Flash Memory Technology," *Proc. Int. Electron Devices and Materials Symp.* Hsinchu, Taiwan, 1996, p. 83.
29. S. M. Sze, "Evaluation of Nonvolatile Semiconductor Memory—from Floating-Gate Concept to Single-Electron Memory Cell," in S. Luryi, J. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics*, Wiley-Interscience, New York, 1999.
30. From INSPEC database, National Chiao Tung University, 1999.

PART I

DEVICE FUNDAMENTALS

Bipolar Transistor Fundamentals

ERICH KASPER

University of Stuttgart
Stuttgart, Germany

2.1 INTRODUCTION

The first functioning transistor invented in 1947 by Shockley, Brattain, and Bardeen was based on the bipolar junction of a semiconductor with npn stacking. Later (1961) the first integrated circuit (IC) combined monolithic bipolar transistors and passive components on a chip. This event marked the start of the extraordinary successful industrial development of microelectronics. The realization of a rather old transistor principle—the field-effect transistor—came about with the metal oxide silicon (MOS) transistor in 1962. This transistor-type competed effectively with the bipolar junction transistor (BJT) and, after the development of the complementary MOS (CMOS) circuits, it nearly completely replaced the BJT transistor in digital logic circuits.

Are we therefore considering bipolar transistors mainly for historical reasons? The interest in BJT circuits is strongly continuing for three reasons:

1. In real systems the logic core is connected to the outside world by interface units that often need to handle power and/or analog signals. Because of their excellent power and analog signal capabilities, BJT circuits share a certain market segment in which high system performance is important.
2. The performance of BJT circuits especially considering speed or noise, was increased tremendously by sophisticated self-adjustment schemes based on the technology used in CMOS fabrication.
3. SiGe/Si heterojunctions will improve the properties of several devices. Because it is easy to incorporate, the SiGe technique will find the first applications in heterobipolar transistor (HBT) circuits.

In the following sections the basic principles of the BJT are explained first. Then the performance impact of modern self-adjustment technique are shown. Finally, the technical requirements and benefits of introducing SiGe into Si technology are discussed.

2.2 STRUCTURE AND OPERATING REGIMES

The BJT contains three adjoining, alternatively doped regions labeled the emitter (E), base (B), and collector (C). The middle region (base) is very small compared with the minority-carrier diffusion length for that region. The basic theory of bipolar transistors is well developed since the early work of Ebers and Moll in 1954, and Gummel and Poon in 1970, and it is described in many textbooks.¹⁻³ Figure 2.1 illustrates the basic structure and the circuit symbol for the npn transistor, which is the most common type. For the pnp transistor, the polarities of the currents and voltages change. In the pnp transistor symbol, the arrow is directed toward the base. Note that as a consequence of Kirchhoff's circuit laws there are only two independent voltages and two independent currents. In Figure 2.1 current reference directions were chosen to meet the IEEE standard notation ($\Sigma I = 0$), whereas in many textbooks the physical current flows in the active region are given. For clarification, we consider the polarities for all three possible circuit configuration with a common terminal between input and output as illustrated in Figure 2.2. Common base, common emitter, and common collector are the three possible amplifier types. Table 2.1 gives nomenclatures and polarities of the input and output terminals using the IEEE standard nomenclature. The most often used of the three amplifier types is the common emitter that can be used for current, voltage, and power amplification. The bipolar transistor has four regions of operation or dc bias (Fig. 2.3). The *active region* is defined as having the EB junction forward-biased and the CB junction reverse-biased. In the *saturation region* both the EB and CB junctions are forward-biased, whereas in the *cutoff region* both junctions are reverse-biased. In the fourth region of operation, is the *inverted active region*, the EB junction is reverse-biased and the CB junction is forward-biased. In the inverted active region the collector acts like an emitter and the emitter like a collector. The current gain in inverse operation is small because of unfavorable doping relations

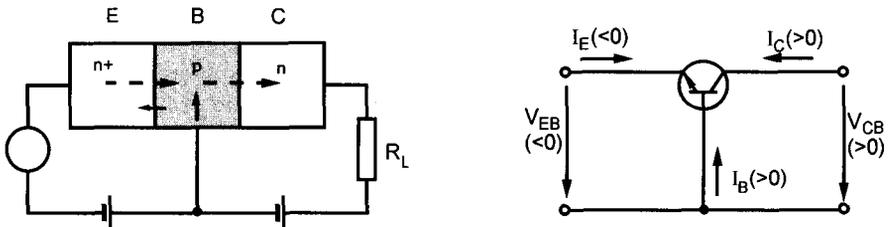


Figure 2.1 Semiconductor structure, symbols, and nomenclatures of the npn transistor.

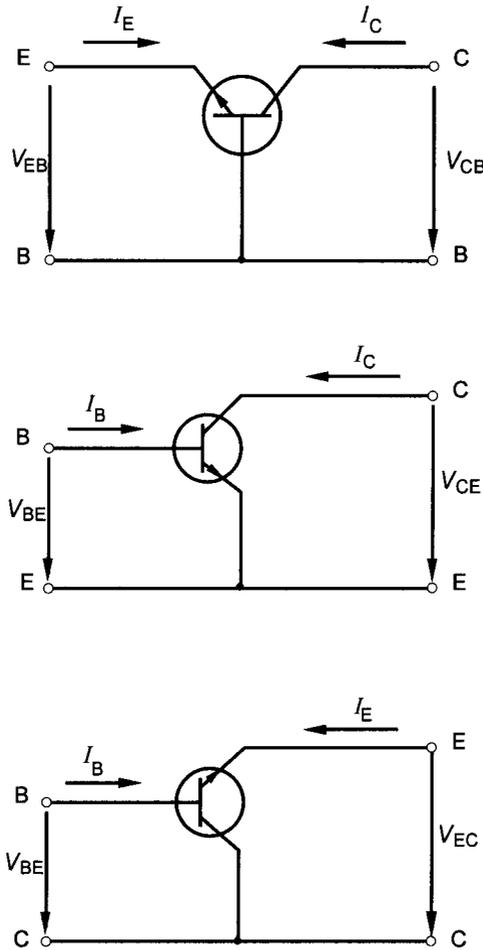


Figure 2.2 Common base, common emitter, and common collector configuration of an npn transistor.

TABLE 2.1 Input and Output Terminals for Common Base, Common Emitter, and Common Collector Configurations^a

Nomenclature	Symbol and Polarity			
	Input current	Input voltage	Output current	Output voltage
Common base	$I_E (-)$	$V_{EB} (-)$	$I_C (+)$	$V_{CB} (+)$
Common emitter	$I_B (+)$	$V_{BE} (+)$	$I_C (+)$	$V_{CE} (+)$
Common collector	$I_B (+)$	$V_{BC} (-)$	$I_E (-)$	$V_{EC} (-)$

^aSee Figure 2.2.

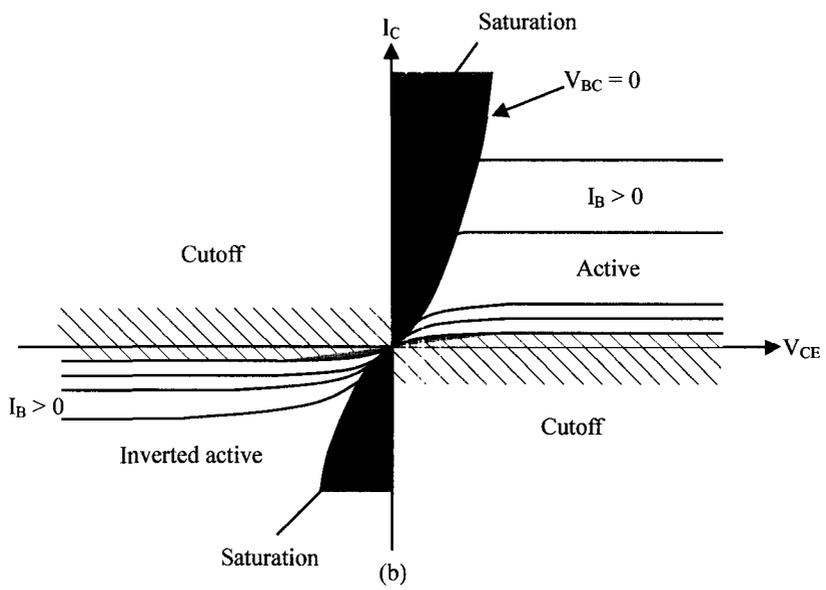
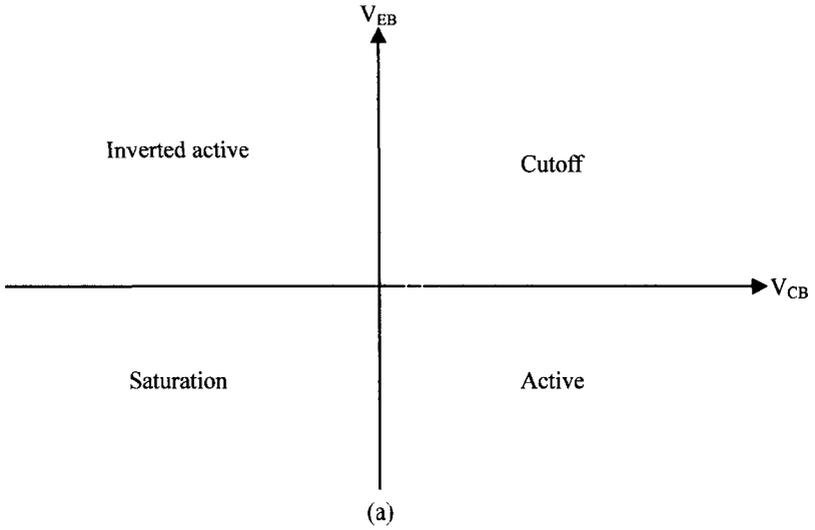


Figure 2.3 Regions of operation for the npn transistor; (a) voltage relations in a common base configuration; (b) output characteristics in a common emitter configuration. (After Neudeck, Ref. 2.)

and geometries. For logic applications the transistor is switched from the saturation region (ON state, high current with low bias voltage) to the cutoff region (OFF state, low current even with high bias voltage). For amplification, the transistor is operated in the active region.

The transistor characteristics are highly nonlinear, but for small-signal analysis the characteristics may be linearized around the static operation point. Consider the current I to be compared of a static part \bar{I} and a small-signal ac part i of frequency ω and phase delay ϕ , which can be written in the usual complex notation as

$$I = \bar{I} + i \exp(j\omega t) = \bar{I} + |i| \exp(j\phi) \exp(j\omega t) \quad (2.1)$$

$$(j = \sqrt{-1}), i = |i| \exp(j\phi) \quad (2.2)$$

Similarly, the voltage V may be separated to a dc part \bar{V} and a complex ac part v . In two-port circuit theory a three terminal-device such as the bipolar transistor is treated with one terminal common to input and output (Fig. 2.4). The small signal ac voltages and currents are connected by linear relations either using the admittance matrix (Eq. 2.3) or the hybrid matrix (Eq. 2.4).

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad (2.3)$$

The elements of the admittance matrix Y are called *input conductance* y_{11} , *inverse transconductance* y_{12} , *transconductance* y_{21} , and *output conductance* y_{22} . All admittance matrix parameters are measured under short-circuit condition, for example

$$y_{11} = \left. \frac{i_1}{v_1} \right|_{v_2=0}$$

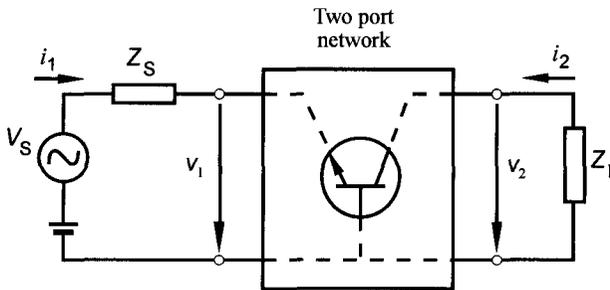


Figure 2.4 Two-port circuit presentation of a transistor in common base configuration with input voltage source V_S and source and load impedances Z_S, Z_L , respectively. The ac input and output voltages and currents are denoted as v_1, v_2 and i_1, i_2 , respectively.

Often a hybrid matrix H is used because the current gain is a parameter of this matrix:

$$\begin{bmatrix} v_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} i_1 \\ v_2 \end{bmatrix} \quad (2.4)$$

The hybrid matrix parameters are measured partly (h_{12}, h_{22}) under open-circuit conditions ($i_1 = 0$) and partly (h_{11}, h_{21}) under short-circuit conditions ($v_2 = 0$). The parameters are called *input resistance* h_{11} , *voltage feedback* h_{12} , *current gain* h_{21} , and *output conductance* h_{22} (under open-circuit conditions, $i_1 = 0$, compared to y_{22} , which is defined under short-circuit conditions, $v_1 = 0$). The parameters of both matrices obey the transformation relations

$$Y = \frac{1}{h_{11}} \begin{bmatrix} 1 & -h_{12} \\ h_{21} & \det H \end{bmatrix} \quad (2.5)$$

with the determinant $\det H$ defined as $h_{11}h_{22} - h_{12}h_{21}$. With $\det Y$ defined as $y_{11}y_{22} - y_{12}y_{21}$, the hybrid matrix H reads as

$$H = \frac{1}{y_{11}} \begin{bmatrix} 1 & -y_{12} \\ y_{21} & \det Y \end{bmatrix} \quad (2.6)$$

As a consequence of Kirchhoff's circuit laws (Eqs. 2.7 and 2.8) the matrix elements of the common emitter or common collector configuration are related to those of the common base configuration. In the following often the basic relations are given for the common base configuration and when needed transformed to the other configurations by Eqs. 2.9–2.12. Kirchhoff's law for bipolar transistors (ac components) is as follows:

$$i_B + i_E + i_C = 0 \quad (2.7)$$

$$V_{CE} + V_{EB} + V_{BC} = 0, \quad V_{BC} = -V_{CB} \quad (2.8)$$

For example, consider the Y matrix for common base and common emitter configurations. In common base the quantities i_1, i_2, v_1, v_2 are given by $i_E; i_C; V_{EB}; V_{CB}$, whereas in common emitter these quantities are expressed by $i_B = -i_E - i_C; i_C; V_{BE} = -V_{EB}; V_{CE} = V_{CB} - V_{EB}$. If we assign the element values for common base, emitter, and collector with the superscripts b, e , and c , respectively, then the relations for common emitter from common base are

$$\begin{bmatrix} y_{11}^e & y_{12}^e \\ y_{21}^e & y_{22}^e \end{bmatrix} = \begin{bmatrix} y_{11}^b + y_{12}^b + y_{21}^b + y_{22}^b & -(y_{12}^b + y_{22}^b) \\ -(y_{21}^b + y_{22}^b) & y_{22}^b \end{bmatrix} \quad (2.9)$$

and for common collector from common base

$$\begin{bmatrix} y_{11}^c & y_{12}^c \\ y_{21}^c & y_{22}^c \end{bmatrix} = \begin{bmatrix} y_{11}^b + y_{12}^b + y_{21}^b + y_{22}^b & -(y_{11}^b + y_{21}^b) \\ -(y_{11}^b + y_{12}^b) & y_{11}^b \end{bmatrix} \quad (2.10)$$

The same calculations can be done for the hybrid matrix H . We obtain for common emitter from common base

$$\begin{bmatrix} h_{11}^e & h_{12}^e \\ h_{21}^e & h_{22}^e \end{bmatrix} = \frac{1}{1 - h_{12}^b + h_{21}^b + \det H^b} \begin{bmatrix} h_{11}^b & \det H^b - h_{12}^b \\ -(\det H^b - h_{21}^b) & h_{22}^b \end{bmatrix} \quad (2.11)$$

and for common collector from common base

$$\begin{bmatrix} h_{11}^c & h_{12}^c \\ h_{21}^c & h_{22}^c \end{bmatrix} = \frac{1}{1 - h_{12}^b + h_{21}^b + \det H^b} \begin{bmatrix} h_{11}^b & 1 + h_{21}^b \\ h_{12}^b - 1 & h_{22}^b \end{bmatrix} \quad (2.12)$$

The frequency limit of modern bipolar transistors is in the microwave regime, even more and more reaching the millimeter-wave regime (>30 GHz). In this high-frequency regime, scattering parameters (s parameter) are extensively used because they are easier to measure at high frequencies. The S matrix is defined by Eq. 2.13 when the incident and reflected waves are denoted by (a_1, a_2) and (b_1, b_2) , respectively:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad (2.13)$$

Later we use the scattering matrix when we discuss the frequency limits of a transistor.

2.3 THE INNER-BOX-SHAPED TRANSISTOR

An idealistic picture of an integrated bipolar transistor is given in Figure 2.5. In this ideal picture the inner transistor between emitter contact and buried layer (subcollector) is box-shaped, connected on top to the poly-Si emitter contact, on both sides to the base, and at the bottom, to the collector via the buried layer subcollector. The box-shaped inner transistor, the depletion layers, the used coordinate system and the energy band diagram are shown in Figures 2.6 and 2.7. In the active region the EB junction is forward-biased. Therefore, and also because of the high doping levels, the depletion layer width is small and can be neglected for most geometric considerations. The depletion layer width l of a pn junction with doping N_A (p side) and N_D (n side) is given by

$$l = \left(\frac{2\varepsilon}{q} \right)^{1/2} \left(\frac{1}{N_D} + \frac{1}{N_A} \right)^{1/2} (V_{bi} - V_j)^{1/2} \quad (2.14)$$

where ε is the dielectric permittivity ($\varepsilon; = 10^{-10}$ (A · s)/(V · m) in Si), q is the electron charge ($1.6 \cdot 10^{-19}$ A · s), V_{bi} is the built-in voltage, and V_j is the voltage

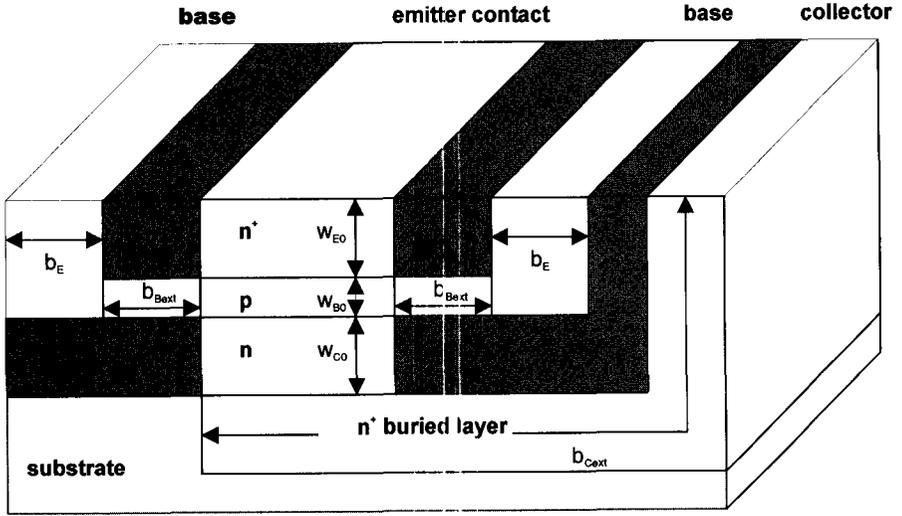


Figure 2.5 Scheme of an integrated bipolar transistor with poly-Si contacts and oxide isolation.

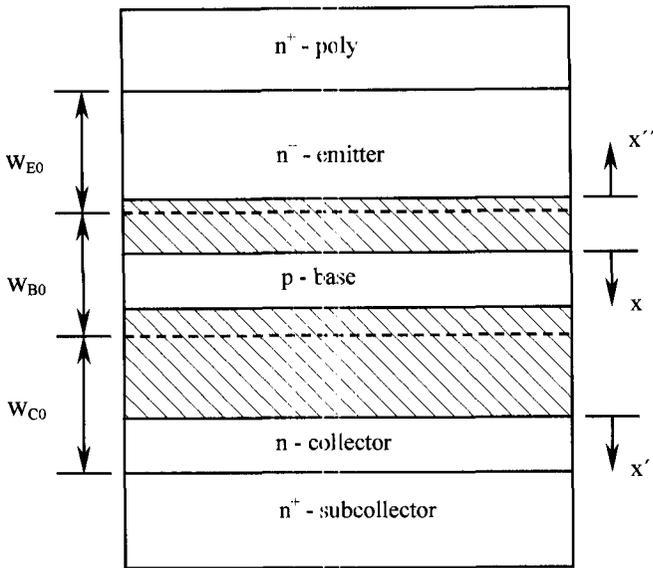


Figure 2.6 Scheme of the inner transistor and its depletion regions. Coordinate systems used are x axis in the base starting from the edge of the EB depletion layer, x' axis in the collector starting from the edge of the BC depletion layer, and x'' axis into the emitter starting from the edge of the BE depletion layer.

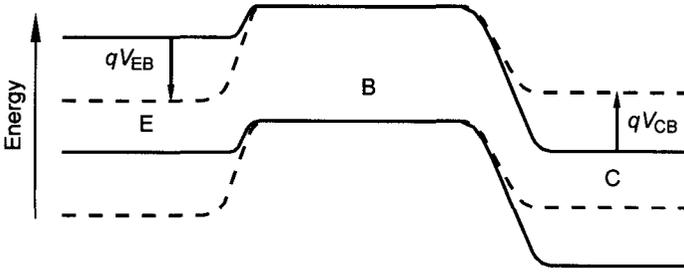


Figure 2.7 Band diagram of the transistor shown in Figure 2.6 (---- equilibrium, — active region).

across the junction. The built-in voltage can be calculated from

$$V_{bi} = V_T \ln \left(\frac{N_A \cdot N_D}{n_i^2} \right) \quad (2.15)$$

with thermal voltage $V_T = (kT/q)$ and intrinsic carrier density n_i . At room temperature the values for V_T , and n_i are 26 meV, and $1.45 \cdot 10^{10} \text{ cm}^{-3}$. The depletion layer extends mainly into the lower doped region

$$l_n \cdot N_D = l_p \cdot N_A, \quad l = l_n + l_p \quad (2.16)$$

with l_n, l_p , the depletion layer widths extending into the n side and the p side, respectively. The BC depletion layer extends mainly into the collector because of its lower doping, but the base extension has to be considered for the Early effect (output resistance). In a well-designed transistor the depletion width of the BC junction is roughly the collector width w_{c0} ; otherwise either the collector series resistance increases ($l < w_{c0}$) or the collector junction capacity C_{jc} is higher than necessary. The right collector doping N_D can be calculated for a given width w_{c0} and a collector-base voltage $|V_{CB}|$ by inserting into Eq. 2.14 $l = w_{c0}, (1/N_A) \rightarrow 0, V_j = -|V_{CB}|$. Table 2.2 shows several width and voltage values and the corresponding doping values at which *punchthrough* occurs, meaning that the collector epilayer (w_{c0}) is fully depleted. With decreasing collector width, the breakdown voltage BV_{CB0} is reduced, the collector transit time is reduced (interesting for ultra-high-speed transistors), and the collector junction capacity C_{jc} is increased.

Let us now consider the minority carrier concentration at equilibrium or with applied voltages V_{EB} and V_{CB} (common base). The coordinate axis of Figure 2.6 and the band diagram of Figure 2.7 will be used. In a widely used approximation the minority-carrier concentration at the edge of the depletion region is dependent only on the equilibrium concentration and the applied voltage V_j (positive polarity at forward bias):

$$n(x=0) = n_0 \exp \left(\frac{V_j}{V_T} \right), \quad n_0 = \frac{n_i^2}{N_A} \quad (2.17)$$

TABLE 2.2 Selected Collector Widths w_{c0} and Voltages $V_{bi} + |V_{CB}|$ and Their Corresponding Doping Values N_{DC} for Punchthrough and Breakdown Voltages BV_{CB0}

w_{c0} (nm)	$V_{bi} + V_{CB} $ (V)	N_{DC} (cm^{-3})	BV_{CB0} (V)
1460	5	$3 \cdot 10^{15}$	46
620	4	$1.25 \cdot 10^{16}$	22
345	3	$3.1 \cdot 10^{16}$	14
140	2	$1.25 \cdot 10^{17}$	7.5
65	1	$3.1 \cdot 10^{17}$	3.9

This approximation is exactly valid with vanishing currents and minority-carrier levels much lower than the majority-carrier level. At high current injection the approximation becomes invalid. Further, we assume a neutral base width w_B much smaller than the corresponding electron diffusion length L_B in the base whereas conveniently the emitter width w_E is assumed to be much larger than the corresponding diffusion length L_E . The diffusion length L is connected to the carrier mobility μ and lifetime τ by

$$L^2 = V_T \mu \tau \tag{2.18}$$

For low doped material ($< 10^{17} \text{ cm}^{-3}$) the lifetime τ is dominated by recombination centers that lead, typically, to lifetimes from 100 μs to many ms and diffusion lengths in the millimeter range. Above a doping level of 10^{18} cm^{-3} Auger recombination dominates with

$$1/\tau = G \cdot n^2 \text{ (Auger coefficient } G \sim 10^{-31} \text{ cm}^6/\text{s}) \tag{2.19}$$

Auger recombination reduces the lifetime τ to 10 μs (10^{18} cm^{-3}), to 100 ns (10^{19} cm^{-3}), and 1 ns (10^{20} cm^{-3}). The diffusion lengths (holes) L_p are reduced to 65 μm (10^{18} cm^{-3}), to 4.5 μm (10^{19} cm^{-3}), and 0.35 μm (10^{20} cm^{-3}). In modern transistor designs these assumptions ($w_E \gg L_E, W_C \gg L_C$) will seldom be fulfilled. Fortunately, the resulting currents depend on a length that will equal L for $w \gg L$ and equal w for $w \ll L$. By keeping these equalities in mind, the standard solutions can be used.

The electron distribution in the p base is given as linear function of x . The exact solution contains hyperbolic functions that may be linearized for small arguments ξ :

$$\begin{aligned} \sinh \xi &= \frac{1}{2} (\exp(\xi) - \exp(-\xi)) = \xi + \frac{\xi^3}{6} + \dots \\ \cosh \xi &= \frac{1}{2} (\exp(\xi) + \exp(-\xi)) = 1 + \frac{1}{2} \xi^2 + \dots \\ \xi \coth \xi &= \xi \frac{\cosh \xi}{\sinh \xi} = 1 + \frac{1}{3} \xi^2 + \dots +, \xi \ll 1 \end{aligned} \tag{2.20}$$

The equilibrium electron density n_0 is given by (n_i^2/N_B) (Eq. 2.17; N_B base doping), and the linearized electron concentration in the base is given by

$$n(x) - n_0 = n_0 \left[\exp\left(-\frac{V_{EB}}{V_T}\right) - 1 \right] - n_0 \frac{x}{w_B} \left[\exp\left(-\frac{V_{EB}}{V_T}\right) - \exp\left(-\frac{V_{CB}}{V_T}\right) \right] \quad (\text{neutral base}) \quad (2.21)$$

On the emitter and collector sides the hole density decreases exponentially from their values at the edges of the EB and BC junctions:

$$p(x'') - p_{E0} = p_{E0} \left[\exp\left(-\frac{V_{EB}}{V_T}\right) - 1 \right] \exp\left(-\frac{x''}{L_E}\right) \quad (\text{neutral emitter}) \quad (2.22)$$

with $p_{E0} = n_i^2/N_E$ (N_E emitter doping).

$$p(x') - p_{C0} = p_{C0} \left[\exp\left(-\frac{V_{CB}}{V_T}\right) - 1 \right] \exp\left(-\frac{x'}{L_C}\right) \quad (\text{neutral collector}) \quad (2.23)$$

with $p_{C0} = n_i^2/N_C$ (N_C collector doping).

When a device is operated at punchthrough conditions (collector depleted), choose the subcollector doping as N_C .

For technological reasons and to improve the transistor speed, the base doping profile monotonically decreases from the emitter edge to the base edge. A transistor with such a doping profile is called a *drift transistor*, since a built-in electric field enhances the electron drift in the base. The built-in field E is given by

$$E = V_T \frac{d(\ln N_B)}{dx} \quad (2.24)$$

The steady-state solution for the active region ($p(x = w_B) = 0$) is given by

$$n(x) = \frac{J_n}{q \cdot V_T \cdot \mu_n} \cdot \frac{1}{N_B(x)} \int_x^{w_B} N_B(x) dx \quad (2.25)$$

where J_n is the minority-carrier current through the base.

The principal distribution of minority carriers in the neutral regions of the transistor is shown in Figure 2.8.

2.3.1 Static Characteristics

Device modeling requires information about the electrical potential, carrier generation or recombination, and carrier transport. This is usually given by a

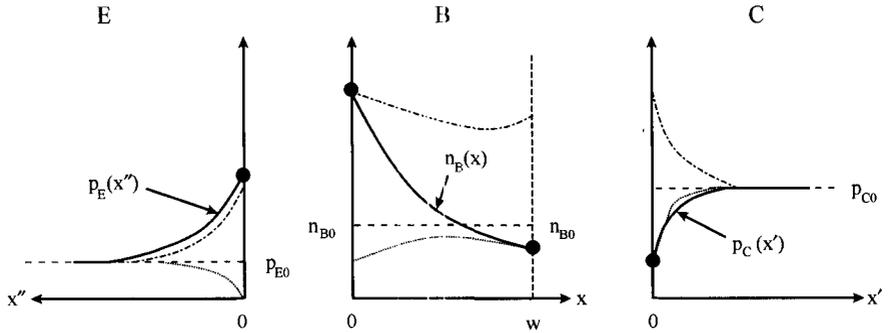


Figure 2.8 Minority-carrier distribution in an npn transistor (not to scale) for the active region (—), the saturation region (- - - - -), and the cutoff region (· · · · ·). (After Neudeck, Ref. 2.)

coupled set of five partial-differential equations (Eqs. 2.26–2.30). In the absence of magnetic fields and with single device dimensions much smaller than the utilized electromagnetic wavelengths, the Poisson equation (Eq. 2.26) describes the electric potential Φ

$$\Delta\Phi = -\frac{\rho}{\epsilon} \tag{2.26}$$

(Laplace operator $\Delta \equiv \text{div} \cdot \text{grad} \equiv d^2/dx^2$ for one-dimensional problems).

The electric charge density ρ is given by the sum of mobile charge densities (p, n) and fixed charge densities (N_A^-, N_D^+):

$$\rho = q(N_D^+ + p - N_A^- - n) \tag{2.27}$$

At room temperature, practically complete ionization of the dopant atoms can be assumed ($N_A^- = N_A, N_D^+ = N_D$). At lower temperatures, partial ionization has to be considered using Fermi statistics. But for higher dopings ($> 5 \cdot 10^{17}/\text{cm}^3$) bandgap shrinkage and impurity band conduction have to be included to avoid significantly wrong results.

The conservation of particles is described by continuity equations, which for holes and electrons are

$$\begin{aligned} \frac{\partial p}{\partial t} + \frac{1}{q} \text{div} J_p &= G_p - R_p \\ \frac{\partial n}{\partial t} - \frac{1}{q} \text{div} J_n &= G_n - R_n \end{aligned} \tag{2.28}$$

The right side of Eq. 2.28 equals zero when particle numbers are conserved. Generation and recombination are described by the right side terms $G-R$

$$\begin{aligned} G_p - R_p &= \frac{p(x, t) - p_0}{\tau_p} \\ G_n - R_n &= \frac{n(x, t) - n_0}{\tau_n} \end{aligned} \quad (2.29)$$

This equation is a valid approximation for low minority-carrier densities and in the absence of light absorption or particle impact.

The mobile-carrier transport is usually divided into diffusive transport and an electric field transport (drift-diffusion model).

$$\begin{aligned} J_p &= qp\mu_p E - qV_T\mu_p \cdot \text{grad } p \\ J_n &= qn\mu_n E + qV_T\mu_n \cdot \text{grad } n \end{aligned} \quad (2.30)$$

with mobility μ , the diffusion coefficient was already written as $V_T\mu$ (Einstein relation) and the electric field strength $E \equiv -\text{grad } \Phi \equiv d\Phi/dx$ for a one-dimensional problem. For very short dimensions ($< 0.1 \mu\text{m}$), the carriers can obtain velocities higher than the saturation velocity, a phenomenon called *velocity overshoot*.⁴ This is treated in the hydrodynamic model⁵ or directly with Monte Carlo simulations.^{6,7}

The coupled differential equations are solved numerically for appropriate boundary conditions at the device terminals.⁸⁻¹⁵ For circuit simulations, simpler compact models are used, including a modified Gummel-Poon model for the circuit simulator SPICE (Simulation Program For IC Emphasis) (SGP)¹⁶ and a model devoted to high current densities (HICUM)¹⁷. Other compact models are described in Refs. 18 and 19. A recent comparison of the benefits of some models especially for radiofrequency (RF) circuit design is given in Ref. 20.

For discussion of the principal effects, we consider mainly the active region of the transistor. The pn junctions can be treated as diodes and the minority currents passing the base as current sources. When the base is thin ($w \ll L$) and the emitter doping is high, the current gain α_0 in common base configuration approaches unity because the base current composed of holes back injected into the emitter and of holes recombined with injected electrons vanishes compared to the injected electron current:

$$\alpha_0 = \alpha_E \alpha_T = \frac{I_C}{-I_E} = \frac{-I_E - I_B}{-I_E} = 1 - \frac{I_B}{|I_E|} \quad (2.31)$$

Both components, the emitter injection efficiency α_E and the transport factor α_T , are only somewhat smaller than unity. For the minority-carrier distribution shown in Figure 2.8, the emitter injection efficiency describing the relation between injected electrons and back-injected holes is given by

$$\alpha_E = 1 - \frac{\mu_p w_B N_B}{\mu_n L_p N_E} \quad (2.32)$$

where μ and L are the mobilities and diffusion lengths of electrons (n) and holes (p) in the base and emitter, respectively.

The transport factor α_T describing the recombination losses is given by

$$\alpha_T = 1 - \frac{1}{2} \left(\frac{w_B}{L_n} \right)^2 \tag{2.33}$$

As demonstrated by Eqs. 2.32 and 2.33, the essential requirements for obtaining a near-unity current gain α_0 demand large N_E/N_B and L_n/w_B ratios. The current gain β_0 in common emitter configuration is related to α_0 by

$$\beta_0 = \frac{I_C}{|I_B|} = \frac{\alpha_0}{1 - \alpha_0} \tag{2.34}$$

In reality the current gain may also depend on current (which means on V_{EB}) and on output voltage V_{CB} . The cause for the current dependence is best seen in the Gummel plot (Fig. 2.9) where the currents ($\ln I_B$, $\ln I_C$) are plotted against the emitter–base terminal voltage V_{EB} . Deviations from the ideal behavior are shown at both very low and very high currents. At low current densities the deviation is caused by the additional contribution to the base current from recombination in the EB depletion layer. At high current densities, temperature effects must be avoided by careful pulsed measurements. The remaining effects can be assigned to the influence of series resistances, high current injection, or current crowding. Series resistances

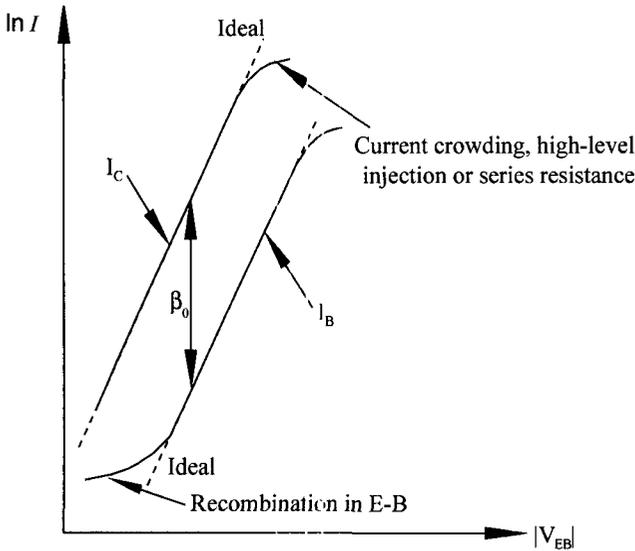


Figure 2.9 Gummel plot. Collector current ($\ln I_C$) and base current ($\ln I_B$) versus emitter–base voltage V_{EB} ($V_{CB} = 0$).

modify the internal voltages V'_{EB} , V'_{CB} compared to the terminal voltages V_{EB} , V_{CB} with polarities in active region, $V_{EB}(-)$, $V_{CB}(+)$:

$$\begin{aligned}\Delta V_{EB} &= V'_{EB} - V_{EB} = I_E r_E + I_B r_B \cong I_C \left(r_E + \frac{r_B}{\beta_0} \right) \\ \Delta V_{CB} &= V'_{CB} - V_{CB} = I_B r_B - I_C r_C \cong I_C \left(-r_C + \frac{r_B}{\beta_0} \right)\end{aligned}\quad (2.35)$$

Sometimes it may be difficult to find the series resistance correction because of high-level injection occurring at high current densities. At a certain current density level the assumption of low-level injection will be violated. Extrapolation yields voltage V_{EB}^* and current J_C^* at which the minority carriers n ($x = 0$) equal the doping level N_A :

$$\begin{aligned}|V_{EB}^*| &= 2V_T \ln \left(\frac{N_A}{n_i} \right) \\ J_C^* &= qV_T \frac{\mu_n}{w_B} N_B\end{aligned}\quad (2.36)$$

The current, which ideally follows an exponential increase

$$J_C = qV_T \frac{\mu_n n_i^2}{w_B N_B} \exp \left(\frac{V}{V_T} \right), \quad (V = -V_{EB}; \text{ see Table 2.1}) \quad (2.37)$$

will then increase with an ideality factor $\eta \rightarrow 2$ ($J_C \sim \exp(V/\eta V_T)$).

At a critical current density J_{crit} , the injected carriers compensate for the ionized donors in the BC depletion layer, which will cause an extension of the high field region and later at J_{Kirk} , a pushout of the neutral base in the collector region (Kirk effect):

$$\begin{aligned}J_{\text{crit}} &= qv_s N_C \\ J_{\text{Kirk}} &= J_{\text{crit}} + \frac{2v_s \epsilon (V_{CB} + V_{bi})}{w_C^2}\end{aligned}\quad (2.38)$$

where v_s is the saturation velocity. The base pushout or Kirk effect results in a decrease in current gain and speed.

To evaluate the dependence of current gain on collector–base voltage, we record the output characteristics (Fig. 2.10), which shows the current source behaviour. In a common base configuration (Fig. 2.10a) the output current I_C is mainly determined by the parameter input current I_E ($I_C = \alpha_0 I_E$). The close proportionality ($\alpha_0 \cong 1$) extends from the saturation region ($I_{CB} \ll 0$) to the avalanche breakdown of the BC junction the upper limit of which is characterized by the breakdown voltage BV_{CB0} .

In a common emitter configuration (Fig. 2.10b), as a consequence of the higher gain, the finite output conductance that is caused by the modulation of the neutral

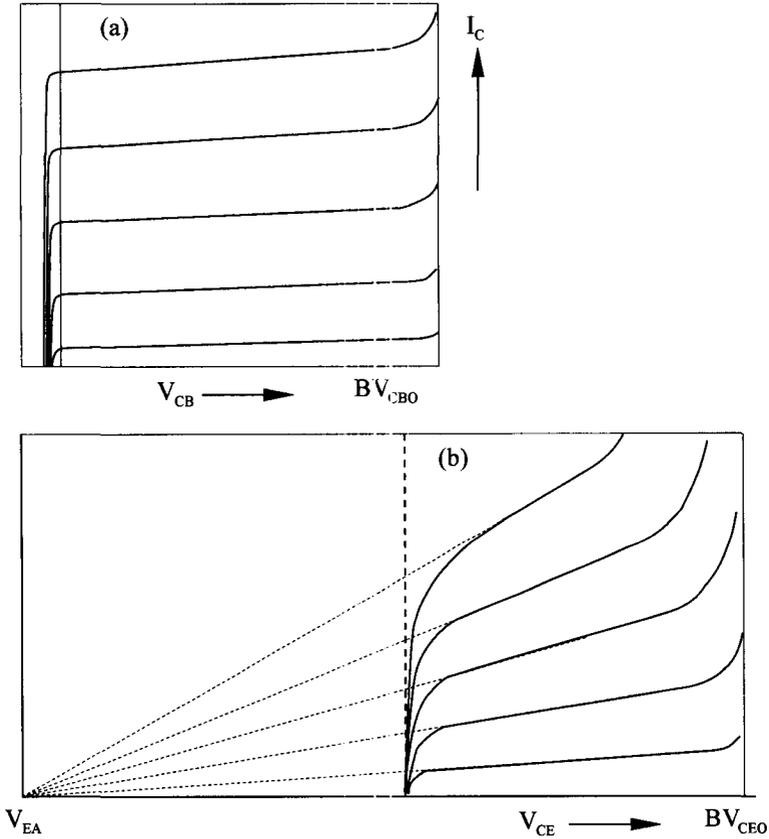


Figure 2.10 Output characteristics for a npn transistor: (a) common base configuration I_C versus V_{CB} ; (b) common emitter configuration I_C versus V_{CE} ; $BV_{CB0} > BV_{CE0}$.

base width w_B (Early effect) is easier to see. The Early voltage V_{Ea} at which the current trajectories meet the abscissa (Fig. 2.10b) is given by

$$V_{Ea} = \frac{qN_B w_B^2}{\epsilon} \quad (2.39)$$

As this equation 2.39 shows, low Early voltages V_{Ea} are a problem of high-frequency transistors that need a small base width w_B .

The most important problem connected with small base widths is caused by the voltage drop from the outer edge to the center of the base. For wide emitter fingers, current crowding at the edges of the base results. The base sheet resistance R_{SB} is given by

$$R_{SB} = (q\mu_B N_B w_B)^{-1} \quad (2.40)$$

The inner base (without its external connections) can be treated as a lateral base resistance connected to the EB junction. The edges become more forward-biased, leading to high current densities. Let the emitter length l_E be larger than the emitter width b_E , and let us assume base contacts on both sides of the emitter stripe. For base currents small enough to hold the voltage drop below V_T , a simple expression is obtained for the inner base resistance $^{21}r_{bi}$:

$$r_{bi} = \frac{1}{12} R_{SB} \frac{b_E}{l_E} \quad (2.41)$$

For larger voltage drops, correction functions are given in Refs. 22 and 23. For higher current densities, the reduction of the base sheet resistance by injected carriers has to be considered.

The inner base resistance limits the emitter stripe width w_E to avoid current crowding. From the condition $r_{bi} I_B \ll V_T$ we obtain

$$V_T \gg r_{bi} \frac{J_C}{\beta_0} b_E l_E \quad (2.42)$$

with J_C the collector current density. Combination with Eq. 2.41 yields as a condition for insignificant current crowding

$$b_E^2 \ll \frac{12\beta_0 V_T}{R_{SB} J_C} \quad (2.43)$$

2.3.2 High-Frequency Behavior

The high-frequency response of a device may be modeled by a device simulation using the time-dependent semiconductor equations. However, for applications in simple circuits, this exact approach is too time-consuming. Compact models based on equivalent circuits are used for circuit analysis instead. Practically, there is also a strong need to characterize the speed of a device without reference to a specific circuit. The transit frequency f_T describing the frequency limits of current gain and the maximum oscillation frequency f_{max} describing power–gain frequency limits are generally used. In the following paragraphs we give an example for an exact modeling under simplified boundary conditions, present equivalent circuit models, and discuss the factors that influence the frequency limits.

For an exact time-dependent solution of the minority carrier distribution in the base, the carrier density $n(x,t)$ is assumed to consist of a stationary part $\bar{n}(x)$ and a frequency (ω)-dependent part $\hat{n}(x,t)$:

$$n(x,t) = \bar{n}(x) + \hat{n}(x,t) = \bar{n}(x) + \hat{n}(x)\exp(j\omega t) \quad (2.44)$$

with the boundary conditions $L_n \gg w_B$, $n(x = w_B) = 0$, and $\bar{n}(x = 0) = n_0 \exp(V_{EB}/V_T)$. Solutions of Eq. 2.44 are of the form

$$\hat{n}(x) = \frac{\text{const.}}{\sinh\left(\frac{w'}{\tilde{L}_n}\right)} \sinh\left(\frac{w-x}{\tilde{L}_n}\right) \quad (2.45)$$

with a complex diffusion length

$$\tilde{L}_n = \frac{L_n}{(1 + j\omega\tau_n)^{1/2}} \quad (2.46)$$

The conductance for small signals is given by

$$\frac{i_E}{v_{EB}} = \left(1 + j\omega \frac{w_B^2}{3V_T\mu_n}\right) g_{ed} \quad (2.47)$$

when using Eq. 2.20. The small-signal conductance of the EB junction is determined by

$$g_{ed} = \frac{I_C}{V_T} \quad (2.48)$$

where the indices *ed* shows that the current is mainly driven by diffusing carriers injected from the emitter. The result (Eq. 2.47) demonstrates that the input current i_E gets out of phase from the input voltage v_{EB} with increasing frequency $\omega = 2\pi f$. The phase difference amounts to $\omega\tau_B$, where the base charging time is more generally written as

$$\tau_B = \frac{w_B^2}{2V_T\mu_n K_{bi}} \quad (2.49)$$

with K_{bi} a constant describing the charging time reduction by the built-in field E_{bi} in drift transistors

$$K_{bi} = 1 + \left(\frac{E_{bi}w_B}{V_T}\right)^{3/2} \quad (2.50)$$

As shown by Eqs. 2.49 and 2.50, the built-in field can significantly reduce the time constant τ_B . However, with decreasing base width w_B the built-in field potential for τ_B reduction shrinks (Eq. 2.50). In Eq. 2.49 the number $\frac{1}{2}$ was used as numerical prefactor instead of the number $\frac{1}{3}$ as suggested by the exact Eq. 2.47. The commonly used numerical prefactor $\frac{1}{2}$ stems from quasistatic charge control models, which

probably overestimate the base delay. With the same procedure used for the emitter current i_E (Eqs. 2.45–2.47), the collector current can be calculated from Eq. 2.45. The collector current $i_C = qV_T\mu_n(dn/dx)$ (where $x = w_B$) at the BC junction follows without phase delay the input voltage v_{EB} . The results can also be interpreted by an input consisting of a parallel conductance g_{ed} and a capacity $C_{ed} = g_{ed}\tau_B$ and an output current source $\alpha_0 g_{ed} v_{BE}$ that has to meet the relation $i_C = \alpha_0 i_E$ for low frequencies:

$$C_{ed} = g_{ed}\tau_B \quad (2.51)$$

For the common base configuration the equivalent circuit shown in Figure 2.11 results. The Early effect is modeled by a conductance ηg_{ed} between collector and base, where η is the voltage feedback from output v_{CB} to input v_{EB} ($\eta \ll 1$).

$$\eta = -V_T \frac{1}{w_B} \frac{dw_B}{dV_{CB}} \quad (2.52)$$

The output current source $\alpha_0 g_{ed} v_{BE}$ is in phase with the input voltage v_{EB} . This is valid if we assume that the BC junction is a perfect drain of the diffusing minority carriers. In reality, the electron speed in the BC junction is limited to v_s , the saturation velocity. The corresponding delay is the collector depletion layer transit time τ_C

$$\tau_C = \frac{w_{C0} - w_C}{2v_s} \quad (2.53)$$

with $w_{C0} - w_C$ the width of the BC depletion layer. This leads to an additional phase delay $\exp(j\omega\tau_C)$ for the current source, which for simplicity is not shown in Figure 2.11. The frequency-dependent current gain α in common base configuration is then given by

$$\alpha(\omega) = \frac{\alpha_0}{1 + j\frac{\omega}{\omega_x}} \quad (2.54)$$

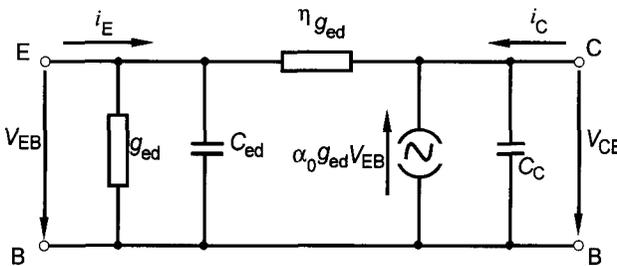


Figure 2.11 Common base equivalent circuit with an input diode represented by parallel conductance g_{ed} , capacity C_{ed} , output current source $\alpha_0 g_{ed} v_{EB}$, and collector depletion layer capacity C_C . Voltage feedback is described by the small conductance ηg_{ed} ($\eta \ll 1$).

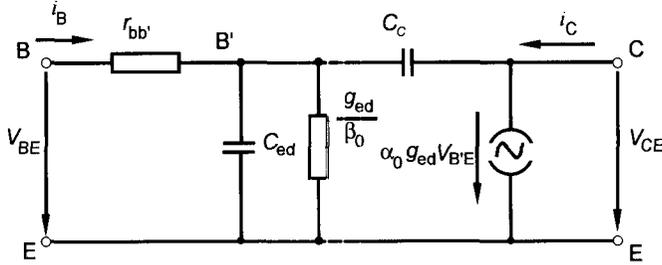


Figure 2.12 Simplified high-frequency, common emitter equivalent circuit of the bipolar transistor.

with the alpha cutoff frequency f_α

$$\omega_\alpha = 2\pi f_\alpha = \frac{1}{\tau_B} \tag{2.55}$$

at which the gain $|\alpha|$ has fallen to $1/\sqrt{2}$ of its low-frequency value. For the common emitter configuration, a simplified high-frequency equivalent circuit is shown in Figure 2.12. The situation is somewhat more complicated because the collector depletion layer capacity couples the output C with the input B . The current gain β in a common emitter configuration is therefore current-level-dependent by the load delay τ_g :

$$\tau_g = \frac{C_C + C_E}{g_{ed}} = \frac{V_T}{I_C} (C_C + C_E) \tag{2.56}$$

At high frequencies, the current gain $|\beta|$ rolls off one decade (20 dB) per decade frequency increase

$$|\beta| \cong \frac{1}{\omega \tau_{ec}} \tag{2.57}$$

with an emitter–collector delay τ_{ec} given by the sum of the individual contribution (Eqs. 2.49, 2.53, 2.56, 2.59)

$$\tau_{ec} = \frac{1}{\omega_T} = \frac{1}{2\pi f_T} = \tau_g + \tau_B + \tau_C + \tau'_C \tag{2.58}$$

and with an often neglected fourth term—the charging of the collector space charge C_C when a collector series resistance r_C —is given as

$$\tau'_C = C_C r_C \tag{2.59}$$

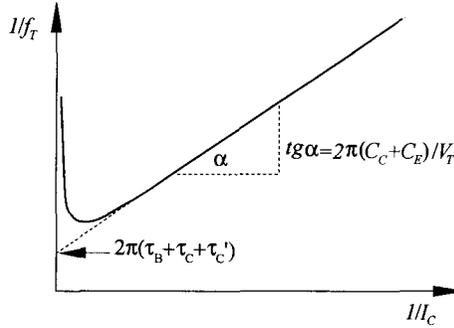


Figure 2.13 Current dependence of the transit frequency f_T shown as $1/f_T = 2\pi\tau_{ec}$ versus $(1/I_C)$.

In Eq. 2.58 we have neglected the injection of minority carriers in the emitter (τ_E) and also neglected the emitter depletion layer capacity (C_E) compared to the diffusion capacity C_{ed} , which is larger at high current injection. The current gain decreases from its low-frequency value β_0 to $\beta_0/\sqrt{2}$ at the β cutoff frequency f_β :

$$\frac{1}{\omega_\beta} = \frac{1}{2\pi f_\beta} = \frac{\beta_0}{\omega_T} \tag{2.60}$$

Another frequency number—the maximum oscillation frequency— f_{max} characterizes the power–gain frequency limits. The unilateral gain U in a feedback amplifier approaches unity at f_{max} .

$$f_{max} = \left(\frac{f_T}{8\pi r_{bb} C_C} \right)^{1/2} \tag{2.61}$$

The maximum oscillation frequency f_{max} increases with decreasing emitter strip width b_E because both the base resistance r_{bb} and the collector capacity C_C decrease. The current dependence of f_T is often represented by a $(1/f_T)$ – $(1/I_C)$ plot shown in Figure 2.13.

2.3.3 Thermal Effects

Temperature Dependence of the Collector Current

The collector current of bipolar junction transistors can be expressed under low current injection as

$$J_c(T) = \frac{qn_i^2(T)D_b(T)}{\int_0^w N(x)dx} \exp\left(\frac{q|V_{EB}|}{kT}\right) \tag{2.62}$$

where $n_i(T)$ is the effective intrinsic density and $D_b(T)$ is the diffusion constant in the base region. Starting with the temperature dependence of the minority-carrier

diffusion constant in the base region of an npn transistor, the Einstein relation $D_b(T) = (kT/q)\mu_b(T)$ is valid.

For low doping concentrations and low electric fields the mobility is affected by the effects of charge scattering with acoustic and optical phonons, resulting in the relation

$$\mu_b = \mu_{\max} \left(\frac{T}{300\text{K}} \right)^\zeta \tag{2.63}$$

Experiments suggest a temperature coefficient of $\zeta = -2.5$ for electrons and $\zeta = -2.2$ for holes.²⁴ If scattering at impurities is taken into account, Eq. 2.63 is extended to

$$\mu_b = \mu_0 + \frac{\mu_{b\max}(T) - \mu_0}{1 + \left(\frac{N}{C_r} \right)^\alpha} - \frac{\mu_1}{1 + \left(\frac{C_s}{N} \right)^\beta} \tag{2.64}$$

where $\mu_{b\max}(T)$ follows the description given by Eq. 2.63. Table 2.3 lists suggested parameters for the temperature range 200 K < T < 460 K.

Although the diffusion constant principally has a positive temperature coefficient resulting from the Einstein relation, its coefficient is dominated by the negative temperature coefficient of the carrier mobility.

For high doping levels, the mobility is dominated by impurity scattering with a temperature coefficient $\zeta = +1.5$ (Eq. 2.63). For medium doping levels, a temperature-independent mobility is an acceptable approximation.

Carrier-carrier scattering and velocity saturation in the base region are not factored into this calculation. The basic relation for calculating the intrinsic density is

$$n_i^2(T) = N_c(T)N_v(T)\exp\left(-\frac{E_g}{kT} + \frac{\Delta E_g}{kT}\right) \tag{2.65}$$

where $N_c(T)$ and $N_v(T)$ are the effective densities of band of the conduction and the valence band, respectively; $E_g(T)$ is the bandgap; and ΔE_g is the bandgap narrowing due to high impurity concentrations. The effective densities of the band state result from an integration of the product of the density of state functions and the Fermi

TABLE 2.3 Parameters for Mobility Calculations (Eq. 2.64)

	μ_0 [cm ² /(V · s)]	μ_{\max} [cm ² /(V · s)]	μ_1 [cm ² /(V · s)]	ζ	C_r (cm ⁻³)	C_s (cm ⁻³)	α	β
Electrons	52.2	1417	43.4	-2.5	$9.68 \cdot 10^{16}$	$3.43 \cdot 10^{20}$	0.68	2.00
Holes	44.9	470.5	29.0	-2.2	$2.23 \cdot 10^{17}$	$6.10 \cdot 10^{20}$	0.719	2.00

distribution functions, which are, in the case of nondegeneracy, approximated by Maxwell–Boltzmann functions. Therefore, the integration over the respective energy levels results in

$$N_{c,v}(T) = \frac{1}{4} \left(\frac{2m_{e,h}^* kT}{\pi \hbar^2} \right)^{3/2} \quad (2.66)$$

The density of state effective masses m_e^* and m_h^* are also temperature-dependent. This dependence can be neglected, compared to the temperature dependence. The effective mass m_e^* varies from 1.08 to 1.13 m_0 over the temperature range from 250–500 K, while m_h^* varies from 1.10 to 1.29 m_0 in the same temperature range.²⁵ In fact, in this calculation the effective masses are considered to be constant over temperature. Taking into account the temperature dependence of the bandgap E_g and the bandgap narrowing

$$E_g(T) = E_{g0} - E_{gT} \cdot T = 1.206 \text{ eV} - 2.73 \cdot 10^{-4} \frac{\text{eV}}{\text{K}} \cdot T \quad (2.67)$$

$$\Delta E_g(T) = \frac{3q^2}{16\pi\epsilon} \sqrt{\frac{q^2 N}{\epsilon kT}} = \frac{\Delta E_{g0}}{\sqrt{T}} \quad (2.68)$$

we obtain for the collector current

$$J_c(T) = J_{\text{const}} \left(\frac{T}{300\text{K}} \right)^4 \left[a + b \left(\frac{T}{300} \right)^\zeta \right] \exp \left[\frac{1}{kT} \left(qV_{EB} - E_{g0} + E_{gT} \cdot T + \frac{\Delta E_{g0}}{\sqrt{T}} \right) \right] \quad (2.69)$$

where

$$J_{\text{const}} = \frac{\mu_0 k^4 (m_e^* \cdot m_h^*)^{3/2}}{2\pi^3 \hbar^6 \int_0^w N(x) dx (300\text{K})^{-4}}$$

is the temperature-independent current density and

$$a = 1 - \frac{1}{1 + (N/C_r)^\alpha} - \frac{\mu_1/\mu_0}{1 + (C_S/N)^\beta}$$

depending on impurity concentration and $b = (\mu_{\text{max}}/\mu_0)$. The temperature dependence of the collector current can be expressed as

$$\frac{1}{J_c(T)} \frac{dJ_c(T)}{dT} = \frac{4}{T} + \frac{\zeta b (T/300)^{\zeta-1}}{300 \cdot (a + b(T/300)^\zeta)} - \frac{1}{kT^2} \left[qV_{EB} - E_{g0} + \frac{3\Delta E_{g0}}{2\sqrt{T}} \right] \quad (2.70)$$

Suppose that an npn transistor with a uniform base doping of $N = 1 \cdot 10^{18} \text{ cm}^{-3}$ is supplied with $V_{EB} = 0.7 \text{ V}$ at room temperature $T = 300 \text{ K}$. The impurity concentration in the base region results in $\Delta E_{g0} = 0.381 (\text{eV/K})^{-1/2}$ and thus in a bandgap narrowing of $\Delta E_g = 22 \text{ meV}$. The temperature coefficient at this operation point is $(1/J_c)(dJ_c/dT) \approx 7\% / \text{K}$. Experiments²⁵ show that ΔE_g is rather temperature-independent between 280 and 400 K. The temperature coefficient is not strongly influenced by this deviation from the assumption of Eq. 2.68.

Thermal Instabilities, Second Breakdown

Supposing a locally existing rise in current density, there is an increase in the density of the dissipated power, that results in an elevated temperature in the space charge zone between the base and the collector. This leads to an increase of the intrinsic carrier density and of the current density, resulting in an increase in dissipated power and, thus, in a positive thermoelectric feedback.

This thermal instability takes place when the local elevated temperature reaches a certain trigger temperature, which is in the order of the intrinsic temperature. The intrinsic temperature is the temperature at which the intrinsic concentration equals the impurity concentration.

Intrinsic conduction can be seen as short circuit between collector and emitter, resulting in breakdown of the collector-emitter voltage. In some cases, where the current is limited by the external circuitry, the resulting decrease in the dissipated power leads to a decrease in the hot-spot temperature and in an electrothermal oscillation. However, in most cases, intrinsic conduction leads to the diffusion of impurities and, consequently, to permanent transistor malfunction.

Because the thermal conductivity of semiconductors is temperature dependent, heat conduction worsens with elevated temperatures. This effect also contributes to the positive electrothermal feedback. Coupled electrothermal simulation shows that emitter-current crowding in power transistors is weakened by electrothermal interaction.²⁶ Thus, breakdown occurs in the center of the cylindrical emitter and not at the edge.

Basically, this electrothermal feedback can be suppressed by emitter series resistances. Common techniques are special doping profiles, small contact holes, or external series resistances resulting in negative electrical feedback of the respective emitter finger. The current of a common emitter configuration, stabilized by an external emitter series resistance R_E , can be expressed as

$$J_c = J_{c0} \exp \left\{ \frac{qV_{BE} - qJ_c[(R_E + r_E)/\alpha + r_B(1/\alpha - 1)] - E_g(T_j)}{kT_j} \right\} \quad (2.71)$$

where α is the base current gain, r_B and r_E the internal base and emitter series resistances respectively, and E_g the bandgap at the junction temperature, which is calculated as $T_j = T_0 + R_{th} \cdot I_c V_{CE}$. Here R_{th} is the steady-state thermal resistance in K/W (kelvins per watt), measured from the junction to ambient.

Another possible approach to stabilize the temperature distribution, and by that way the current density, of a multiemitter bipolar transistor is the use of thermal

shunts. The emitter fingers are connected by so-called thermal lenses, which consist mainly of metal bridges with a thicknesses greater than 10 μm . These bridges thermally couple the emitters to each other. The advantages of thermal shunt technology are that the device's cutoff frequency and efficiency are not reduced.

Bipolar transistors can be rendered more resistant to failure from avalanche injection by incorporating graded impurity density profiles within the collector region. When the current density increases (high injection), the electric field is pushed in direction of the nn^+ -junction (Kirk effect). With linearly graded doping in the epitaxial layer, the critical current density can be increased by several factors.

To predict thermally caused second breakdown of bipolar transistors, there is a need to use either fully coupled device simulation, namely, a simulator, that solves Poisson, electron and hole continuity, and the heat diffusion equation, or thermoelectric circuit simulators,²⁷ operating with electrothermal compact models. Each of these devices must have a thermal pin connected to a thermal module representing heat conduction in solids by solving the heat diffusion equation $\nabla\lambda(T)\nabla T = c_v(\delta T/\delta t) - p_v$, where $\lambda(T)$ is the temperature-dependent thermal conductivity [in silicon $\lambda(T) = 1.5486 (T/300\text{ K})^{-4/3} \text{ W}/(\text{K}\cdot\text{cm})$,²⁸ c_v is the specific heat capacity per volume in $\text{J}/(\text{K}\cdot\text{cm}^3)$, and p_v is the density of dissipated power in W/cm^3 . Reference 27 describes a three-dimensional (3D) thermal model representing nonlinear heat conduction in silicon, a two-dimensional (2D) model representing heatspreading in the leadframe, and a thermal resistance layer representing coupling between silicon and leadframe. In Ref. 29, thermal compact models representing the packaging are added to the thermal module. All thermal equations and interface and boundary conditions are built into a circuit simulator, thus solving the fully coupled electrothermal problem.

Influence of Thermal Diffusion and Heat Flux on Device Performance

With the ever decreasing device geometries of BJTs, the common drift diffusion model is often no longer valid. Generally, the current density in semiconductors can be written as

$$\vec{j} = \sigma \cdot \left(\nabla \frac{E_F}{q} - P \nabla T \right) \quad (2.72)$$

where $\sigma = \sigma_n + \sigma_p = q\mu_n n + q\mu_p p$ is the conductivity, E_F the Fermi energy, and P the thermoelectric power. The current density and thermoelectric power can be split into an electron and hole part, resulting in

$$\vec{j} = \vec{j}_n + \vec{j}_p = -q\mu_n n(\nabla\Phi_n + P_n\nabla T) - q\mu_p p(\nabla\Phi_p + P_p\nabla T) \quad (2.73)$$

The gradients of the quasi-Fermi potentials $\nabla\Phi_{n,p}$ represent the common drift and diffusion currents. Additionally, a thermal gradient in a semiconductor with Dirichlet boundary conditions (applied voltage sources) leads to a change in current

density. On the other hand, supposing no current or a constant current flow, a thermal gradient results in a change in the descent of the quasi-Fermi potentials.

Generally, there is always a thermal gradient arising from Joule or recombination heating at the collector–base junction. The thermoelectric power of a nondegenerated semiconductor can be expressed as:

$$P = \frac{q\mu_n n}{\sigma} P_n + \frac{q\mu_p p}{\sigma} P_p \quad (2.74)$$

$$P_n = -\frac{k}{q} \left(\frac{5}{2} - s + \frac{E_c - E_F}{kT} \right) = -\frac{k}{q} \left(\frac{5}{2} - s + \ln \frac{N_c}{n} \right) \quad (2.75)$$

$$P_p = \frac{k}{q} \left(\frac{5}{2} - s + \frac{E_F - E_v}{kT} \right) = \frac{k}{q} \left(\frac{5}{2} - s + \ln \frac{N_v}{p} \right) \quad (2.76)$$

where $s = \frac{1}{2}$ for scattering with phonons and $s = -3/2$ for scattering with impurities (the mean free time between collisions is expected to be $\tau \propto E^{-s}$). The thermoelectric power of electrons is negative, while that of holes is positive, an effect that is often used to determine the conduction type of semiconductor materials. Assume a current flowing along the spatially decreasing temperature, then the thermal diffusion of electrons is along the gradient, thus reducing the current density, while the thermal diffusion of holes along the above-assumed gradient contributes to the current flow. The thermoelectric power of electrons and holes of a 10^{18} cm^{-3} n- or p-doped material at room temperature (300 K) is $P_n = -0.631 \text{ mV/K}$ and $P_p = 0.546 \text{ mV/K}$, respectively. Supposing an impurity level of 10^{15} cm^{-3} , the thermoelectric powers are $P_n = -1.227 \text{ mV/K}$ and $P_p = 1.141 \text{ mV/K}$, respectively.

Concerning energy transport, the power flux density can be written as

$$\vec{S}_{\text{tot}} = -\kappa \nabla T + P_n T \vec{j}_n + P_p T \vec{j}_p + \Phi_n \vec{j}_n + \Phi_p \vec{j}_p \quad (2.77)$$

$$\kappa = \lambda + \left(\frac{k}{q} \right)^2 T \left[\sigma \left(\frac{5}{2} - s \right) - \frac{\sigma_n \sigma_p}{\sigma} \left(\frac{E_g}{kT} + 2 \left(\frac{5}{2} - s \right) \right) \right] \quad (2.78)$$

The first term in Eq. 2.77 represents heat transport due to a thermal gradient. The transport via the lattice, given by the first term in Eq. 2.78, is, in most cases, the dominating contribution. The thermal conductivity of the lattice in silicon is $\lambda \cong 1.55 \text{ W/(K} \cdot \text{cm)}$ at room temperature. Another contribution to energy transport due to thermal gradients is heat transport via carriers, shown in Eq. 2.78 by the second term. The constant, relating the product of temperature and electrical conductivity to the thermal conductivity, called the *Lorenz number*, equals $(k/q)^2 (5/2 - s) \approx 1.5 \cdot 10^{-8} (\text{V/K})^2$, considering only phonon scattering ($s = \frac{1}{2}$). The last term in Eq. 2.78 represents heat transport via mixed conduction, which can be quite large when $E_g \gg kT$.

The energy transport due to the fourth and fifth term on the right-hand side of Eq. 2.77 is the product of the particle flux and the respective electrochemical potential. To explain the meaning of the second and third term of Eq. 2.77, we assume no contribution due to thermal gradients.

By inserting Eq. 2.75 in Eq. 2.77, the energy transport by electrons under the assumptions given above can be approximated by

$$\vec{S}_n \approx (P_n T + \Phi_n) \vec{j}_n = - \frac{(5/2 - s)kT + E_c}{q} \vec{j}_n.$$

Electrons possess a potential energy, represented by the conduction band, and an additional kinetic transport energy of $(\frac{5}{2} - s)kT \approx 52$ meV at room temperature in silicon. Assume a metal–semiconductor contact and a current flow from the metal to the semiconductor. This current flow results in a decrease in the potential-energy level for electrons flowing from the semiconductor to the metal. The gain in potential energy, which equals the difference from conduction band energy of the semiconductor to the Fermi energy of the metal, heats the lattice. The kinetic energy remains the same.

For holes there is another contribution besides the difference in potential energy between the Fermi level and the valence band. At the contact, an electron–hole pair is generated, each carrier having the kinetic energy given above. For current flow from the metal to the semiconductor, the necessary energy for constant energy flow across the contact is two times the kinetic energy plus the Fermi–valence band difference, which is

$$[2(5/2 - s)kT + (E_F - E_v)] \frac{j_p}{q} = \left(5 - 2s + \ln \frac{N_v}{N_A} \right) \frac{kT}{q} j_p$$

This energy must be provided by the lattice, thus cooling the contact. This effect, resulting from current flow across inhomogeneous materials, is called the *Peltier effect*.

The influence of modeling the flux components, driven by the thermal gradients, namely, the thermal diffusion (TD) components of the current densities and the heat flux components (HF) of the energy flux densities have been investigated.³⁰ The authors conclude that it is important to take into account the thermal driving forces, based on a device simulation of a 50-nm-base-width npn transistor, by varying the influence of the contributions of the thermal gradients. This results in a significant variation of output current.

The decrease of the thermal gradient contribution in the current–density equation leads to an increase in collector current mainly in the active region, whereas a decrease of the influence of the thermal gradient in the energy flux equation mainly affects the saturated region.

Modeling of One-Dimensional Heat Transfer in Bipolar Transistors: Analytical Approach

Figure 2.14 shows the schematic cross section of an interdigitated BJT. Heat is dissipated at $x = x_1$, while the silicon surface is adiabatic with $\text{grad } T = 0$ (Neumann boundary condition) and the bottom of the chip is at ambient temperature, which is constant (Dirichlet boundary condition). T is the difference between actual temperature and ambient temperature. The thermal conductivity of silicon $\lambda(T)$ can be expressed as $\lambda(T) = [1/(a + bT_{\text{abs}} + cT_{\text{abs}}^2)]$, where T_{abs} is the absolute temperature, and $a = 0.03 \text{ (K} \cdot \text{cm)}/\text{W}$, $b = 1.56 \cdot 10^{-3} \text{ cm}/\text{W}$, $c = 1.65 \cdot 10^{-6} \text{ cm}/(\text{W} \cdot \text{K})$ or in another model,²⁸ as $\lambda(T) = 1.5486 \text{ W}/(\text{K} \cdot \text{cm})(T_{\text{abs}}/300 \text{ K})^{-4/3}$. The nonlinear thermal diffusion equation

$$c_v \frac{\delta T}{\delta t} = \text{div}(\lambda(T) \text{grad } T) + p_v \tag{2.79}$$

where c_v is the specific thermal capacitance of silicon per volume ($c_v = 1.63 \text{ J}/(\text{K} \cdot \text{cm}^3)$), p_v is the density of dissipated power per volume in W/cm^3 , and T is the elevated temperature with respect to ambient temperature. This equation can be transformed into a stationary linear one by applying the Kirchoff transformation

$$\vartheta(T) = \int_{T_0}^T \frac{\lambda(T')}{\lambda(T_0)} dT' \tag{2.80}$$

This results in the differential equation

$$\frac{c_v}{\lambda(T)} \frac{\delta \vartheta}{\delta t} = \Delta \vartheta + \frac{p_v}{\lambda(T_0)},$$

which is linear, concerning the stationary case. The Kirchoff transformation is a well suited method to apply Green's functions in monolayer structures or structures with the same exponential behavior of the thermal conductivity. In the following equations the temperature dependence of the thermal conductivity of the semiconductor is neglected by setting $\lambda(T) = \lambda(T_0)$.

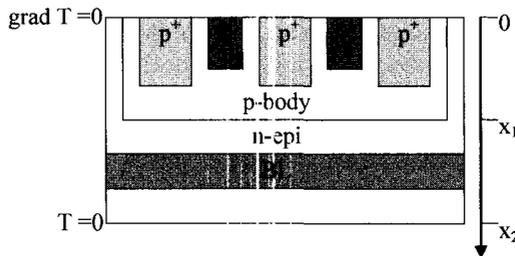


Figure 2.14 Multiemitter bipolar transistor as example of one-dimensional heat conduction to the bottom of the die.

The pulse response of a Dirac pulse generated at $x = x_1$ and $t = 0$ is given as

$$z(x, t) = \frac{1}{\lambda A \alpha x_2} \sum_{n=1,3,5..}^{\infty} \left(\cos \frac{|x - x_1| n \pi}{x_2} + \cos \frac{|x + x_1| n \pi}{x_2} \right) \exp \left(-\frac{n^2 \pi^2}{4 \alpha x_2^2} t \right) \quad (2.81)$$

where $\alpha = c_v / \lambda$ is the reciprocal of the thermal diffusivity. Now, for a general power signal $p(t)$, generated at $x = x_1$, the transient elevated temperature can be calculated by convolution:

$$T(x, t) = p(x_1, t) \otimes z(x, t) = \int_0^t p(t') z(x, t - t') dt' \quad (2.82)$$

The response to an applied power step $p(t) = P_0 s(t)$, where $s(t)$ is the Heaviside function, is referred to as Z_{th} curvature in packaging literature and offered by packaging vendors to decide whether a package is suitable to conduct heat generated by a power pulse, to the heatsink efficiently. This Z_{th} curvature is expressed as

$$Z_{th} = \frac{T(t)}{P_0} = \frac{x_2}{\lambda A} \sum_{n=1,3,5..}^{\infty} \frac{\cos \frac{|x - x_1| n \pi}{x_2} + \cos \frac{|x + x_1| n \pi}{x_2}}{\frac{n^2 \pi^2}{4}} \left(1 - \exp \left(-\frac{n^2 \pi^2}{4 \alpha x_2^2} t \right) \right) \quad (2.83)$$

In steady state Eq. 2.83 can be simplified to $Z_{th} = (x - x_2) / \lambda A$. In case of an infinite heat sink ($x_2 = \infty$) or short timescales, where the thermal heat front does not reach the boundary, the thermal pulse response at $x = x_1$ can be simplified to $z(t) \cong (1 / \lambda A) (1 / \sqrt{4 \pi \alpha t})$, while the thermal step response at $x = x_1$ is $Z_{th} \cong (1 / \lambda A) (t / \alpha \pi)^{-1/2}$

Compact modeling of one-dimensional heat conduction: The behavior of a linear system can be completely described by either the transfer function or, in the time domain, the step response. The thermal step response is the time-dependent temperature at the location of an applied power pulse, known as Z_{th} curvature (see Fig. 2.15). For modeling a given temperature step response using a Foster filter (Fig. 2.16), we follow an approach that is suitable for modeling heat conduction in materials whose thermal conductivity can be considered constant:³¹

The temperature step response of a Foster-filter is given by

$$T(t) = \sum_{n=1,2,3..}^{\infty} P R_{thn} \left(1 - \exp \left(-\frac{t}{\tau_n} \right) \right)$$

where R_{thn} and τ_n are the thermal resistance and the thermal time constant of the n th RC component, respectively, while P is the applied power pulse.

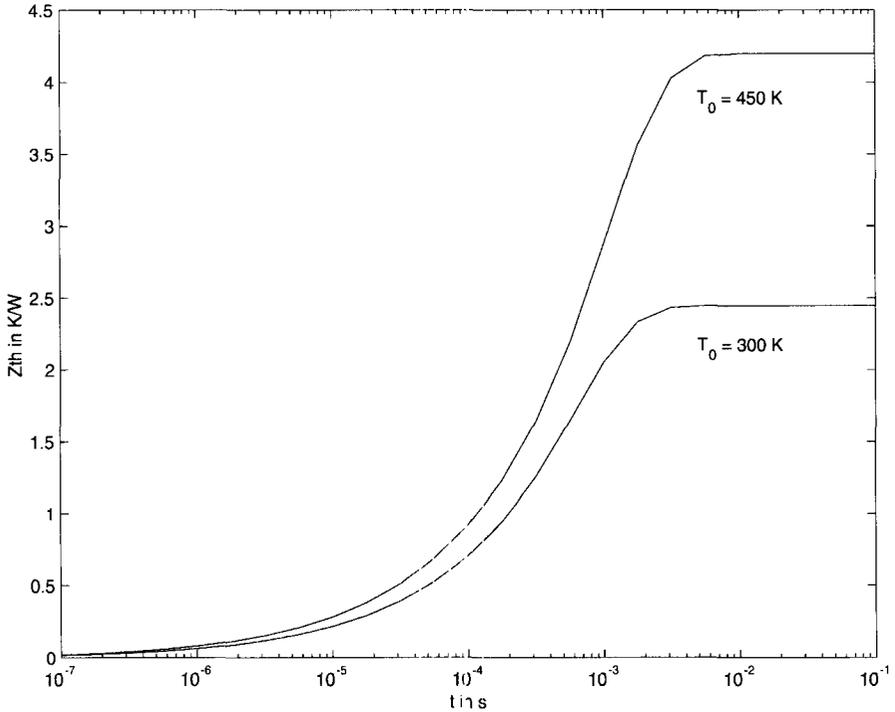


Figure 2.15 Z_{th} curvature of a 1-mm² power bipolar transistor, integrated on a 380- μ m substrate. Self-heating is not considered. The lines correspond to 450 and 300 K ambient temperature.

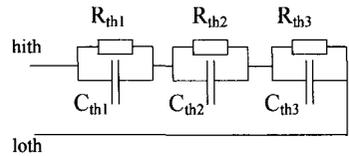


Figure 2.16 Foster filter of third order representing the essential poles and zeros of the step response.

If we assume an infinitely dense Foster filter and a logarithmic timescale $\zeta = \ln t$ and $\Theta = \ln \tau$, the response can be written as $T(\zeta) = \int_{-\infty}^{+\infty} Pr(\Theta)(1 - e^{-e^{\zeta-\Theta}})d\Theta$, where $r(\Theta)$ is the time-constant-dependent density of the thermal resistance. If this function is differentiated with respect to ζ , the result is

$$\frac{\partial T(\zeta)}{\partial \zeta} = \int_{-\infty}^{+\infty} Pr(\Theta)e^{-e^{(\zeta-\Theta)}+(\zeta-\Theta)}d\Theta \tag{2.84}$$

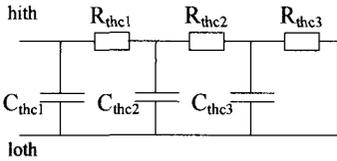


Figure 2.17 A Cauer filter of third order. This filter is used to model physically exact one-dimensional heat conduction in solids.

The time-constant-dependent density of the thermal resistance $r(\zeta)$ can be integrated, discretized, and associated with the respective Foster RC component by

$$R_{thn} = \frac{1}{P} \int_{\zeta_i - \Delta\zeta/2}^{\zeta_i + \Delta\zeta/2} r(\xi) d\xi \quad \text{and} \quad C_{thn} = \frac{e^{\zeta_i}}{R_{thi}} \quad (2.85)$$

Depending on the number of discretization points, the order of the filter and, therefore, the accuracy of the model can be chosen.

The Foster filter models the essential poles and zeros of the thermal step response. However, to obtain a physically correct behavior for timescales outside the measured or simulated range, it is necessary to use the Cauer filter, which models, analog to an electrical line, one-dimensional heat conduction in solids. The transformation of the Foster network into a Cauer network (Fig. 2.17) is a standard procedure in system theory.

Safe Operation Area (SOA)

The designer needs a criterion to decide whether the specification and the lifetime of the single device can be guaranteed. This criterion is called the *safe operation area*. Every single $I_C - V_{CE}$ curve must be within the area defined by the criterions given below.

The operation of the bipolar transistor is limited by

1. The maximum collector current (I_{max}), given by the upper limit in the diagram. This current is limited by
 - The melting point of the bonding wires: The melting point of Al99.99 is given as 660°C, of AlSi1 as 600–655°C, of Au-99.99 as 1063°C and of Cu as 1083°C. The critical current density depends on the diameter and length of the bonding wire. For Al, the critical current density is in the order of $j \approx 10^4 - 10^5$ A/cm², for Au the critical current density is up to one order higher than that for Al.
 - The melting point of the metallization is in the order of 600°C.
 - Electromigration aspects are as follows. For mono- and polycrystalline films with grain sizes exceeding 1 to 2 μm, the maximum current density³² is $j \approx 10^5$ A/cm². The main scattering effect is the lattice scattering. For thin layers (50–500 nm), and grain sizes of 20–60 nm scattering at grain boundaries and surface scattering dominate. Electromigration starts at about $j \approx 10^3$ A/cm² and can be improved by sophisticated deposition

technologies to up to $j \approx 10^6$ A/cm². The thermal activation energy, necessary for electromigration in pure Al is low, due to the low melting point (660°C). AlSiCu offers, due to the copper, a higher activation energy. The jI_{th} product, which is the product of critical current density and length of the conducting metal, is given for Cu as $jI_{th} \approx 1200$ A/cm for temperatures in the range of 170–200°C.³³ At higher temperatures the product decreases to $jI_{th} \approx 400$ A/cm at about 280°C.

2. The maximum collector–emitter voltage (V_{max}) results from
 - Avalanche breakdown
 - Second breakdown
 - Punchthrough
 - The use of the compact model describing the electrical characteristics of the device based on the electric field
3. Static dissipated power, resulting in an increase of the device’s junction temperature
 - $P_{max} = (T_{max} - T_{ambient})/R_{th}$. Depending on the package, the mounting conditions [printed-circuit board (PCB), cooling fin, metal plate etc.], the heat dissipating area and the dissipated power, the R_{th} value can be between 0.6 K/W (TO218 on a thermal chuck) and 300 K/W (SO8 on a PCB). If the package is floating in air, the R_{th} value can be even worse.
 - The P_{max} -line is a -45° line in the logarithmic diagram (Fig. 2.18).

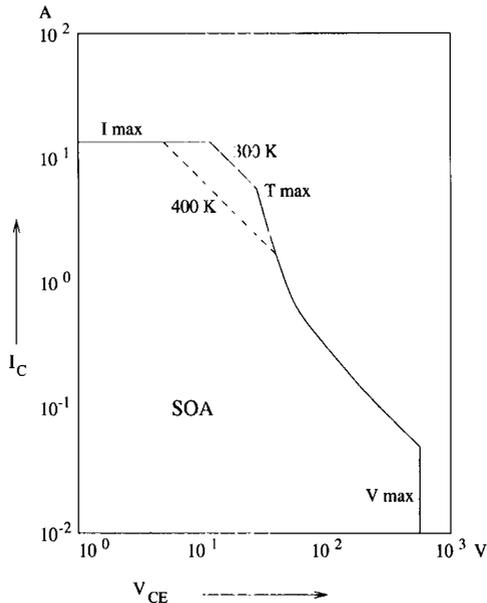


Figure 2.18 Safe operation area of a power bipolar transistor.

- A change in ambient temperature results in a parallel shifting of the -45° line (in Fig. 2.18 the ambient temperature is shifted from 300 to 400 K).
4. Static dissipated power as a function of the applied collector–emitter voltage
 - The -45° line becomes steeper with increasing V_{CE} . This results because second breakdown in the active region is not homogeneous. The current density is concentrated on parts of the device, and this leads to a locally high elevated temperature. The inhomogeneity increases with the applied collector–emitter voltage, thus the maximum dissipated power is a function of V_{CE} .
 5. Transient dissipated power
 - Dynamically generated power leads to an extension of the SOA. The respective junction temperature resulting from short (submillisecond) pulses can be calculated analytically by treating the silicon as a semi-infinite heatsink. Longer pulses generally need numerical solutions. In some cases, only fully coupled electrothermal simulation offers predictable results. Transient conditions result in an elevated maximal current and in parallel shifting of the -45° line. Nevertheless the maximum allowed voltage (V_{\max}) remains constant.

Keeping all these conditions in mind, the circuit designer determines the operation point and design rules of critical devices.

2.4 SELF-ADJUSTED TRANSISTOR STRUCTURES

The lateral dimensions of modern transistors have shrunk to submicron dimension. For bipolar transistors, small emitter fingers mean low power consumption, higher speed (f_{\max} increases), and negligible current crowding, allowing for the higher base sheet resistances of thinner bases (f_T increases). At least as important is to get as near as possible to a box-shaped transistor with good external contacts. The collector area defines the collector depletion layer capacity C_C and the highly doped external base reduces parasitic base resistance components. To avoid tunneling, the external base must not overlap the emitter region. Sophisticated self-alignment techniques were developed to allow devices to come close to the idealized structures. A good overview about fundamental techniques and results is given in Ref. 34.

2.4.1 Polysilicon Base and Emitter Contacts

Figures 2.19 and 2.20 compare a conventional bipolar transistor with a self-aligned polysilicon emitter transistor. Common to both devices is a p^- substrate, an n^+ buried layer as a subcollector that connects to the collector contact (C). On top of the substrate, a n^- epitaxial layer is deposited that contains collector, base, and emitter. Outside the transistor area and between the base (B) and collector (C) contact an

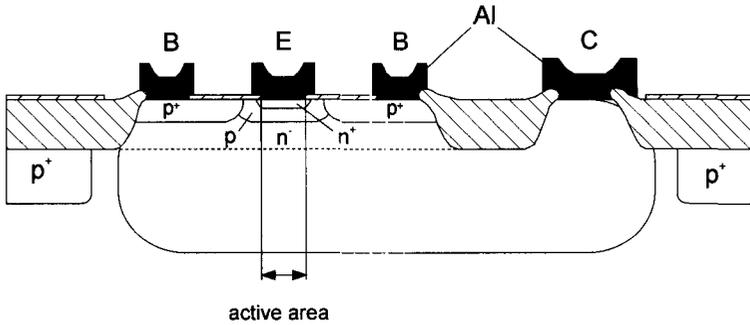


Figure 2.19 Conventional bipolar transistor. (After Treitinger and Miura-Mattauch, Ref. 34.)

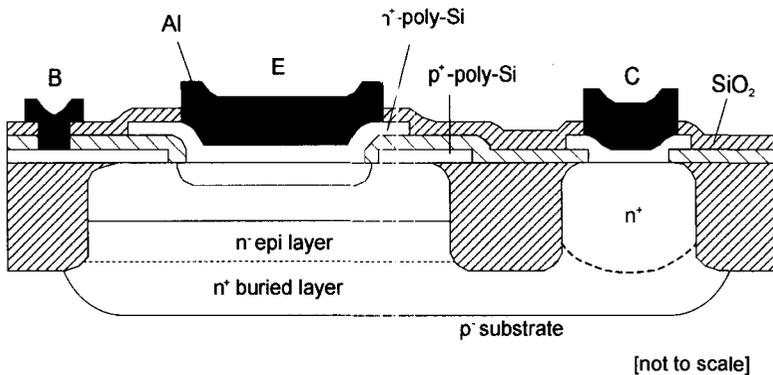


Figure 2.20 Double poly-Si emitter transistor. (After Treitinger and Miura-Mattauch, Ref. 34.)

oxide isolation is performed. In the conventional device the active regions are formed by ion implantation. The external contact region, which deliver high parasitical capacitances and resistance, cannot be reduced because of the limitations in the alignment accuracy of the lithographic process. The conventional device suffers from large base–collector (BC) and collector–substrate (CS) capacitances and large extrinsic base resistances. In spite of these handicaps, reasonably good results have been obtained with conventional techniques. Some of the disadvantages have been overcome by the self-aligned polysilicon technology.

Polysilicon technology does not mean replacement of the monocrystalline transistor areas. In polysilicon technology a highly doped polysilicon region is placed between the device and the metallic contact. This region serves as a source for the diffusion yielding very shallow and steep doping profiles, such as for an arsenic emitter with emitter depths less than 100 nm. The common technique now is to use poly for the base and emitter contact (double-poly or inside-spacer technique). A single-poly technique (outside spacer) is also available.

Double-Poly Transistor with Inside Spacer

Figure 2.21 shows the essential steps of the double-poly technique.³ In Figure 2.21a, into a layer stacking of the p⁻ substrate, n⁺ buried layer and, n epitaxial layer, a trench will be etched by anisotropic etching. A nitride layer with thin pad oxide and a masking layer for trench etching is deposited on the n⁻ epitaxial layer. After mask definition, trench etching follows using anisotropic dry etching such as reactive ion etching (RIE). The trenches penetrate both the epitaxial and buried layers to reach the p⁻ substrate. In Figure 2.21b, in combination with local oxidation (LOCOS) the transistor is isolated. In Figure 2.21c, the active base region is formed by ion implantation or diffusion. With boron ion implantation the surface is preamorphized by Si⁺ implantation to reduce channeling. In Figure 2.21d, a p⁺ poly Si/SiO₂ sandwich is deposited and a window is etched in the base/emitter area. The p⁺ base contact is created by diffusion from the p⁺ polysilicon. In Figure 2.21e, a thin spacer (0.15–0.25 μm) is created at the sides of the p⁺ poly Si/SiO₂ stacking by deposition and backetching. An n⁺ poly Si film is deposited and structured. After outdiffusion of the emitter region (n⁺) a structure with very small self-aligned separation of the n⁺ emitter from the p⁺ base contact is obtained. In Figure 2.21f, contact layer and Al metallization for E, B, C contacts are formed.

Self-alignment results not only in small transistor areas and short distances between external and internal base but also in small emitter fingers because with inside-spacer technique the emitter stripe width is smaller (twice the spacer) than the

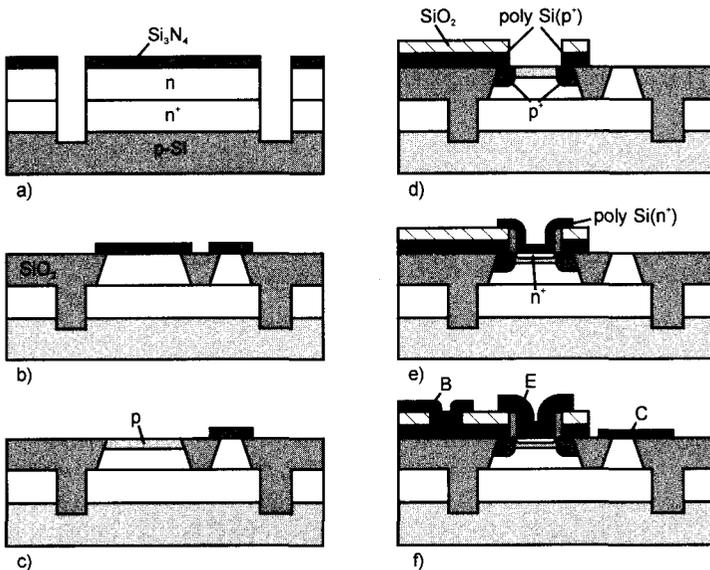


Figure 2.21 Fabrication sequence for a double-polysilicon bipolar transistor: (a) trench etching; (b) trench filling with oxide; (c) active base layer formation; (d) diffusion of base contact from the p⁺ poly-Si; (e) inside spacer, n⁺ poly-Si, emitter diffusion; (f) metal contacts. (After von Münch, Ref. 3.)

lithography dimensions. The basic process sequence may be combined with other techniques such as trench isolation, which was discussed in the example (Figs. 2.19, 2.20), which consisted of a silicide/poly-Si (salicide) sandwich for low-resistance interconnects, selective collector implantation (SIC) to form a high doped pedestal on the subcollector, in situ doping of the poly-Si (IDP) to reduce the thermal load compared to the implanted poly-Si, and epitaxial growth of the base profile (epitaxial base transistor). The epitaxial growth can be either selective within oxide windows or can be of differential crystallinity³⁵ with single crystalline active transistor areas and poly-Si layers on top of the oxide, or it can be uniform. From the many variants of the double-poly process, let us discuss in some detail the SMI process which uses a self-aligned metal/in situ-doped poly (SMI) technology for the base contact.³⁶ In this technology (Fig. 2.22) a thin B-doped poly-Si (IBDP), in situ boron-doped poly-Si (Fig. 2.22a) is covered by a stack SiO_2 /undoped poly-Si/ SiO_2 where the undoped poly-Si layer is later used as sacrificial layer as in surface micromechanics. Then the emitter window, the inside spacer and the n-doped poly-Si (Fig. 2.22b, IPDP, in situ phosphorus-doped poly-Si) is formed and protected by a thick SiO_2 cover. The sacrificial poly-Si is etched off (Fig. 2.22c). Then (Fig. 2.22d) a metal layer (tungsten) is selectively grown on the p^+ poly Si, refilling the space under the protecting overhang. Using this technique, a low-resistivity base connect is formed within a spacer thickness near the intrinsic base. The tungsten layer has a sheet resistivity of $2 \Omega/\text{square}$ compared with the typical sheet resistivities of intrinsic base, extrinsic base, p^+ -poly connect of $10,000 \Omega/\text{square}$, and $50 \Omega/\text{square}$, respectively. The single poly-Si process utilizing an outside spacer technique is schematically shown in Figure 2.23. At the position of the later emitter, a silicon nitride cover with an outside spacer allow the self-adjusted formation of the p^+ extrinsic base. After the nitride cover is removed, the n^+ poly Si is deposited and

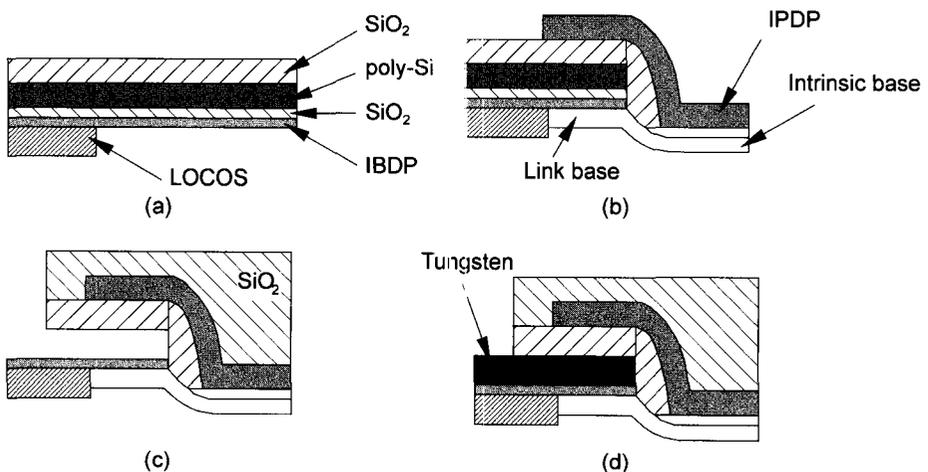


Figure 2.22 Double-poly-Si process following technology (self-aligned metal/in situ doped poly-Si (SMI)). (After Onai et al., Ref. 36.)

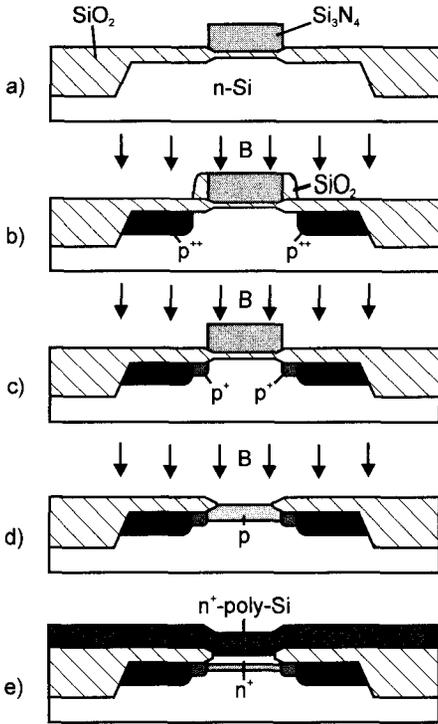


Figure 2.23 Fabrication sequence for a single poly emitter bipolar transistor; (a) selective oxidation (SiO_2) and nitride (Si_3N_4) mask above the later-fabricated emitter; (b) outside spacer formation (SiO_2) and implantation of the base contact p^{++} ; (c) removal of the spacer and implantation of the base link (p); (d) removal of the nitride and active base implantation; (e) n^+ -poly Si deposition and emitter diffusion. (After von Münch, Ref. 3.)

annealed to outdiffuse the shallow emitter region. The n^+ poly on top of the shallow emitter influences the current gain and the emitter resistance. The current gain can be increased at the cost of emitter resistance by a thin, natural-oxide layer (< 1 nm) between single and poly crystalline Si. The current gain increase is probably caused by a tunneling barrier at this boundary.

2.4.2 Operation of Self-Adjusted Transistors at High Current Density Conditions

The small dimensions of self-adjusted transistors and low supply voltages allow high current densities because of less stringent thermal limitations and a higher critical current density before the onset of the Kirk effect in low BV_{CE0} transistors because of higher collector doping N_C . At high current densities some simplifying assumptions of transistor modeling are no longer valid. In Refs. 37 and 38, various compact models are compared with one- and two-dimensional device simulations. Some results are presented for a transistor with the vertical dimensions summarized in Table 2.4. The transfer characteristics (I_C vs. V_{BE}) are shown in Figure 2.24. Deviations from the ideal exponential characteristics (ideality factor $\eta = 1$) caused by high current injection are seen above $V_{BE} = 0.78$ V, with severe deviations above 0.86 V. This is in rather good agreement with the assessment (Eq. 2.34) of high carrier injection ($n > N_B$) above. V_{BE}^* . At the BC junction, carrier distribution and

TABLE 2.4 Metallurgical Depths w_{E0} , w_{B0} , w_{C0} of the Transistor Used for Modeling^a

w_{E0} (μm)	w_{B0} (μm)	w_{C0} (μm)	N_C (cm^{-3})	R_{SB0} ($k\Omega$)	Q_{p0} ($\text{fC}/\mu\text{m}^2$)
0.090	0.125	0.400	$2.4 \cdot 10^{16}$	15	4.394

^aFor the base the integral numbers sheet resistivity R_{SB0} and hole charge Q_{p0} at thermal equilibrium are given. The collector doping N_C is uniform. The exact profiles of emitter doping ($N_E \cong 1.5 \cdot 10^{20} \text{ cm}^{-3}$) and base doping ($N_B = 1.5 \cdot 10^{18}$ to 10^{17} cm^{-3}) are extracted from experimental SIMS profiles.

Source: Schröter, Ref. 38.

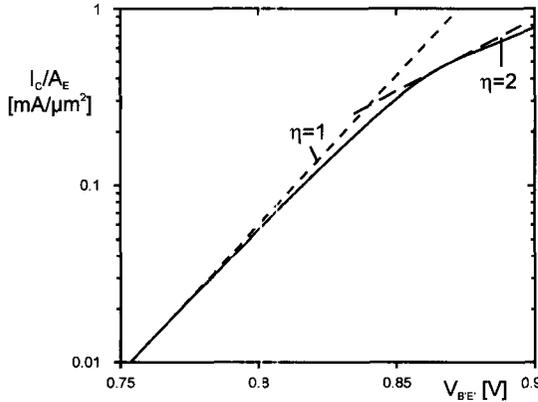


Figure 2.24 Transfer characteristics calculated for the transistor of Table 2.4. Collector current ($\log I_C$) as function of the base-emitter voltage V_{BE} . (After Schröter, Ref. 38.)

electric field are totally changed when the collector current exceeds a critical value I_{crit} ($I_{crit} = 0.42 \text{ mA}/\mu\text{m}^2$ in the example). Figure 2.25 shows the electric field (Fig. 2.25a) and the electron–hole densities (Fig. 2.25b) in the collector region. At small collector currents ($I_C \ll I_{crit}$) the depletion layer with negligible mobile carrier is created (Schottky approximation, curve 1, Fig. 2.25). At very high current densities ($I_C > I_{crit}$) the depletion layer no longer exists because of the high density of injected electrons that are partly neutralized by a high hole density (curve 2 in Fig. 2.25). The maximum electric field moves from the metallurgical BC junction to the collector–subcollector junction. The base pushout (Kirk effect) causes a strong degradation of current gain and frequency. This is clearly shown in a τ_{ec} versus $1/I_C$ (Fig. 2.26) plot where τ_{ec} decreases with increasing current up to Kirk effect deviation ($\Delta\tau$) from the ideal curve. In this plot the increase s is a measure of the junction capacitances $C_E + C_C$. At high currents ($1/I_C \rightarrow 0$) the capacities are loaded and unloaded rapidly, whereas at low currents ($1/I_C \rightarrow 100 \mu\text{m}^2/\text{mA}$) the load delay dominates.

The last figures were obtained with one-dimensional device simulation using the transistor described in Table 2.4. Two-dimensional device simulations³⁸ for transistors with small finger widths b_E exhibit a softer increase of delay time τ_{ec} for high current densities ($I_C > I_{crit}$). This is caused by a current spreading below the emitter finger that reduces the real current densities below the nominal calculated I_C/A_E .

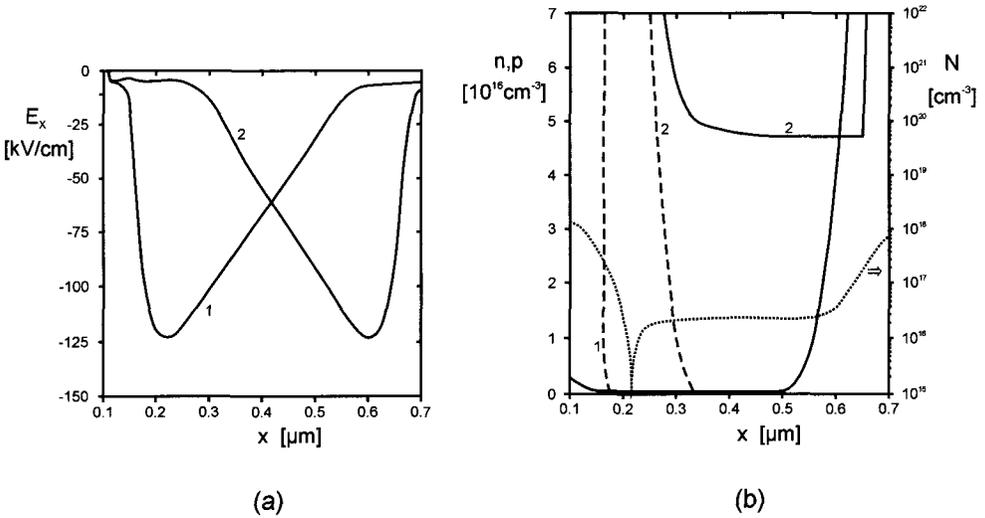


Figure 2.25 (a) Electric field and (b) electron density (—) and hole density (----) in the collector region ($x = 0.225\text{--}0.625\ \mu\text{m}$). Curve 1 is for low current density ($I_C = 0.01\ \text{mA}/\mu\text{m}^2$) and curve 2, for high current density ($I_C = 0.835\ \text{mA}/\mu\text{m}^2$). For comparison, the netto doping $N(|N_D - N_A|)$ is also given ($\dots\dots$). (After Schröter, Ref. 38.)

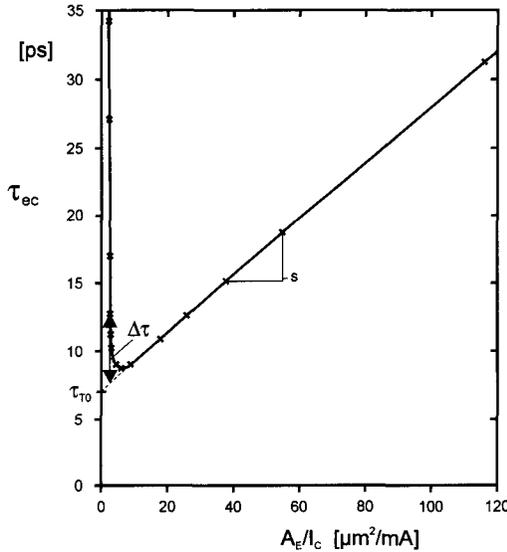


Figure 2.26 Delay time $\tau_{ec} = (\frac{1}{2} \pi f_T)$ as function of the inverse current density A_E/I_C .

2.4.3 Device and Circuit Results

Self-alignment and careful vertical profile control have reduced parasitic elements and improved speed and noise properties. Frequency limits f_T, f_{max} increased well into the millimeter-wave region (>30 GHz) with leading-edge developments up to

TABLE 2.5 Parameters of Self-Adjusted BJT^a

Emitter mask size	A_E	$0.6 \times 20 \mu\text{m}^2$
Spacer width	S	150 nm
Current gain	β	90
EB breakdown	BV_{EB0}	2.0 V ($I = 1 \mu\text{A}$)
CB breakdown	BV_{CB0}	10.0 V ($I = 1 \mu\text{A}$)
CE breakdown	BV_{CE0}	2.9 V ($I = 1 \mu\text{A}$)
EB junction capacitance	C_{JE}	81.6 fF (at zero bias)
CB junction capacitance	C_{JC}	44.0 fF (at zero bias)
CS capacitance	C_{JS}	91.0 fF (at zero bias)
Cutoff frequency	f_T	36 GHz ($V_{BC} = 0 \text{ V}$, $I_C = 10.5 \text{ mA}$)
Maximum oscillation frequency	$f_{T \text{ max}}$	38 GHz ($V_{BC} = 0 \text{ V}$, $I_C = 10.5 \text{ mA}$)
Total base resistance	r_b	23Ω ($V_{BC} = 0 \text{ V}$, $I_C = 10.5 \text{ mA}$)
Base sheet resistance	R_{pinch}	12 k Ω square

^a β , f_T , and $f_{T \text{ max}}$ are maximum values with respect to collector current.

Source: Aufinger et al., Ref. 40.

transit frequencies³⁹ $f_T = 100 \text{ GHz}$ and well-controlled production technology of $f_T = 25 \text{ GHz}$. Typical device parameters for a near-production technology⁴⁰ are given in Table 2.5. The emitter size is given as $0.6 \times 20 \mu\text{m}^2$ with a spacer width of 150 nm. From this an effective emitter width b_E of 0.3–0.4 μm may be assessed. Low base resistance ($\approx 1 \text{ k}\Omega/\mu\text{m}$ emitter length) and low CB capacitance ($\approx 6 \text{ fF}/\mu\text{m}^2$ emitter area) and outbalanced frequency data ($f_T \cong f_{T \text{ max}} = 37 \text{ GHz}$) are shown. The limited voltage range follows from the collector design. CB breakdown (voltage) is $BV_{CB0} = 10 \text{ V}$ and CE breakdown BV_{CE0} is reduced to 2.9 V because of the current gain. Transit-time analysis indicates junction capacitances of $C_E + C_C = 224 \text{ fF}$. This capacitance is composed of the near current, independent-collector junction ($C_C = 44 \text{ fF}$, up to the Kirk effect regime) and the forward-biased emitter–base junction ($C_E \cong 180 \text{ fF}$ compared to $C_E \cong 80 \text{ fF}$ at zero bias). Therefore, the load delay of the capacitances is mainly determined by the emitter–base junction. Excellent noise figures are reported with 1 dB up to 2 GHz and less than 2 dB up to 7 GHz. The measured minimum noise figure F_{min} agreed very well with the model,⁴¹ which is given in the form

$$F_{\text{min}} = 1 + (av + e) + [2(av + b) + (av + e)^2]^{1/2} \quad (2.86)$$

with the single components ($\tau_f = \tau_{ec} - \tau_g = \tau_B + \tau_C + \tau_C$, Eq. 2.58)

$$a = g_m r_b = \frac{I_C}{V_T} r_b; \quad b = \frac{1}{9} (\omega \tau_f)^2 + \frac{1}{2\beta}$$

$$d = \frac{2}{3} \omega \tau_f + \frac{\omega(C_E + C_C)}{g_m}; \quad e = \frac{4}{30} (\omega \tau_f)^2 + \frac{1}{\beta}$$

$$v = d^2 + e^2 + 2b \quad (2.87)$$

TABLE 2.6 Characteristics of Advanced Self-Adjusted BJT with an Emitter Size of $0.2 \times 2 \text{ mm}^2$

Base	RVD ^a	I/I ^b
A_E (μm^2)	0.2×2	0.2×2
d_{epi} (μm)	0.3	0.4
β	500	200
BV_{CE0} (V)	2.5	2.5
BV_{CB0} (V)	6.4	7.0
BV_{EB0} (V)	6.2	5.2
C_E (fF)	2.9	2.5
C_C (fF)	4.8	4.0
R_E (Ω)	30	25
R_C (Ω)	25	25
V_E (V)	8.0	3.5
f_T (GHz)	82	59
r'_{bb} (Ω)	240	390

^aRapid vapor-phase doping.^bIon-implanted base.

Source: After Ohue et al., Ref. 39.

It was stated that in the considered frequency range the same result was obtained when using the Hawkins expression.⁴² In either case, Ref. 40, demonstrates impressively the low RF noise of self-adjusted bipolar transistors, and shows that the thermal noise of the base resistance and the shot noise of the collector and base currents contribute to the observed results. For comparison, Table 2.6 exhibits the data of double-poly transistor fabricated with the advanced SMI technology. The emitter size was $0.2 \times 2 \mu\text{m}^2$. The base resistance (we consider the RVD version, left column of Table 2.6) was very small ($\approx 0.5 \text{ k}\Omega/\mu\text{m}$ emitter length), the collector capacity ($C_C \approx 12 \text{ fF}/\mu\text{m}^2$ emitter area) was somewhat larger than the zero-bias emitter capacity ($C_E \approx 7.5 \text{ fF}/\mu\text{m}^2$). Under forward bias, the emitter capacity will dominate slightly. The rather large collector capacity (for a box-shaped transistor C_C is clearly smaller than C_E because of the lower collector doping) is caused partly by an emitter stripe that is too small. Indeed, the highest transit frequencies ($f_T = 100 \text{ GHz}$) were obtained with a $0.6\text{-}\mu\text{m}$ stripe but for which no data were given. Disadvantages of an advanced vertical scaling are low breakdown voltages ($BV_{CE0} = 2.5 \text{ V}$) and low Early voltages ($V_{Ea} = 8 \text{ V}$).

Applications that use bipolar circuits are, for example, core processing units (CPUs) of mainframe computers, high-frequency measurement equipment, satellite- and mobile-communication systems, optical fiber links, fast signal processors, and amplifiers and power-consuming peripherals. The requirements for many of these applications are very different, but there are some common measures. For logic applications the propagation delay per logic gate is connected to the speed of a communication system (Gb/s). A rough assessment may be found in Ref. 43. In bipolar circuits, which are relatively immune to fanout, the propagation delay per

gate should be less than one-third of the clock period that is the reciprocal of the bit rate. After this rule of thumb, communication systems of 8, 16, 32, and 64 Gb/s require propagation delays of 40, 20, 10, and 5 ps, respectively. The individual components of an optical fiber link have been investigated in more detail.^{44,45} On the transmitter side, the multiplexer, and laser driver, and on the receiver side, the preamplifier, main amplifier, decision circuit, clock recovery, frequency divider, and demultiplexer were evaluated. With a production technology (Siemens B6HF, $f_T = 25$ GHz) all essential components of a 10-Gb/s system could be realized partly with much higher potential (e.g., multiplexer with 30 Gb/s). With a laboratory technique ($f_T = 35$ GHz, 50 GHz) some components for 40-Gb/s systems were developed, such as a 50-Gb/s multiplexer,⁴⁵ a 35-GHz static frequency divider,⁴⁷ and a 20-Gb/s XOR; gate.⁴⁸ With the $f_T, f_{max} = 50$ -GHz devices, ECL (emitter coupled logic) gate delays of 16 ps were obtained. With SMI double-poly devices, gate delays of 14.3 ps³⁶ and 12 ps⁴⁹ were obtained and a 45-GHz dynamic frequency divider was realized with $f_T = 50$ GHz, $f_{max} = 73$ -GHz devices.

2.5 HETEROJUNCTION BIPOLAR TRANSISTOR

The emitter injection efficiency α_T (common base configuration) given in Eq. 2.32 is valid only for junctions within one material. For a pn heterojunction, the emitter efficiency is changed because the barriers for electrons and holes differ (Fig. 2.27), leading to an improved emitter efficiency when the emitter has a larger bandgap. In an abrupt type I heterojunction ($\Delta E_C < 0, \Delta E_V > 0$, Fig. 27a) a disturbing heterobarrier appears at the interface that is not visible for a type II heterointerface ($\Delta E_C, \Delta E_V > 0$, Fig. 2.27b). For the type I we assume a graded interface, which avoids this barrier. In the wide-bandgap emitter the intrinsic carrier density n_{iE} is much smaller than that in the base n_{iB} , which changes the emitter

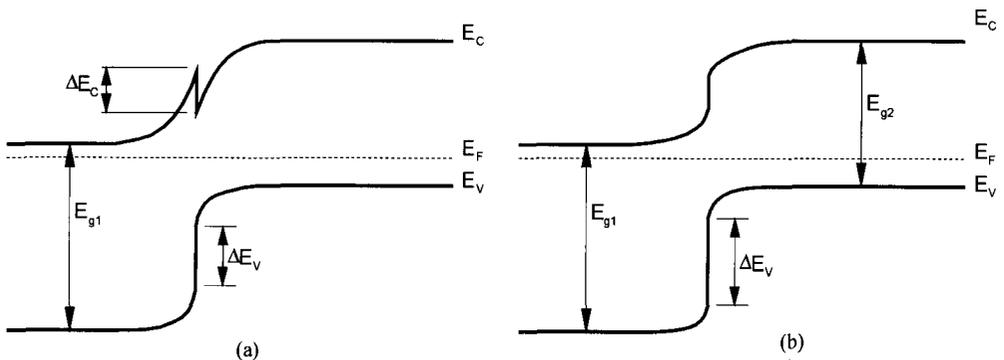


Figure 2.27 Band diagram of an abrupt pn heterojunction with wide-bandgap n emitter and narrow bandgap p base (a). For the usual type I heterojunction an energy spike or notch appears at the interface, which is not seen for the type II heterojunction (b) (E_C , conduction band edge; E_V , valence band edge; E_F , Fermi energy).

efficiency α_E to

$$\alpha_E = 1 - \frac{1}{\beta_E} = 1 - \frac{\mu_p w_B N_B n_{iE}^2}{\mu_n L_p N_E n_{iB}^2}$$

$$\beta_E = \frac{\mu_n L_p N_E n_{iB}^2}{\mu_p w_B N_B n_{iE}^2} \quad (2.88)$$

which reduces to Eq. 2.32 when $n_{iB}^2 = n_{iE}^2$. Remember that the transport properties are those of the minority carriers (μ_p, L_p in emitter, μ_n in base) and the hole diffusion length L_p in the emitter is assumed to be smaller than the emitter width w_E , otherwise the width w_E has to be used instead of L_p . The ratio n_{iE}^2/n_{iB}^2 can be written as

$$\frac{n_{iE}^2}{n_{iB}^2} = \frac{(N_C N_V)_E}{(N_C N_V)_B} \exp\left(\frac{\Delta E_g}{kT}\right) \quad (2.89)$$

where N_C, N_V are the effective density of states of conduction and valence band, respectively, with bandgap difference ΔE_g between emitter and base. Often for simplicity the different effective densities of states for base and emitter are set equal.⁵⁰ Sturm discusses the validity and limits of this assumption.⁵¹ If the base doping is nonuniform, the product $N_B w_B$ is replaced by the base Gummel number G_B :

$$G_B = \overline{N_B w_B} = \int_{x=0}^{x=w_B} N_B(x) dx \quad (2.90)$$

With nonuniform bandgap (compositional gradient) the collector saturation current is^{52,53}

$$J_{C0} = \frac{q}{V_T \mu_n} \left(\int_{x=0}^{x=w_B} \frac{N_B(x)}{n_{iB}^2(x)} dx \right)^{-1} \quad (2.91)$$

The current gain β (common emitter) of an HBT is roughly higher by a factor $\exp(\Delta E_g/kT)$ than that of a BJT of the same structural data (the transport factor α_T is set to unity in this assessment). The neglect of the transport factor is not allowed for heterostructures with reduced material quality (low L) or transistors with very high current gains (superbeta transistors):

$$\beta = \frac{\beta_E}{1 + \frac{1}{2} \beta_E (w_B/L_n)^2} \quad (2.92)$$

For instance, consider a transport factor $\alpha_T = 1 - 0.5 \cdot 10^{-3}$ that reduces the current gain of a usual transistor ($\beta_E = 100$) insignificantly to $\beta = 95$ whereas that of a superbeta transistor ($\beta_E = 1000$) is reduced to $\beta = 670$.

The heterobipolar transistor offers, as a first choice, an enhanced current gain (see Eqs. 2.88–2.92) when the doping profile is unchanged. In practice, the real advantage of the HBT is not to achieve high current gain but to trade gain for freedom in the base profile engineering, such as low base widths and high base dopings. In the HBT, the current gain β is guaranteed by the bandgap difference not solely by the base doping which is much smaller than the emitter doping in BJTs.

In a HBT the doping level N_A can be increased by a proper choice of ΔE_g without sacrificing the emitter efficiency α_E near unity. The HBT concept gives more freedom for the layer design of a bipolar transistor. This freedom can be used to obtain higher current gain β , a lower Early voltage V_{Ea} , a higher transit frequency f_T , and a lower base sheet resistivity R_{bi} . In the following section the classic concept of a wide-bandgap emitter is treated. Later sections deal with SiGe as a narrow-bandgap base material used in the already developed drift transistor to future strain-adjusted HBTs with strong doping-level inversion.

2.5.1 Wide-Bandgap Emitter

A wide variety of semiconductor materials exists with bandgaps larger than the gap of Si (1.12 eV at room temperature). However, for a broad application in silicon-based circuits the compatibility with silicon technology should be considered where group IV materials would be preferred. The search for appropriate group IV materials follows two different routes. One route relies on single crystalline materials. The bandgap (Table 2.7) of the diamond lattice group IV materials decreases with increasing atomic number.

Diamond (C) and SiC would be wide-bandgap candidates from Table 2.7. The β -SiC has indeed found much attention as a wide-bandgap emitter material for silicon based transistors.³⁴ Problems connected with the crystalline β -SiC/Si heterointerface include high growth temperatures for SiC, the large lattice mismatch (SiC lattice constant $a_0 = 0.436$ nm is much smaller than the Si lattice constant $a_0 = 0.543$ nm), and a type I band offset. A type I band offset results in an electron energy spike at the interface for the n/p^+ junctions of an HBT. Usually, this spike is avoided by a gradual transition, which is not possible in the SiC/Si system. The other route utilizes the bandgap modulation with phase changes (hydrogenated amorphous silicon a-Si:H) or with strong localization–quantization (microcrystalline silicon μ -Si). The heterophase boundary (a-Si/Si or μ -Si/Si) can be used for the HBTs

TABLE 2.7 Indirect Bandgap E_g (eV), Lattice Constant a_0 (nm), and Dielectric Constant ϵ for Group IV Materials (Diamond Lattice)

	C	β -SiC	Si	Ge	Sn
E_g	5.48	2.2	1.12	0.66	—
a_0	0.357	0.436	0.543	0.566	0.649
ϵ	5.7	6.5	11.9	16.2	(24)

because μ -Si ($E_g = 1.4 \text{ eV}$) and a-Si ($E_g = 1.7 \text{ eV}$) exhibit larger bandgaps than single crystalline silicon ($E_g = 1.2 \text{ eV}$). Doping of a-Si:H and low mobility and, therefore, high emitter resistances cause the main problems within this heterophase route. Absorption in a-Si is much stronger than in crystalline Si, which could be used in some optoelectronic applications of HBTs. The remaining problems of wide-bandgap emitter solutions with technology, parasitics, and material quality caused a shift to narrow-bandgap base solutions, which are described in the next sections.

2.5.2 Narrow-Bandgap Base

Instead of a wide-bandgap emitter, a narrow-bandgap base may be used to increase the emitter efficiency α_E . From Table 2.7 one can see Ge as possible material for a narrow-bandgap base. Indeed, the lattice mismatch between Si and Ge restricts technological meaningful solutions to SiGe alloys with rather small Ge contents and thicknesses. In high-frequency bipolar transistors, the base transit time τ_B contributes essentially to the total transit time ($\frac{1}{2}\pi f_T$) measured by the transit frequency f_T . For a pure diffusion current, the base transit time τ_B can be reduced by an internal electric field (carrier drift). A Ge content gradient through the base provides the necessary electrical field. With homogeneous doping (Fig. 2.28), the mean electric field strength is roughly given by $\Delta E_g/qw_B$, leading to reduced transit times.

The main advantages of the SiGe drift base are an easy implementation into existing technologies and a low Ge-content stable structure. The first industrial realizations of high-speed SiGe transistors and integrated circuits followed this concept.⁵⁴ The implementation into existing bipolar technologies leaves the drift transistor with a drawback common to high-frequency BJTs. Base sheet resistivities increase with decreasing base widths because of low base doping and high emitter doping.

However, together with a very small emitter width, reduced external base resistances and reduced parasitics using sophisticated self-alignment techniques⁵⁵ this drift approach has already led to considerable improvement in transistor performance. One has to consider that in a BJT the high doping shrinks (bandgap narrowing)⁵⁶ unfavorably the emitter bandgap, which can be counterbalanced by even a small amount of Ge in the base. From many examples of this technique we mention some devices that combine an epitaxial SiGe base with the double-poly, self-aligned emitter base structure to obtain 74 GHz f_{\max} ,⁵⁷ a 60-Gbit/s multiplexer⁵⁸ and a 42-GHz static frequency divider.⁵⁹ Low parasitic capacitances allow high maximum oscillation frequency at low currents demonstrated by less than 10-fJ power-delay products of an ECL ring oscillator gate.⁶⁰ Integration of high-speed analog⁶¹ and digital integrated circuits for optical networks is demonstrated utilizing super-self-aligned selectively grown SiGe base bipolar transistors.⁶²

Obviously a narrow-bandgap base results in improved emitter efficiency α_E similar to that of the wide-bandgap emitter. The only difference appears at the base-collector interface, which is then a heterojunction. Therefore, this concept is called a

double-heterojunction bipolar transistor (DHBT). The alloy $\text{Si}_{1-x}\text{Ge}_x$ offers the desired smaller bandgap E_g for a silicon-based DHBT:

$$E_g = (1.17 - 0.896x + 0.396x^2) \text{ eV} \quad (T = 4.2 \text{ K, Si unstrained, SiGe compressed}) \quad (2.93)$$

But in pseudomorphic structures on Si (no misfit dislocations) the obtainable bandgap differences ΔE_g are limited to below 150–300 meV because of the lattice mismatch η between SiGe and Si. The lattice constant $a(x)$ of a SiGe increases with Ge content x :

$$a(x) = (0.5431 + 0.01992x + 0.002733x^2) \text{ nm} \quad (2.94)$$

Lattice mismatched material may grow completely strained up to a critical thickness, which strongly decreases with increasing mismatch f . For typical base widths (15–50 nm) in high-frequency HBTs the critical thickness criterion limits the choice of SiGe alloys to Si-rich ($x = 0.3$ – 0.2) ones. Nevertheless, it was demonstrated that the improved emitter efficiency allows a complete inversion of the doping levels in such strained SiGe-DHBTs. Instead of very high doping of the emitter as in bipolar junction transistors (BJT) now, the base is doped to much higher doping levels than the emitter (Fig. 2.28). The DHBT-concept provides, therefore, very thin base layers with acceptable or even improved base sheet resistivities (typically 1–7 k Ω square). This leads to excellent high-frequency properties, low noise, or high current gain (up to 5000 was obtained at room temperature), and high Early voltages. Let us consider now in more detail the doping profile of a SiGe-DHBT (Fig. 2.29). The whole active structure on top of an As-doped subcollector (Sub-C in Fig. 2.29) was grown in one step by molecular beam epitaxy (MBE).⁶⁴ It consists of the n-doped collector (C), the highly p-doped SiGe base (B), followed by a low-doped emitter (LDE) and a highly doped emitter (HDE). A low-doped emitter means lower (10^{18} cm^{-3}) doping than in the BJT. Clear doping-level inversion (base doping, $6 \cdot 10^{19} \text{ cm}^{-3}$) is allowed by a higher Ge content ($x = 0.28$) as used for drift

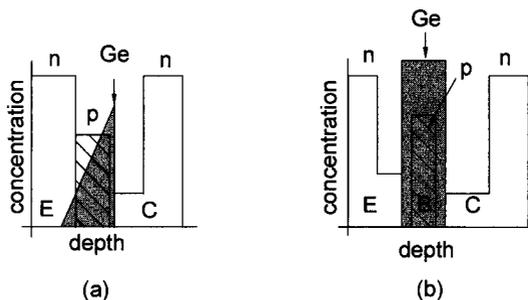


Figure 2.28 Scheme of a drift field transistor with a (a) graded SiGe base of the (b) double heterojunction bipolar transistor.

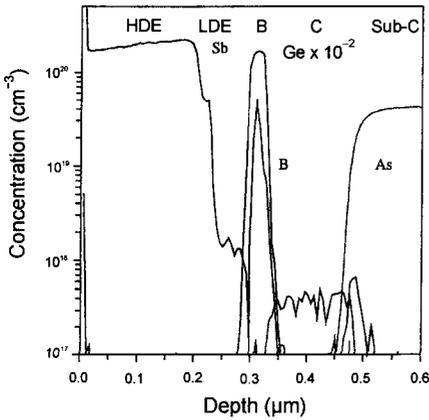


Figure 2.29 Doping profile [secondary-ion mass spectrometry (SIMS)] of a 100-GHz SiGe-DHBT grown by MBE (molecular beam epitaxy). From the surface (left) to the substrate (right) follow the high-doped emitter (HDE), the low-doped emitter (LDE), the boron-doped SiGe base (B), the Sb-doped collector (C), and the As-doped subcollector (Sub-C). (After Kasper et al., Ref. 63.)

transistors. Doping-level inversion avoids tunneling currents that otherwise would degrade the characteristics of a n^+/p^+ junction. The high base doping significantly reduces the Early effect⁵³ because the high base doping suppresses base-width modulation leading to excellent current gain–Early voltage products. For microwave power generation, the high frequencies, low output conductances (high Early voltage), good thermal properties of the Si substrate and the reduced current gain at enhanced temperatures (the term $\exp(\Delta E_g/kT)$ decreases with increasing temperature), let us expect good power-added efficiency (PAE) without thermal shunt precautions in the contacts. In summary, the SiGe-DHBT would be an advantageous device principle for operation with high current gain, at high frequencies, with low output conductance, with low noise and high power-added efficiency. But there are some technological and structural restraints that have to be considered. Technologically, the growth of the complete structure in one step could be advantageous, but otherwise the self-adjustment techniques have to be modified. Structurally, the SiGe base is elastically strained, which raises the question about the stability of the structure. Some principal considerations about stability of strained layers are given in the next section.

2.5.3 The SiGe/Si Material System

The higher lattice constant of a $\text{Si}_{1-x}\text{Ge}_x$ alloy compared to Si causes a lattice mismatch f

$$f = \frac{a(x) - a_{\text{Si}}}{a_{\text{Si}}} \quad (2.95)$$

which, only for thin layers, results in a defect free elastically strained layer. We call this layer *pseudomorphic*. Above a critical thickness t_c strain relaxes by the generation of a misfit dislocation network at the interface film/substrate (relaxed layer). Here we consider only the pseudomorphic layer. Without any proof (for more

information, see, e.g., Ref. 64), we give some rules for strained layers that can help the device engineer to handle this new material.

Pseudomorphic SiGe Layers

1. Up to the critical thickness t_{cm} the strained layer is stable even at high-heat process cycles (naturally, outdiffusion occurs as in other device structures). The critical thickness t_{cm} (the indices c, m mean critical and Matthews–Blakeslee, respectively). Their investigation results in a simple theory about the phenomenon and is given by the implicit relation

$$\left(\frac{t_{cm}}{b}\right)f = 5.78 \cdot 10^{-2} \ln\left(\frac{t_{cm}}{b}\right) \quad (2.96)$$

where b is the Burgers vector length (= 0.384 nm in Si).

2. A pseudomorphic layer covered by a Si cap (e.g., the Si emitter on the SiGe base) is even more stable than a single layer, up to a thickness of approximately $2t_{cm}$
3. At low growth temperatures (e.g., 550°C), kinetic limitation extend the thickness of pseudomorphic layers to larger thicknesses than in equilibrium (metastable regime). For a temperature of 550°C, we note this metastable critical thickness as t_{cp} (the index p means People-Bean, who discovered this phenomenon)

$$\left(\frac{t_{cp}}{b}\right)f^2 = \frac{1}{200} \ln\left(\frac{t_{cp}}{b}\right) \quad (2.97)$$

Even this metastable material when capped with silicon is rather stable against moderate heat cycles.

2.6 SUMMARY AND FUTURE TRENDS

In this chapter the principal function of the bipolar transistor, the inner transistor behavior, and recent technological trends were treated. Sophisticated self-alignment techniques and the incorporation of a SiGe alloy base have extended the performance of bipolar devices and circuits enormously. From a technical viewpoint bipolar circuits are often preferred solutions for analog, mixed analog/digital, and power-consuming peripherals because of their low stable threshold, high transconductance and speed, and comparatively low dynamic power consumption. Otherwise, there are well-known deficiencies with bipolar circuits such as manufacturability and static power consumption. Also, the technically attractive combination of bipolar circuits and CMOS circuits (BiCMOS) suffers from additional costs and technical compromises with respect to the bipolar part. The economic pressure of CMOS technology on the bipolar market is expected to continue.

For future developments, we will concentrate on two topics that will probably define the role bipolar technology can play. The first question concerns manufacturability; the second one addresses the possible impact of silicon-based heterostructures on other technologies. The complexity of bipolar circuits is less than that of CMOS circuits, and the fabrication is rather sophisticated. Which trends are visible that not only promise technical benefits but also simplify fabrication and lower power consumption? At the moment, self-adjustment techniques and epitaxial deposition processes are developed separately. Specially adopted self-alignment techniques together with a single-step epitaxial definition of the complete active transistor structure have an enormous potential to improve competitive manufacturing of bipolar circuits. The highest speed of a bipolar transistor is always obtained at high current densities near the Kirk limit. The decrease of speed with decreasing current is proportional to the emitter–base junction capacity C_E (at least in near ideal box-shaped transistors). This capacity can be decreased by inverted doping levels in HBTs (low-doped emitter, high-doped base). Not the maximum speed, but relatively high speed can then be obtained with lower current levels, which reduces the static power consumption.

In bipolar transistors some Ge doping of the base, at least a few percent Ge to counterbalance the emitter bandgap narrowing (BGN) caused by high doping will be more widely used. The broad availability of Ge techniques will raise the issue of how other Si devices can extend their performance with heterojunctions.⁶⁵ In conclusion we assume that technical solutions such as layers with high Ge content or so-called virtual substrates that defines the elastic strain distribution are available and manufacturable. Then, the most exciting targets would be complementary hetero-field-effect transistors (C-HFET), charge injection transistors (CHINT), and silicon-based optoelectronic receivers for fiber communication. Hetero-field-effect transistors with Ge channels for holes and Si channels for electrons would offer—as the only system we know—symmetrical, high mobility for electrons and holes, around $2000 \text{ cm}^2/(\text{V} \cdot \text{s})$.

ACKNOWLEDGMENT

The author benefited from discussions and with the cooperation of many colleagues, especially U. König, H.-M. Rein, J. Slotboom, and L. Treitinger. Help in manuscript preparation from G. Digele and W. Zhao is also acknowledged.

REFERENCES

1. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.
2. W. G. Neudeck, *The Bipolar Junction Transistor*, 2nd ed., Addison-Wesley, Reading, MA, 1989.
3. W. Von Münch, *Einführung in die Halbleitertechnologie* (in German), B. G. Teubner, Stuttgart, 1993.

4. B. Meinerzhagen and W. L. Engl, "The Influence of the Thermal Equilibrium Approximation of the Accuracy of Classical Two-Dimensional Numerical Modelling of Silicon Submicrometer MOS Transistors," *IEEE Trans. Electron Devices* **ED-35**, 689–697 (1988).
5. C. M. Snowden, *Introduction to Semiconductor Device Modelling*, World Scientific, 1986.
6. Y. Park, D. H. Navon, and T. Tang, "Monte-Carlo Simulation of Bipolar Transistors," *IEEE Trans. Electron Devices* **ED-31**, 1724–1729 (1984).
7. H. H. Ou and T. W. Tang, "Numerical Modelling of Hot Carriers in Submicrometer Silicon BJT's," *IEEE Trans. Electron Devices* **ED-34**, 1533–1539 (1987).
8. W. Fichtner et al., "The Impact of Supercomputers on IC Technology Development and Design," *Proc. IEEE*, **72**, 96–112 (1984); H. K. Gummel, "A Self-Consistent Iterative Scheme for One-dimensional Steady State Transistor Calculations," *IEEE Trans. Electron Devices* **ED-11**, 455–465 (1964); J. W. Slotboom, "Computer-aided Two-dimensional Analysis of Bipolar Transistors," *IEEE Trans. Electron Devices*, **ED-20** (1973).
9. W. L. Engl et al., "Device Modelling," *Proc. IEEE* **71**, 10–33 (1983).
10. J. A. Greenfield, and R. W. Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE Trans. Electron Devices* **ED-27**, 1520–1532 (1980).
11. R. J. Goossens and R. W. Dutton, "Device CAD in the 90's—at the Crossroads," *IEEE Circ. Dev. Mag.* 19–26 (1992).
12. M. Schröter, "Transient and Small-Signal High-Frequency Simulation of Numerical Device Models Embedded in an External Circuit," *COMPEL* **10**, 4, 377–387 (1991).
13. D. J. Roulston, "Numerical Simulation of Bipolar Devices Using BIPOLE: Overview of Numerical Methods and SPICE Parameter Generation," *Proc. NASECODE VII*, Copper Mountain (USA), 1991, pp. 108–110.
14. T. Toyabe, H. Masuda, Y. Aoki, H. Shukuri, and T. Hagiwara, "Three Dimensional Device Simulator CADDETH with Highly Convergent Matrix Solution Algorithm," *IEEE Trans. CAD CAD-4*, 482–488 (1985).
15. B. Neinhüs, P. Graf, S. Decker, and B. Meinerzhagen, "Examination of Transient Drift-Diffusion and Hydrodynamic Modeling Accuracy," *Proc. 27th ESSDERC*, H. Grünbacher, ed., Editions Frontieres, Paris, 1997, pp. 188–191.
16. S. Antognetti and K. Massobrio, *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York, 1987.
17. M. Schröter and H.-M. Rein, "Transit Time of High-Speed Bipolar Transistors in Dependence on Operating Point, Technological Parameters, and Temperature," *Proc. IEEE Bipolar Circuits and Technology Meeting*, Minneapolis, 1989 pp. 250–253.
18. H. C. De Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna/New York, 1990.
19. H. Jeong and J. G. Fossum, "A Charge-Based Large-Signal Bipolar Transistor Model for Device and Circuit Simulation," *IEEE Trans. Electron Devices*, **ED-36**, 124–131 (1989).
20. H. C. De Graaff, "State of the Art in Compact Modeling with Emphasis on Bipolar RF Circuit Design," *Proc. 27th ESSDERC* by H. Grünbacher, ed., Editions Frontieres, Paris, 1997, pp. 14–23.

21. H. N. Gosh, "A Distributed Model of Junction Transistor and Its Application in the Predication of the Emitter-Base Diode Characteristic, Base Impedance, and Pulse Response of the Device," *IEEE Trans. Electron Devices* **ED-12**, 513–531 (1965).
22. H.-M. Rein, and M. Schröter, "Base Spreading Resistance of Square-Emitter Transistors and Its Dependence on Current Crowding," *IEEE Trans. Electron Devices* **ED-36**, 770–773 (1989).
23. M. Schröter, "Simulation and Modeling of the Low-Frequency Base Resistance of Bipolar Transistors in Dependence on Current and Geometry," *IEEE Trans. Electron Devices* **ED-38**, 538–544 (1991).
24. C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, "A Physically Based Mobility Model for Numerical Simulation of Nonplanar Devices," *IEEE Trans. CAD* **7**(11), 1164–1171 (1988); D. B. M. Klaassen, "A Unified Mobility Model for Device Simulation," *Solid-State Electron.* **35**, 953 (1992).
25. M. A. Green, "Intrinsic Concentration, Effective Densities of States and Effective Mass in Silicon," *J. Appl. Phys.* **67**(6), 2944–2954 (1990); D. B. M. Klaassen, J. W. Slotboom, and H. C. de Graaff, "Bandgap Narrowing in Si Bipolar Transistor," *Solid-State Electron.* **35**, 125 (1992).
26. L. L. Liou, T. Jenkins, and C. I. Huang, "Effect of Base Potential Distribution on Thermal Runaway and the Power Limitation of the Heterojunction Bipolar Transistor with Circular Dot Geometry," *Proc. IEEE Hong Kong Electron Device Meeting*, 1996, pp. 49–52.
27. G. Digele, S. Lindenkreuz, and E. Kasper, "Fully Coupled Dynamic Electro-Thermal Simulation," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.* **5**(3), 250–257 (1997).
28. S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, New York, 1984.
29. G. Digele, S. Lindenkreuz, and E. Kasper, "Electro-Thermal Interaction on Circuit Level under the Influence of Packaging," *Proc. 27th Eur. Solid-State Device Research Conf.*, 1997, pp. 460–464.
30. M. Stecher, B. Meinerzhagen, I. Bork, and W. L. Engl, "On the Influence of Thermal Diffusion and Heat Flux on Bipolar Device and Circuit Performance," *Simul. Semicond. Devices Processes* **5**, 49–52 (1993).
31. V. Székely and T. van Bien, "Fine Structure of Heat Flow Path in Semiconductor Devices: A Measurement and Identification Method," *Solid State Electron.* **31**(9), 1363–1368 (1988).
32. H. Baumgärtner and R. Gärtner, *ESD-Elektrostatistische Entladungen*, R. Oldenbourg, München/Vienna, 1997.
33. R. Frankovic and G. H. Bernstein, "Temperature Dependence of Electromigration Threshold in Cu," *J. Appl. Phys.* **81**, 3 (1997).
34. L. Treitinger and M. Miura-Mattausch, *Ultra Fast Silicon Bipolar Technology*, Springer-Verlag, Berlin/Heidelberg, 1988.
35. E. Kasper, H.-J. Herzog, and K. Wörner, *J. Crystal Growth* **81**, 458–462, (1987).
36. T. Onai, E. Ohue, M. Tanabe, and K. Washio, "Self-Aligned Metal/IDP Si Bipolar Technology Featuring 14ps/70GHz," *IEDM*, 1995, *Tech. Digest Papers*, 1995, 699–702 pp.

37. M. Schröter, "A Survey of Present Compact Models for High-Speed Bipolar Transistors," *Frequenz* **47**, 178–190 (1993).
38. M. Schröter, "Physikalische Modelle für schnelle Silizium-Bipolartransistor—Eine Vergleichende Übersicht," *Habilitationsschrift, Ruhr-Universität Bochum*, 1993.
39. E. Ohue, Y. Kiyota, T. Onai, M. Tanabe, and K. Washio, "100-GHz f_T Si Homojunction Bipolar Technology," *Proc. VLSI Conf.*, 1996, pp. 2/9–3/9.
40. K. Aufinger, J. Böck, T. F. Meister, and J. Popp, "Noise Characteristics of Transistors Fabricated in an Advanced Silicon Bipolar Technology," *IEEE Trans. Electron Devices* **43**(9), 1533–1538 (1996).
41. A. Van der Ziel and G. Bosman, "Accurate Expression for the Noise Temperature of Common Emitter Microwave Transistors," *IEEE Trans. Electron Devices* **ED-31**, 1280–1283 (1984).
42. R. J. Hawkins, "Limitations of Nielsens and Related Noise Equations Applied to Microwave Bipolar Transistors, and a New Expression for the Frequency and Current Dependent Noise Figure," *Solid-State Electron.* **20**, 191–196 (1977).
43. P. K. Tien, "Propagation Delay in High Speed Silicon Bipolar and GaAs HBT Digital Circuits," *Int. J. High Speed Electron.* **1**, 101–124 (1990).
44. H.-M. Rein, "Hochgeschwindigkeits-Schaltungen in Silizium-Bipolartechnologie" (in German), *Informationstechnik* **34**, 209–219 (1992).
45. H.-M. Rein, and M. Möller, "Design Considerations for Very-High-Speed Si-Bipolar IC's Operating up to 50 Gb/s," *IEEE, Solid-State Circ.* **31**(8), 1076–1090 (1996).
46. M. Möller, H.-M. Rein, A. Felder, J. Popp, and J. Böck, "50 Gb/s Time-Division Multiplexer in Si-Bipolar Technology," *Electron. Lett.* **31**, 17, 1431–1433 (1995).
47. J. Böck, A. Felder, T. F. Meister, M. Franosch, K. Aufinger, M. Wurzer, R. Schreiter, S. Boguth, and L. Treitinger, "A 50 GHz Implanted Base Silicon Bipolar Technology with 35 GHz Static Frequency Divider," *Symp. VLSI Tech. Digest Papers*, 1996, pp. 108–109.
48. M. Wurzer, A. Felder, J. Popp, and J. Böck, "20 Gb/s Silicon bipolar XOR Gate for Optical Transmission Systems," *Proc. 21st ECOC'95*, Brussels, 1995, pp. 1035–1037.
49. K. Washio, E. Ohue, M. Tanabe, and T. Onai, "Self-Aligned Metal/IDP Si Bipolar Technology with 12-ps ECL and 45-GHz Dynamic Frequency Divider," *Proc. ESSDERC*, Baccaroni/Rudan, ed., Editions Frontieres, Paris, 1996, pp. 807–810.
50. A. Gruhle, "SiGe Heterojunction Bipolar Transistors," in J. F. Luy and P. Russer, eds., *Silicon-Based Millimeter-Wave Devices*, Springer, Berlin, 1994.
51. J. Sturm, "Si/SiGe/Si Heterojunction Bipolar Transistor," in E. Kasper ed., *Properties of Strained and Relaxed Silicon Germanium*, EMIS Datareviews Series, Vol. 12, INSPEC, IEE, London, 1995.
52. H. Krömer, "Two Integral Relations Pertaining to the Electron Transport through a Bipolar Transistor with a Nonuniform Energy Gap in the Base Region," *Solid-State Electron.* **28**, 1101–1103 (1985).
53. E. J. Prinz and J. Sturm, "Analytical Modeling of Current Gain–Early Voltage Products in SiGe HBTs," *Proc. IEDM91*, 1991, pp. 853–856.
54. E. Crabbe et al., *Proc. IEDM 93, Tech. Digest, Papers* 1993, p. 83.
55. C. Y. Chang and S. M. Sze, *ULSI Technology*, McGraw-Hill, New York, 1996.

56. D. B. M. Klassen, J. W. Slotboom, and H. C. de Graaff, "Unified Apparent BGN in n- and p-Type Silicon," *Solid-State Electron* **35**, 125–129 (1992).
57. T. F. Meister, H. Schäfer, M. Franosch, W. Molzer, K. Aufinger, U. Scheler, C. Walz, M. Stolz, S. Boguth, and J. Böck, "SiGe Base Bipolar Technology with 74 GHz f_{\max} and 11 ps Gate Delay," *Proc. IEDM* **95**, 1995, pp. 739–742.
58. M. Möller, H.-M. Rein, A. Felder, and T. F. Meister, "60 Gbit/s Time-Division Multiplexer in SiGe-Bipolar Technology with Special Regard to Mounting and Measuring Technique," *Electron. Lett.* **33**(8), 679–680 (1997).
59. M. Wurzer, T. F. Meister, H. Schäfer, H. Knapp, J. Böck, R. Stengl, K. Aufinger, M. Franosch, M. Rest, M. Möller, H.-M. Rein, and A. Felder, "42GHz Static Frequency Divider in a Si/SiGe Bipolar Technology," *Proc. ISSCC97*, 1997, pp. 122–212.
60. M. Kondo, K. Oda, E. Ohue, H. Shimamoto, M. Tanabe, T. Onai, and K. Washio, "Sub-10-fJ ECL/68- μ A 4.7-GHz Divider Ultra-Low-Power SiGe Base Transistors with a Wedge-Shaped CVD-SiO₂ Isolation Structure and a BPSG-Refilled Trench," *IEDM'96 Technical Digest*, 1996, pp. 245–248.
61. D. L. Hareme et al., "Si/SiGe Epitaxial Base Transistors," *IEEE Trans.* **ED-42**, 455–482 (1995).
62. F. Sato, H. Tezuka, M. Soda, T. Hashimoto, T. Suzaki, T. Tatsumi, T. Morikawa, and T. Tashiro, "A 2.4Gb/s Receiver and a 1:16 Demultiplexer in One Chip Using a Super Self-Aligned Selectively Grown SiGe Base (SSSB) Bipolar Transistor," *IEEE Solid-State Circ.* **31**(10), 1451–1457 (1996).
63. E. Kasper, H. Kibbel, H.-J. Herzog, and A. Gruhle, "Growth of 100GHz SiGe-Heterobipolar Transistor (HBT) Structures," *Jpn. J. Appl. Phys.* **33**, 2415–2418 (1994).
64. E. Kasper, *Properties of Strained and Relaxed Silicon Germanium*, in E. Kasper, ed., EMIS Datareviews Series, Vol. 12, INSPEC, IEE, London, 1995.
65. E. Kasper, "Silicon Germanium Heterojunction—Extending the Performance of Si Devices," *Curr. Opin. Solid State Mat. Sci.* **2**, 48–53 (1997).

PROBLEMS

- 2.1 The collector current I_C in the saturation region of operation is decreased compared to its value in the active region (constant I_B). Calculate the output characteristics in common emitter configuration for a box shaped transistor (one-dimensional solution) neglecting the Early effect.
- 2.2 The inner base resistance r_{bi} causes current crowding in large area transistors. Give a relation for the geometric dimensions of the desired transistor without current crowding.
- 2.3 Reformulate the equations for current gain α_0 assuming that the recombination lifetime is governed by Auger recombination. For a numerical example choose for the diffusion length $D = D_p = D_n = 4 \text{ cm}^2/\text{s}$, for the Auger coefficient $G = 10^{-31} \text{ cm}^6/\text{s}$, and for the base Gummel number $N_{BWB} = 3 \cdot 10^{13}/\text{cm}^2$.

- 2.4** The neutral base modulation causes an output conductance that may be characterised by the Early voltage V_{EA} . Consider the influence of the base doping on the Early voltage and on the Early voltage current gain product $V_E\beta_0$.
- 2.5** The emitter base junction is forward-biased. The injection of holes into the emitter is calculated as $J_p = q \cdot (D_p/L_p) \cdot (N_{i,E}^2/N_E)\exp(V/V_T)$ for a uniformly doped emitter. In an HBT the emitter is often designed as a thin low-doped emitter ($w_1 \ll L_1, N_1$) followed by a thick high-doped emitter contact ($w_2 \ll L_2, N_2$). Calculate the back-injected hole current for this structure.
- 2.6** The transit frequency f_T of a heterobipolar transistor (HBT) is increased by a higher doped and thinner base than possible in silicon bipolar transistors (BJT). Investigate the improvements in transit frequency and base resistivity by an HBT.
- 2.7** At high frequencies the currents i_E, i_B, i_C are out of phase. Give phase relationships as function of frequency.
- 2.8** The common emitter current gain β decreases toward unity at the transit frequency $f_T, |\beta(f_T)| = 1$. Explain the current dependence of the transit frequency and compare the result with Figure 2.13.
- 2.9** A good high-frequency transistor should exhibit both a high transit frequency f_T and a high maximum oscillation frequency f_{max} . The maximum oscillation frequency f_{max} is strongly dependent on the lateral transistor dimensions. Explain the critical geometrical dimensions and give a relation for the geometry of a box shaped transistor with $f_T = f_{max}$. (*Hint*: Neglect the weak lateral geometry dependence of f_T .)
- 2.10** Explain the essential features of an SiGe-HBT doping profile in the example of Figure 2.29.

MOSFET Fundamentals

HON-SUM PHILIP WONG

IBM Thomas J. Watson Research Center
Yorktown Heights, NY

3.1 INTRODUCTION

The MOSFET has been the mainstream microelectronics device technology from the VLSI era of the 1980s through the ULSI era of the 1990s. The dominance of CMOS technologies is expected to continue well into the twenty-first century.¹ Basic MOSFET operations have been described in many excellent textbooks such as Muller and Kamins² and Sze,³ as well as monographs such as Ko.⁴ This chapter provides an introduction to the fundamentals of MOSFETs in the ULSI era.

Section 3.2 outlines the derivations of the drain current models at various levels of abstractions and approximations. It provides the link between device physics and circuit behavior. Discussions on “ideal” MOSFET characteristics provide the basis on which nonideal effects such as short-channel effects, hot-carrier effects, and other parasitic effects are developed in subsequent sections of the chapter.

The ULSI era is punctuated by the expectation that improvements in circuit and device performances can be achieved by device scaling.⁵ The concept and practice of device scaling are described in Section 3.3. Historical trends of CMOS technologies and projections into the future are covered in this section. The control of short-channel effects is of paramount importance in the design of future generations of devices. The basic concepts of short-channel behavior are described in Section 3.4. The section starts with simple models and progresses into more elaborate models that describes the short-channel behavior for modern devices. This section concludes with a brief summary of the concept of the characteristic scale length of MOSFETs. This characteristic scale-length concept is very useful in evaluating various device design options for devices that are structurally significantly distinct from conventional bulk MOSFETs (see Section 3.7).

Transport properties of carriers in the channel dominate the current driving capability of MOSFETs. This is covered in Section 3.5. The dependence of carrier

mobility and velocity on normal electric fields as well as on longitudinal electric fields are discussed in this section. Parasitic effects that cause the MOSFET to deviate from first-order models are discussed in Section 3.6. These various parasitic effects become significant effects as devices are scaled to smaller dimensions. Various nonconventional device structures have been proposed to alleviate the short-channel effects and parasitic effects of MOSFETs to allow devices to scale down to smaller geometries. These device structures are reviewed in Section 3.7. The evolution of the bulk CMOS is first reviewed to provide a background against which other, more revolutionary, device structures are compared. Most of the new device structures are based on SOI (silicon-on-insulator) or its variants. Finally, this chapter closes with a discussion on the limits of CMOS device scaling. There is a general feeling that, as CMOS technology approaches the turn of the century into the GSI (gigascale integration) era with nanometer-sized gate lengths, CMOS will face formidable challenges to meet its historical pace of performance improvement, and the potential for scaling CMOS will come to an end. This last section provides an overview of the challenges ahead and some suggested solutions.

3.2 MOSFET DEVICE PHYSICS: FIRST-ORDER MODELS

In this section, we develop the first-order models for the MOSFET. Starting from the basic equations of Poisson's equation and the current-continuity equations from Maxwell's equations, we proceed to make various approximations to arrive at a set of simple analytic equations that describe the drain current of the MOSFET. These simple descriptions of the MOSFET are then refined in subsequent sections to include other important physical effects.

Figure 3.1 shows the cross section of an n-channel MOSFET with the voltage terminals and the space coordinates defined. The substrate voltage (V_B) is arbitrarily set to be the voltage reference. The one-dimensional energy band diagram perpendicular to the Si/SiO₂ interface is shown in Figure 3.2.

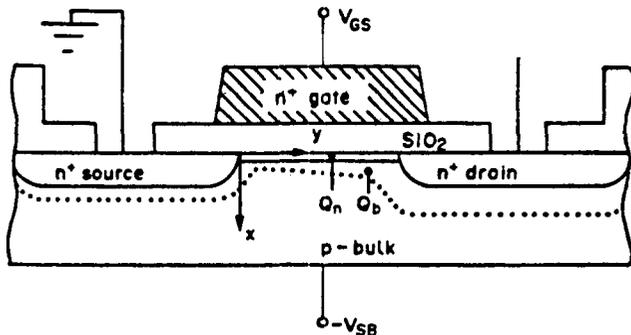


Figure 3.1 Cross section of an n-channel MOSFET with the voltage terminals and the space coordinates defined.

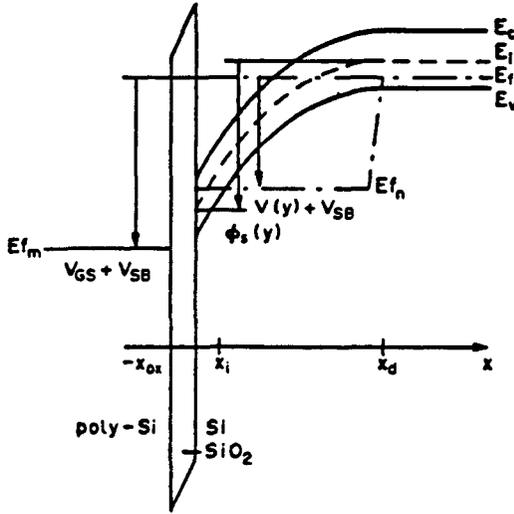


Figure 3.2 Energy-band diagram in the direction perpendicular to the Si/SiO₂ interface corresponding to Figure 3.1.

3.2.1 One-Dimensional Drain Current Model

Assuming that the current flow is essentially one-dimensional from the source to the drain (this assumption will be inadequate for short-channel devices), the drain current is described by

$$J_n(x, y) = q\mu_n nE + qD_n \frac{dn}{dy} \quad (3.1)$$

Assuming a Boltzmann distribution for the minority carrier (electrons for n-channel and holes for p channel),

$$n = n_{pB} \exp^{q(\phi - V - V_{SB})/kT} \quad (3.2)$$

and

$$\frac{dn}{dy} = \frac{q}{kT} n \left(\frac{d\phi}{dy} - \frac{dV}{dy} \right) \quad (3.3)$$

Together with the Einstein relation, $(kT/q)\mu_n = D_n$, Eq. 3.1 reduces to

$$J_n(x, y) = -q\mu_n n \frac{dV}{dy} \quad (3.4)$$

Integrating over all the three spatial dimensions, one obtains^{6,7}

$$\int_0^W dz \int_0^L dy \int_0^{x_w} dx J_n(x, y) = -q \int_0^W dz \int_0^{V_{DS}} dV \int_0^{x_w} dx \mu_n(x, y, z) n(x, y, z) \quad (3.5)$$

and

$$I_D = q \left(\frac{W}{L} \right) \int_0^{V_{DS}} dV \int_0^{x_w} dx \mu_n(x, y, z) n(x, y, z) \quad (3.6)$$

with the mobility approximated by a weighted average

$$\mu_n(y) \simeq \frac{\int_0^{x_w} \mu_n n dx}{\int_0^{x_w} n dx} \quad (3.7)$$

the drain current can then be written as^{6,7}

$$I_D = q \left(\frac{W}{L} \right) \int_0^{V_{DS}} dV \mu_n \int_0^{x_w} n dx = - \left(\frac{W}{L} \right) \int_0^{V_{DS}} dV (y) \mu_n(y) Q_n(y) \quad (3.8)$$

where

$$Q_n(y) = -q \int_0^{x_w} n(x, y) dx \quad (3.9)$$

To obtain the drain current, we need to know the two-dimensional (2D) charge density Q_n in Eq. 3.8, which is derived in the next section.

3.2.2 Charge Density: One-Dimensional Poisson Equation Solution

We will now derive the 2D charge density Q_n using the solution to the Poisson equation:

$$\nabla^2 \phi(x, y, z) = \frac{-\rho(x, y, z)}{\epsilon_s} \quad (3.10)$$

$$\rho(x, y, z) = q(N_D^+ - N_A^- + p_p - n_p) \quad (3.11)$$

where ϵ_s is the dielectric constant of the semiconductor, the subscript p in the electron density (n_p) and hole density (p_p) refers to the respective quantity in a p-type semiconductor, N_A^- is the ionized acceptor-dopant density, and N_D^+ is the ionized donor-dopant density.

Referring to Figure 3.2 and assuming a three-dimensional (3D) density of states and the Maxwell–Boltzmann statistics for the carrier density, the electron and hole densities are described by

$$n_p = n_{p0} e^{-q(V_{SB}+V)/kT} = n_{pB} e^{q\phi/kT} e^{-q(V+V_{SB})/kT} = n_{pB} e^{q(\phi-V-V_{SB})/kT} \quad (3.12)$$

$$p_p = p_{p0} = p_{pB} e^{-q\phi/kT} \quad (3.13)$$

where n_{p0} is the electron concentration in p-type semiconductor at equilibrium, n_{pB} is the electron concentration in p-type semiconductor in the bulk (i.e., no band ending or $\phi = 0$), and p_{p0} and p_{pB} are the corresponding values for holes.

Assuming a one-dimensional solution perpendicular to the Si/SiO₂ interface (which is essentially the same one-dimensional current flow assumption as in the gradual-channel approximation), Eqs. 3.10–3.11 reduce to

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_s} [N_D^+ - N_A^- + p_{pB} e^{-(q\phi/kT)} - n_{pB} e^{q(\phi-V-V_{SB})/kT}] \quad (3.14)$$

The boundary conditions for solving Eq. 3.14 are

1. $\phi(x \rightarrow \infty) = 0$
2. $d\phi/dx|_{x \rightarrow \infty} = 0$

For SOI, the Poisson equation must be solved with boundary conditions unique to its geometry and the preceding boundary conditions may not be valid. When $\rho = 0$, $\phi = 0$, and $V + V_{SB} = 0$, charge neutrality dictates that

$$N_D^+ - N_A^- + p_{pB} - n_{pB} = 0 \quad (3.15)$$

Substituting Eq. 3.15 into Eq. 3.14, we obtain

$$\frac{d^2\phi}{dx^2} = -\frac{q}{\epsilon_s} p_{pB} [e^{-(q\phi/kT)} - 1 - \frac{n_{pB}}{p_{pB}} (e^{q(\phi-V-V_{SB})/kT} - 1)] \quad (3.16)$$

Considering that

$$\frac{1}{2} \frac{\partial}{\partial x} \left[\left(\frac{\partial \phi}{\partial x} \right)^2 \right] = \frac{\partial \phi}{\partial x} \frac{\partial^2 \phi}{\partial x^2} \quad (3.17)$$

we obtain

$$\frac{1}{2} \int_{x=\infty}^x d \left(\frac{d\phi}{dx} \right)^2 = \int_{x=\infty}^x -\frac{q}{\epsilon_s} p_{pB} \left[e^{-(q\phi/kT)} - 1 - \frac{n_{pB}}{p_{pB}} (e^{q(\phi-V-V_{SB})/kT} - 1) \right] d\phi \quad (3.18)$$

$$\begin{aligned} \frac{1}{2} (-\mathcal{E}_s)^2 &= \frac{qp_{pB}kT}{\epsilon_s q} \left[e^{-(q\phi/kT)} + \frac{q\phi}{kT} \right. \\ &\quad \left. - 1 + \frac{n_{pB}}{p_{pB}} \left(e^{q\phi/kT} e^{-[q(V+V_{SB})/kT]} - \frac{q\phi}{kT} - e^{-[q(V+V_{SB})/kT]} \right) \right] \quad (3.19) \end{aligned}$$

The extrinsic Debye length is given by

$$L_D = \sqrt{\frac{kT\epsilon_s}{q^2 p_{pB}}} \quad (3.20)$$

Substituting $p_{pB} = N_A^-$, Eq. 3.19 becomes

$$\begin{aligned} \mathcal{E}_s = \frac{kT\sqrt{2}}{q L_D} \left[e^{-(q\phi/kT)} + \frac{q\phi}{kT} \right. \\ \left. - 1 + \frac{n_{pB}}{p_{pB}} \left(e^{q\phi/kT} e^{-[q(V+V_{SB})/kT]} - \frac{q\phi}{kT} - e^{-[q(V+V_{SB})/kT]} \right) \right]^{1/2} \end{aligned} \quad (3.21)$$

Defining the F function as

$$\begin{aligned} F(\phi, V, V_{SB}, \phi_F) = \left[e^{-(q\phi/kT)} + \frac{q\phi}{kT} \right. \\ \left. - 1 + \frac{n_{pB}}{p_{pB}} \left(e^{q\phi/kT} e^{-[q(V+V_{SB})/kT]} - \frac{q\phi}{kT} - e^{-[q(V+V_{SB})/kT]} \right) \right]^{1/2} \end{aligned} \quad (3.22)$$

the surface electric field \mathcal{E}_s can be expressed as

$$\begin{aligned} \mathcal{E}_s &= \frac{kT\sqrt{2}}{q L_D} F(\phi, V, V_{SB}, \phi_F) \\ &= \frac{\lambda C_{ox}}{\epsilon_s} \sqrt{\frac{kT}{q}} F(\phi, V, V_{SB}, \phi_F) \end{aligned} \quad (3.23)$$

where $\lambda = \sqrt{2\epsilon_s q N_A} / C_{ox}$ and C_{ox} is the gate oxide capacitance. Gauss' law requires that the total charge per unit area in the semiconductor $Q_s = -\epsilon_s \mathcal{E}_s$ and

$$Q_s = -\lambda C_{ox} \sqrt{\frac{kT}{q}} F(\phi, V, V_{SB}, \phi_F) \quad (3.24)$$

The mobile charge per unit area (Q_n) is given by integrating the charge per unit volume over the depletion layer

$$\begin{aligned} Q_n &= q \int_{x=x_w}^{x=0} n_p dx \\ &= q \int_{\phi=0}^{\phi_s} \frac{n_p d\phi}{d\phi/dx} \end{aligned} \quad (3.25)$$

where ϕ_s is the potential at the surface ($x = 0$). Substituting $n_p = n_{pB}e^{q(\phi-V-V_{SB})/kT}$ into Eq. 3.25, we have

$$Q_n = q \int_{\phi_F}^{\phi_s} d\phi \left[\frac{n_{pB}e^{q(\phi-V-V_{SB})/kT}}{\sqrt{2\varepsilon_s q p_{pB}}} \sqrt{\frac{kT}{q}} F(\phi, V, V_{SB}, \phi_F) \right] \quad (3.26)$$

$$= -\frac{1}{2} \lambda C_{ox} \sqrt{\frac{q}{kT}} G(\phi, V, V_{SB}, \phi_F)$$

where

$$G(\phi, V, V_{SB}, \phi_F) = \int_{\phi_F}^{\phi_s} \frac{e^{q(\phi-2\phi_F-V-V_{SB})/kT}}{F(\phi, V, V_{SB}, \phi_F)} \quad (3.27)$$

The total charge in the semiconductor Q_s , and the mobile charge in the inversion layer Q_n must be related to the applied terminal voltages in order to relate the terminal currents to the terminal voltages. Referring to Figures 3.2 and 3.1 and taking a one-dimensional slice along the x direction, the terminal voltages and the potential at the Si/SiO₂ interface (ϕ_s)

$$V_G - V_B - V_{FB} = V_{ox}(y) + \phi_s(y) \quad (3.28)$$

$$V_{GB} = V_{FB} + V_{ox}(y) + \phi_s(y)$$

and

$$Q_s(y) = Q_n(y) + Q_d(y) = -V_{ox}(y)C_{ox} \quad (3.29)$$

where V_{FB} is the flat-band voltage, V_{ox} is the voltage drop across the gate oxide, and Q_d is the charge per unit area in the depletion region of the semiconductor.

3.2.3 Drain Current in the Strong Inversion Approximation

The depletion charge Q_d can be obtained by solving the Poisson equation

$$Q_d(y) = -\sqrt{2\varepsilon_s q N_A \phi_s} \quad (3.30)$$

Substituting Eqs. 3.29 and 3.30 into Eq. 3.28, the mobile charge per unit area (Q_n) is

$$Q_n(y) = -C_{ox}[V_{GB} - V_{FB} - \phi_s(y)] + \sqrt{2\varepsilon_s q N_A \phi_s(y)} \quad (3.31)$$

$$= -C_{ox}[V_{GB} - V_{FB} - 2\phi_{Fp} - V(y)] + \sqrt{2\varepsilon_s q N_A (2\phi_{Fp} + V(y))}$$

with the assumption that $\phi_s = 2\phi_{Fp} + V(y)$. This assumption is only approximate in that it assumes that once the inversion layer is formed, the surface potential ϕ_s is pinned at $2\phi_{Fp}$, which is not valid in many cases.⁷

The drain current can be obtained by substituting Eq. 3.31 into Eq. 3.8 and integrating:

$$I_D = \mu_n \left(\frac{W}{L} \right) \left\{ C_{ox} \left[V_{GB} - V_{FB} - 2\phi_{Fp} - \frac{1}{2}(V_{DB} + V_{SB}) \right] (V_{DB} - V_{SB}) - \frac{2}{3} \sqrt{2\varepsilon_s q N_A} [(2\phi_{Fp} + V_{DB})^{3/2} - (2\phi_{Fp} + V_{SB})^{3/2}] \right\} \quad (3.32)$$

For the case where the source and substrate terminals are grounded ($V_S = V_B = 0$), we obtain

$$I_D = \mu_n \left(\frac{W}{L} \right) \left\{ C_{ox} \left[V_G - V_{FB} - 2\phi_{Fp} - \frac{1}{2}V_D \right] V_D - \frac{2}{3} \sqrt{2\varepsilon_s q N_A} [(2\phi_{Fp} + V_D)^{3/2} - (2\phi_{Fp})^{3/2}] \right\} \quad (3.33)$$

In the limit of $V_D \ll 2\phi_{Fp}$, Eq. 3.33 reduces to a simple expression usually found in circuit applications:

$$I_D = \mu_n \left(\frac{W}{L} \right) \left\{ C_{ox} \left[V_G - V_{FB} - 2\phi_{Fp} - \frac{1}{2}V_D \right] V_D - 2\sqrt{\varepsilon_s q N_A} \phi_{Fp} V_D + \frac{1}{4} \sqrt{\frac{\varepsilon_s q N_A}{\phi_{Fp}}} V_D^2 \right\} \quad (3.34)$$

The threshold voltage (V_T) is often defined as the gate to substrate voltage (V_{GB}), where $\phi_s|_{y=0} = 2\phi_{Fp}$ and $Q_n = 0$ at $y = 0$:

$$V_T = V_{FB} + 2\phi_{Fp} + V_{SB} + \frac{\sqrt{2\varepsilon_s q N_A (2\phi_{Fp} + V_{SB})}}{C_{ox}} \quad (3.35)$$

The drain current equation, Eq. 3.34, can be expressed as

$$I_D = \mu_n \left(\frac{W}{L} \right) C_{ox} \left[(V_G - V_T) V_D - \left(\frac{1}{2} + \frac{\frac{1}{4} \sqrt{\frac{\varepsilon_s q N_A}{\phi_{Fp}}}}{C_{ox}}} \right) V_D^2 \right] \quad (3.36)$$

In the linear region where $V_D \ll V_G - V_T$, Eq. 3.36 reduces to

$$I_D = \mu_n \left(\frac{W}{L} \right) C_{\text{ox}} (V_G - V_T) V_D \quad (3.37)$$

The small-signal conductances (g_d) and transconductances (g_m) are given by

$$\begin{aligned} g_d &= \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G} \\ &= \mu_n \left(\frac{W}{L} \right) C_{\text{ox}} (V_G - V_T) \end{aligned} \quad (3.38)$$

$$\begin{aligned} g_m &= \left. \frac{\partial I_D}{\partial V_D} \right|_{V_D} \\ &= \mu_n \left(\frac{W}{L} \right) C_{\text{ox}} V_D \end{aligned} \quad (3.39)$$

As the drain voltage is increased, the channel will be pinched off. From Eq. 3.31, if we set $Q_n(y=L) = 0$ when $V(y=L) = V_{D\text{sat}}$, we get

$$\begin{aligned} V_{D\text{sat}} &= (V_{GB} - V_{FB} - 2\phi_{Fp}) + \frac{\epsilon_s q N_A}{C_{\text{ox}}^2} \left[1 - \sqrt{1 + \frac{4(V_{GB} - V_{FB} - 4\phi_{Fp})}{(2\epsilon_s q N_A / C_{\text{ox}}^2)}} \right] \\ &= V_{GB} - V_T - V_x \end{aligned} \quad (3.40)$$

where

$$V_x = \frac{\sqrt{4\epsilon_s q N_A \phi_{Fp}}}{C_{\text{ox}}} + \frac{\epsilon_s q N_A}{C_{\text{ox}}^2} \left[1 - \sqrt{1 + \frac{4(V_{GB} - V_{FB} - 4\phi_{Fp})}{(2\epsilon_s q N_A / C_{\text{ox}}^2)}} \right] \quad (3.41)$$

Substituting Eq. 3.40 into Eq. 3.36 with $V_D = V_{D\text{sat}}$, the drain current is

$$I_D = \frac{1}{2} \mu_n C_{\text{ox}} \left(\frac{W}{L} \right) (V_G - V_T)^2 \left[1 - \left(\frac{V_x}{V_G - V_T} \right)^2 \right] \quad (3.42)$$

For $[V_x / (V_G - V_T)]^2 \ll 0$ (i.e. V_G is large or N_A is small)

$$I_D = \frac{1}{2} \mu_n C_{\text{ox}} \left(\frac{W}{L} \right) (V_G - V_T)^2 \quad (3.43)$$

and the small-signal transconductance is

$$\begin{aligned} g_m &= \left. \frac{\partial I_D}{\partial V_D} \right|_{V_D} \\ &= \mu_n \left(\frac{W}{L} \right) C_{\text{ox}} (V_G - V_T) \end{aligned} \quad (3.44)$$

3.2.4 Drain Current in the Subthreshold Region

In the subthreshold region ($\phi_F + V - V_{SB} < \phi_s < 2\phi_F + V + V_{SB}$), the drain current is mainly a diffusion current,³ as opposed to a drift current in the strong-inversion region. The subthreshold drain current can be derived in two ways:

1. Expansion in Taylor's series about the $\phi_F + V + V_{SB} < \phi_s < 2\phi_F + V + V_{SB}$ point to obtain a simplified expression of the mobile charge density $Q_n(\phi_s)$ for Eq. 3.8.
2. Consider the diffusion current component alone.

Sze³ has already derived the subthreshold current using the second method. Here we use the Taylor's series expansion method to derive the subthreshold current.

Using Eq. 3.24 and the depletion approximation $Q_b = -\lambda C_{\text{ox}} \sqrt{\phi_s - (kT/q)}$, the mobile charge density is given by

$$\begin{aligned} Q_n &= Q_s - Q_b \\ &= -\lambda C_{\text{ox}} \sqrt{\phi_s - \frac{kT}{q}} \left\{ \left[1 + \frac{1}{2} \frac{kT/q}{\sqrt{\phi_s - kT/q}} \exp \frac{q(\phi_s - 2\phi_F - V - V_{SB})}{kT} + \dots \right] - 1 \right\} \\ &= -\frac{\frac{1}{2} \lambda C_{\text{ox}}}{\sqrt{\phi_s - kT/q}} \left(\frac{kT}{q} \right) \exp \frac{q(\phi_s - 2\phi_F - V - V_{SB})}{kT} \end{aligned} \quad (3.45)$$

Substituting Eq. 3.45 into Eq. 3.8, the subthreshold current is

$$I_D = \frac{1}{2} \mu_n \left(\frac{W}{L} \right) \left(\frac{kT}{q} \right)^2 \frac{\sqrt{2\epsilon_s q N_A}}{\sqrt{\phi_s - kT/q}} \exp \frac{q(\phi_s - 2\phi_F - V_{SB})}{kT} [1 - e^{-(qV_{DS}/kT)}] \quad (3.46)$$

Equation 3.46 shows that for $V_{DS} > 3kT/q$, the drain current is independent of the drain voltage. This agrees with the diffusion description where the collection of the carrier is efficient for sufficiently large drain voltages, and the current is dependent solely on the carrier density gradient along the channel (from source to drain) and not on the collector efficiency. Note also that the drain current depends exponentially on the surface potential ϕ_s in the subthreshold region.

We need to express the surface potential ϕ_s in terms of the applied terminal voltages to relate the drain current of Eq. 3.46 to the terminal voltages. Starting from Eqs. 3.28 and 3.29, and introducing a finite interface charge density per unit area of $Q_{it}(\phi_s)$, the gate voltage and the surface potential is related by

$$V_{GB} = V_{FB} + \phi_s - \frac{Q_s}{C_{ox}} - \frac{Q_{it}}{C_{ox}} \quad (3.47)$$

If we expand the surface potential (ϕ_s) around the strong inversion approximation value of $\phi_s = 2\phi_F + V + V_{SB}$, Eq. 3.47 becomes

$$\begin{aligned} V_{GS} &= V_T + \phi_s \left(1 + \frac{C_D}{C_{ox}} + \frac{C_{it}}{C_{ox}} \right) - (2\phi_F + V_{SB}) \left(1 + \frac{C_D}{C_{ox}} + \frac{C_{it}}{C_{ox}} \right) - V \frac{C_{it}}{C_{ox}} \\ &= V_T + n\phi_s - n(2\phi_F + V_{SB}) - (n - m)V \end{aligned} \quad (3.48)$$

leading finally to

$$\phi_s = \frac{V_{GS} - V_T}{n} + 2\phi_F + V_{SB} + \left(\frac{n - m}{n} \right) V \quad (3.49)$$

with

$$n = 1 + \frac{C_D}{C_{ox}} + \frac{C_{it}}{C_{ox}} \quad (3.50)$$

$$m = 1 + \frac{C_D}{C_{ox}} \quad (3.51)$$

$$C_{it}(\phi_s) = - \frac{\partial Q_{it}}{\partial \phi_s} \quad (3.52)$$

$$C_D(\phi_s) = - \frac{\partial Q_b}{\partial \phi_s} = \frac{\lambda C_{ox}}{2} \frac{1}{\sqrt{\phi_s - kT/q}} \quad (3.53)$$

Putting Eq. 3.49 into Eq. 3.46, the drain current in the subthreshold region is

$$I_D = \mu_n \left(\frac{W}{L} \right) \left(\frac{kT}{q} \right)^2 C_D(\phi_s) e^{q(V_{GS} - V_T)/nKT} \left(\frac{n}{m} \right) [1 - e^{-(qmV_{DS}/nKT)}] \quad (3.54)$$

The inverse subthreshold slope (sometimes called the *S factor*) is

$$S = \left(\frac{\partial \log I_D}{\partial V_{GS}} \right)^{-1} = \ln(10)n \left(\frac{kT}{q} \right) \quad (3.55)$$

3.3 DEVICE SCALING

3.3.1 Scaling Theory

The concept of device scaling was first introduced by Dennard et al.⁵ and further developed in the early 1980s.⁸ Figure 3.3 illustrates the concept of device scaling whereby the larger device is scaled down by a factor α to yield the smaller device. The theoretical basis of scaling is formulated by considering the Poisson and the current continuity equations:⁸

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = -\frac{q}{\epsilon_s} (N_D - N_A + p - n) \tag{3.56}$$

$$\nabla J_n = 0 \tag{3.57}$$

where the drift diffusion current J_n is given by

$$J_n(x, y, z) = -q\mu_n n \nabla \phi + qD_n \nabla n \tag{3.58}$$

Consider, for the moment, only the electrostatics of device design. In subthreshold conditions, the electron concentration contributes negligibly to the space charge on the right-hand side of Eq. 3.56. Therefore, Eqs. 3.56 and 3.57 can be decoupled and we can refer to Eq. 3.56 only. Consider the variable transformation

$$\phi' = \frac{\phi}{\kappa} \tag{3.59}$$

$$(x', y', z') = \frac{(x, y, z)}{\alpha} \tag{3.60}$$

$$(n', p', N'_A, N'_D) = \frac{(n, p, N_A, N_D) \lambda^2}{\kappa} \tag{3.61}$$

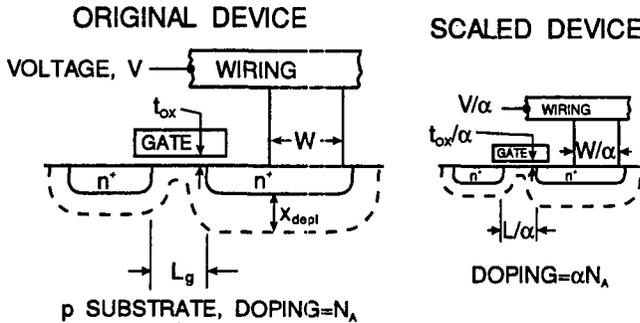


Figure 3.3 Schematic illustration of the scaling of Si technology by a factor α . (Adapted from Davari et al., Ref. 10.)

Equation 3.56 becomes

$$\begin{aligned} \frac{\partial^2(\kappa\phi')}{\partial(\lambda x')^2} + \frac{\partial^2(\kappa\phi')}{\partial(\lambda y')^2} + \frac{\partial^2(\kappa\phi')}{\partial(\lambda z')^2} &= -\frac{q}{\epsilon_s}(N'_D - N'_A + p' - n')\frac{\kappa}{\lambda^2} \\ \frac{\partial^2\phi'}{\partial x'^2} + \frac{\partial^2\phi'}{\partial y'^2} + \frac{\partial^2\phi'}{\partial z'^2} &= -\frac{q}{\epsilon_s}(N'_D - N'_A + p' - n') \end{aligned} \quad (3.62)$$

Equation 3.62 is formally equivalent to Eq. 3.56 and can be interpreted as the Poisson's equation for a scaled device. If the boundary conditions (i.e., potentials at the source, drain, substrate, and gate terminals) are proportionally reduced by the factor κ , solutions of Eqs. 3.56 and 3.62 differ by only a scale factor, and the shape of the electric field pattern is the same in both devices, thus preserving the electrostatic behavior of the device. The intensity of the electric field increases by a factor $\epsilon = \lambda/\kappa$. This electric field increase, as a result of device scaling, has driven much of the device physics research to find ways to mitigate the deleterious effects of higher electric fields in the device and yet sustain device performance improvements for successive technology generations. Table 3.1 summarizes the scaling relations for various scaling scenarios.

There are, however, several physical quantities that do not scale properly:

1. The extrinsic Debye length ($L_D = \sqrt{\epsilon_s kT/q^2 N_A}$) scales as $\sqrt{\kappa}/\lambda$ instead of $1/\lambda$. The Debye length determines the width of the transition region between

TABLE 3.1 Technology Scaling Rules for Three Cases^a

Physical Parameter	Constant Electric Field Scaling Factor	Generalized Scaling Factor	Generalized Selective Scaling Factor
Channel length, insulator thickness	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Wiring width, channel width	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Electric field in device	1	ϵ	ϵ
Voltage	$1/\alpha$	ϵ/α	ϵ/α_d
Doping	α	$\epsilon\alpha$	$\epsilon\alpha_d$
Area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha_w^2$
Capacitance	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Gate delay	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Power dissipation	$1/\alpha^2$	ϵ^2/α^2	$\epsilon^2/\alpha_w\alpha_d$
Power density	1	ϵ^2	$\epsilon^2\alpha_w/\alpha_d$

^a α is the dimensional scaling parameter, ϵ is the electric field scaling parameter, and α_d and α_w are separate dimensional scaling parameters for the selective scaling case. α_d is applied to the device vertical dimensions and gate length, while α_w applies to the device width and the wiring.

Source: After Wong et al., Ref. 11.

neutral and depletion regions (see the solution of Poisson's equation in Eq. 3.21). The Debye length should be less than the depletion region to preserve the electrostatic behavior of the MOSFET.

2. The bandgap energy of silicon does not change with device scaling. This gives rise to a fixed built-in potential between junctions as well as undesirable effects such as band-to-band tunneling and Zener breakdown, thus limiting applicability of the scaling theory in the low-voltage or high-doping regime.
3. The subthreshold slope is determined by the thermodynamics of carrier distribution (Maxwell-Boltzmann distribution or Fermi-Dirac distribution in the classical approximation) and does not increase with device scaling. As a result, the threshold voltage cannot be reduced in proportion to the terminal voltages because of the need to limit the OFF current of the MOSFET. The gate overdrive, $(V_{DD} - V_T)$, therefore, decreases as terminal voltages are reduced.
4. The thickness of the inversion-layer scales as κ/λ in the classical approximation (instead of $1/\lambda$). This thickness is even broader in the presence of quantization of the inversion layer,⁹ giving rise to transconductance degradation as described in more detail in Section 3.6.3.

The second column of Table 3.1 summarizes the scaling relations for constant electric field scaling ($\epsilon = \lambda/\kappa = 1$). In practice, because of previously mentioned difficulties with low voltage, the voltage is seldom scaled as rapidly as the linear dimensions. This can be accommodated by introducing an additional scaling factor ϵ (defined as $\epsilon = \lambda/\kappa$) for the electric field (this ϵ is greater than 1), as summarized under "generalized scaling" in column 3 of Table 3.1. Increasing the electric field requires increasing the amount of doping, and also increases the power dissipation, but it does mitigate some of the low-voltage difficulties. The disadvantage of this scaling is that the increasing electric field increases the energy of the carriers, resulting in such undesirable effects as hot-carrier injection into the gate oxide and impact ionization, which affects device reliability. Indeed, this reliability concern forces the use of lower supply voltages for smaller devices even when power dissipation is not an issue.¹⁰

In recent generations ($L_g < 0.5 \mu\text{m}$) of technology, the wiring is not scaled to the same extent as the gate length, since this improves the wiring yield without degrading the gate delay. This approach, called *selective scaling*, is shown in the final column of Table 3.1 and has two spatial dimension scaling parameters, α_d for scaling the gate length and device vertical dimensions, and α_w for scaling the device width and the wiring. These approaches to scaling and issues related to them are described in more detail in Ref. 10.

The expected power densities and delays of future technology generations have been estimated using these selective scaling rules, and are illustrated in Figure 3.4 down to near the limits of scaling. The high-performance option yields high logic speeds, with loaded delays down to 80 ps even for static CMOS. Still higher speeds are expected for dynamic logic families, but as shown in Figure 3.4a, these high

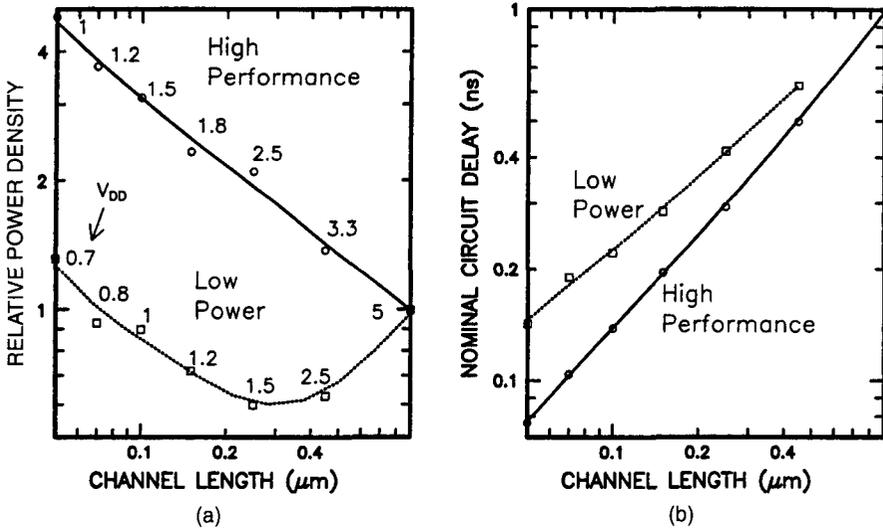


Figure 3.4 (a) Relative power density versus effective channel length for high-performance and low-power technology scaling: (b) corresponding loaded circuit (3-input NAND) delay versus channel length for the same technology scalings. (after Wong et al., Ref. 11 and adapted from Davari et al., Ref. 10.)

speeds are at the expense of high power densities. Note that even though the lower voltages of the low-power path result in an initial savings in power, the power density starts to rise on that path as well for gate lengths less than $0.25 \mu\text{m}$.

3.3.2 Historical Trend and Projections

Scaling theory in conjunction with observations of past industry trends (e.g., Moore's law¹²) has led to the creation of so-called roadmaps for CMOS technology, the most public and widely agreed on is the Semiconductor Industry Association *National Technology Roadmap for Semiconductors* (NTRS)¹ some highlights of which are shown in Table 3.2. The NTRS is a statement of the historical trend as well as a projection of the future device needs and performances as perceived at the time of formulation of the road-map. Note that the effective channel length (not shown in Table 3.2), which determines the current drive, is expected to be 10–30% less than the gate lithography shown in Table 3.2, depending on the manufacturer.

The future DRAM capabilities on this roadmap are based on simple projections of past progress. These numbers require gradually reducing the area required for a single bit down to four lithographic squares, which will require serious innovation, since there is no obvious present path to accomplish this.

The anticipated increase in chip sizes will require and accompany increases in wafer sizes. 300 mm diameter Si wafers are expected to be used in production starting around the beginning of the twenty-first century, and still larger wafers (perhaps 450 mm) are being considered for the future.

TABLE 3.2 1997 SIA Roadmap Highlights

Year first shipped	1995	1997	1999	2001	2003	2006	2009
DRAM (bits/chip)	64 M	256 M	1 G		4 G	16 G	64 G
DRAM chip size (mm ²)	190	280	400	445	560	790	1120
μ P transistors/cm ²	—	3.7 M	6.2 M	10 M	18 M	39 M	84 M
μ P chip size (mm ²)	250	300	340	385	430	520	620
General lithography (μ m)	0.35	0.25	0.18	0.15	0.13	0.10	0.07
Gate lithography (μ m)	0.28	0.20	0.14	0.12	0.10	0.07	0.05
Oxide thickness (nm)	7–12	4–5	3–4	2.4–3.2	2–3	1.5–2	< 1.5
Supply voltage (V)	3.3	1.8–2.5	1.5–1.8	1.2–1.5	1.2–1.5	0.9–1.2	0.6–0.9
V_T 3σ variation (\pm mV)	60	60	50	45	40	40	40
Clock (MHz) (across chip)	300	750	1200	1400	1600	2000	2500

Source: Adapted from NTRS, Ref. 1.

Figures 3.5–3.11 show typical device characteristics and circuit performances for 0.25- and a 0.1- μ m technologies, illustrating the improvement in performance as devices are scaled to a smaller geometry.

3.4 SHORT-CHANNEL EFFECTS

The MOSFET models developed in Section 3.2 assumed that 2D effects due to the proximity of the source–drain and high-drain biases are insignificant. For devices with short channel lengths (when a channel length can be considered “short” is discussed in Section 3.4.2), the drain electric field begins to penetrate into the channel toward the source and the potential distribution begins to require a 2D description. These 2D effects result in the dependence of the threshold voltage on the channel length (and width), drain voltages and substrate voltages, degradation of the subthreshold slope, as well as punchthrough (when the drain depletion region merges with the source depletion region). Furthermore, nonideal device scaling often results in an increase of the electric field in the channel. This leads to such parasitic effects as mobility degradation due to high fields and velocity saturation (see Section 3.5), and hot-carrier effects (see Section 3.6). This section focuses on electrostatic effects such as the dependence of the threshold voltage and punchthrough on channel lengths and drain–substrate biases.

3.4.1 Threshold Voltage

The drain current in the subthreshold region is determined by thermionic emission of carriers over the source to channel barrier. This potential barrier is modulated by the

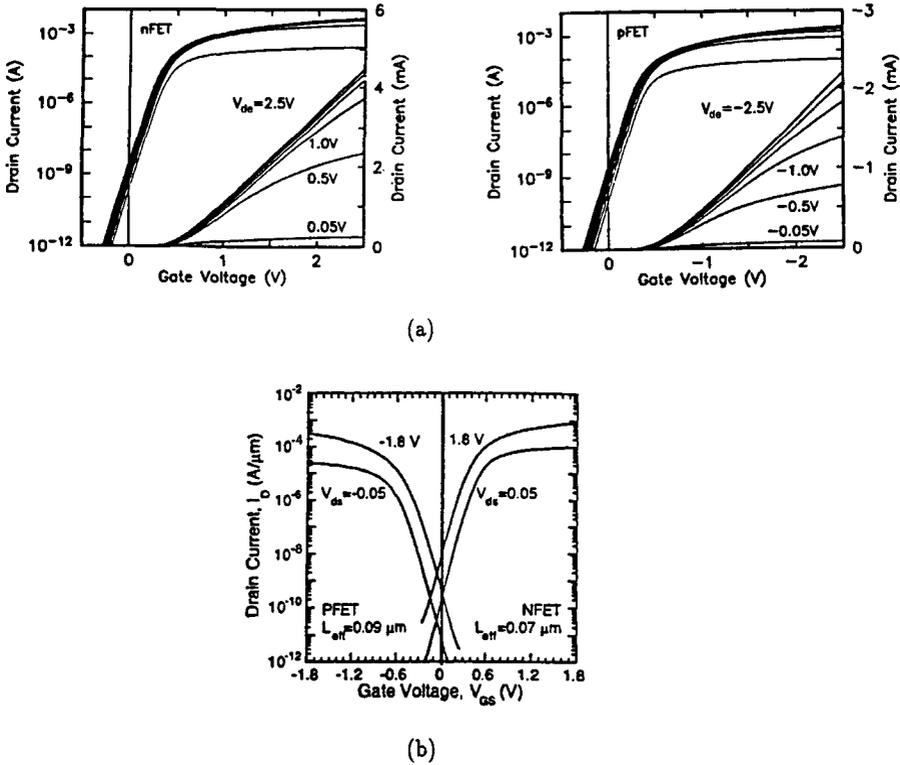


Figure 3.5 Drain current versus gate voltage characteristics: (a) 0.25- μm technology (after Chang et al., Ref. 13); (b) 0.1- μm technology (after Su et al., Ref. 14.)

proximity of the drain junction to the source junction, as well as the application of drain and substrate biases. Figure 3.12 illustrates the surface potential of short-channel MOSFETs for various channel lengths and two drain biases. Two behaviors are evident from Figure 3.12: (1) the potential barrier from the source to the channel (for a constant drain bias) is lowered as the channel length is reduced and (2) the potential barrier from the source to the channel is lowered as the drain bias is increased. Therefore, the threshold voltage of MOSFETs decreases as the channel length is shortened and also as the drain bias is increased. The dependence of the threshold voltage on channel length is usually characterized by a shift of the threshold voltage from the long-channel threshold voltage [$\Delta V_{T(\text{sce})} = V_{T|V_{DS}=100\text{ mV}}(\text{long}) - V_{T|V_{DS}=100\text{ mV}}(\text{short})$] at a low drain bias such as 100 mV. The dependence of the threshold voltage on drain biases [drain-induced barrier lowering (DIBL)] is usually characterized by the shift of the threshold voltage as the drain bias is raised from a low value (e.g., 100 mV) to a high value (e.g., the power supply voltage, V_{DD}) [$\Delta V_{T(\text{DIBL})} = V_{T|V_{DS}=100\text{ mV}} - V_{T|V_{DS}=V_{DD}}$]. It is usually expressed as

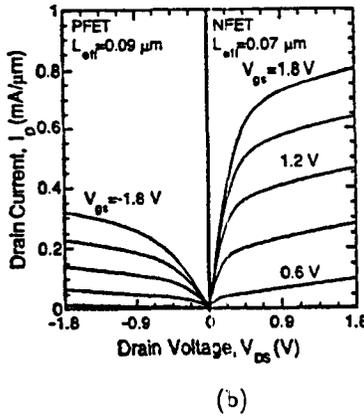
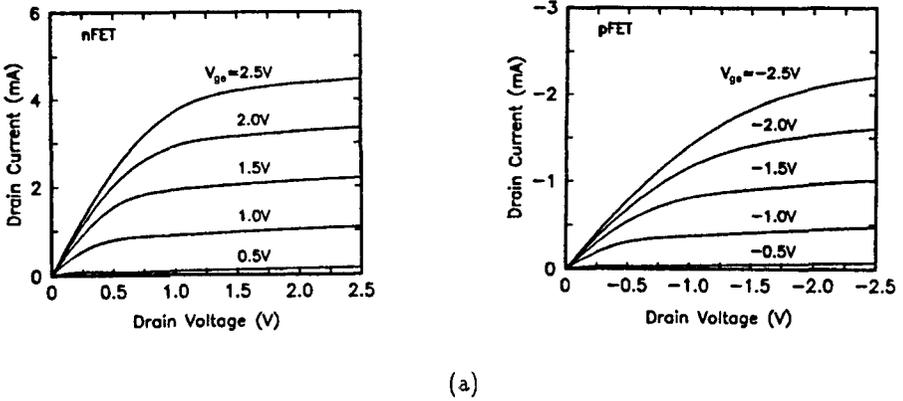


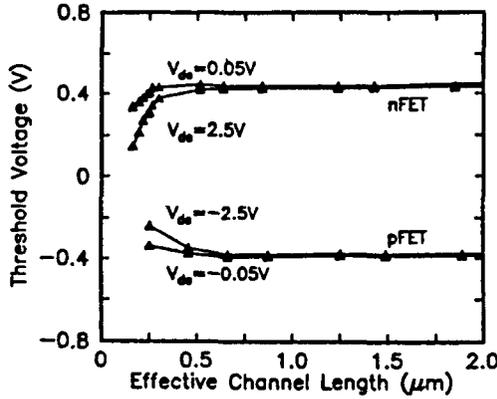
Figure 3.6 Drain current versus drain voltage characteristics: (a) 0.25- μm technology (after Chang et al., Ref. 13); (b) 0.1- μm technology (after Su et al., Ref. 14.)

$\Delta V_T / \Delta V_{DS}$ in units of mV/V. Typical acceptable values are $\Delta V_{T(sce)} = 100 \text{ mV}$ at the minimum channel length and $\Delta V_T / \Delta V_{DS} = 50 \text{ mV/V}$ at the minimum channel length.

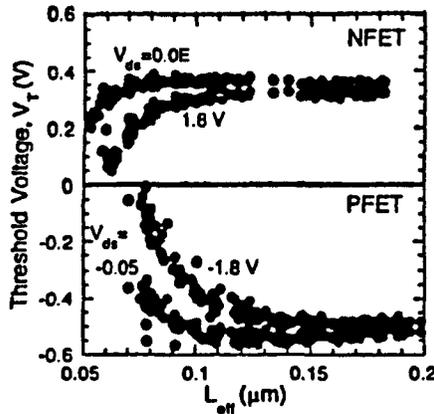
Figure 3.13 shows the typical dependence of the threshold voltage on the effective channel length. We develop a model of the dependence of the threshold voltage on the channel length and drain biases in this section.

Charge-Sharing Model

A simple model of the dependence of the threshold voltage on the channel length is given by Yau¹⁵ and illustrated in Figure 3.14. For a short-channel MOSFET, the depletion charge controlled by the gate is reduced because part of the depletion charge under the gate is controlled by the source–drain junctions. The surface



(a)



(b)

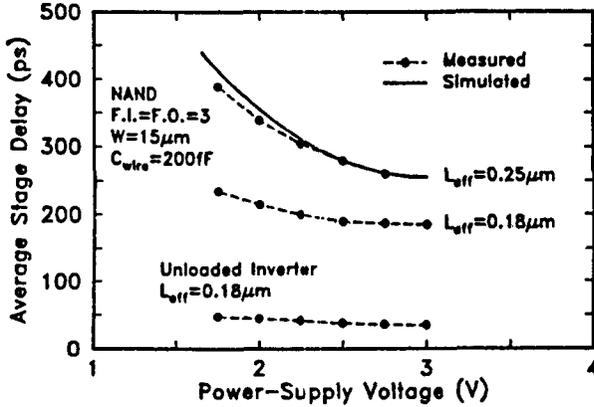
Figure 3.7 (a) Threshold voltage versus effective channel length (after Chang et al., Ref. 13); (b) 0.1- μm technology (after Su et al., Ref. 14.)

potential in the presence of this “charge sharing” may be expressed as

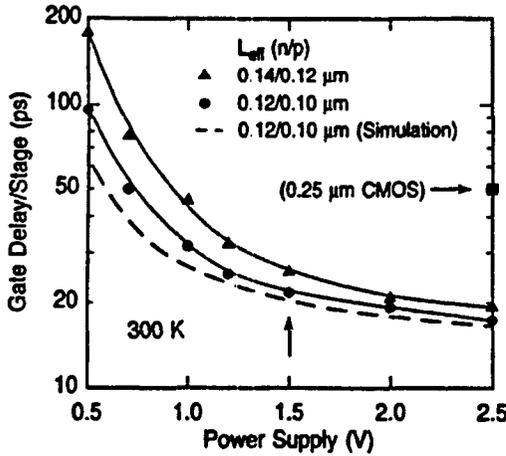
$$V_G = V_{FB} + \phi_s + \frac{Q'_B}{C_{ox}A} \tag{3.63}$$

where Q'_B is the total (not per unit area) depletion charge controlled by the gate electrode and A is the gate area. The threshold voltage is then given by

$$V_T = V_{FB} + 2\phi_{Fp} + \frac{Q'_B}{C_{ox}A} \tag{3.64}$$



(a)



(b)

Figure 3.8 (a) Inverter delay versus the power supply voltage (after Chang et al. Ref. 13); (b) 0.1- μm technology (after Taur et al., Ref. 35.)

For long-channel devices where $L' \gg W_{nt}$, we have $Q_B = Q'_B/A = qN_A W_m$, where W_m is the maximum depletion depth. The surface potential is given by¹⁵

$$V_G = V_{FB} + \phi_s + \frac{1}{C_{ox}} \sqrt{\frac{q\epsilon_s N_A (\phi_s + V_{SB})}{2}} \left(1 + \frac{L - W_D - W_S}{L - y_D - y_S} \right) \quad (3.65)$$

$$y_S = \sqrt{\frac{2\epsilon_s (V_{bi} - \phi_s)}{qN_A}} \quad (3.66)$$

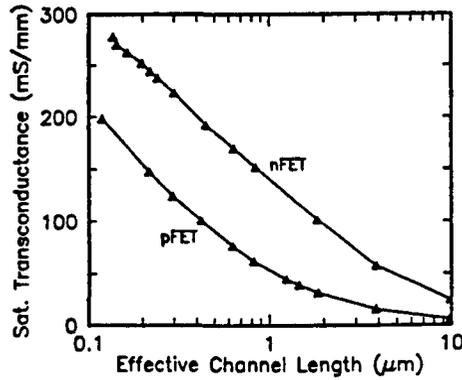
$$y_D = \sqrt{\frac{2\epsilon_s (V_{bi} - \phi_s + V_D)}{qN_A}} \quad (3.67)$$

$$W_S = \sqrt{\frac{2\epsilon_s(V_{bi} + V_{SB})}{qN_A}} \quad (3.68)$$

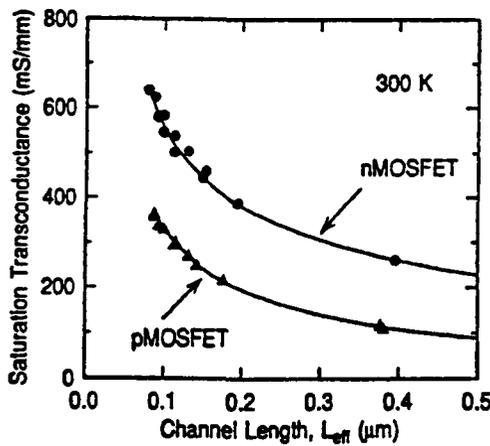
$$W_D = \sqrt{\frac{2\epsilon_s(V_D + V_{bi} + V_{SB})}{qN_A}} \quad (3.69)$$

$$L_{\text{eff}} = L - y_S - y_D \quad (3.70)$$

The effective channel length L_{eff} should be used in the drain current models in Section 3.2. The quantity Q'_B can be computed using purely geometric analysis as



(a)



(b)

Figure 3.9 (a) Saturation transconductance versus effective channel length (after Chang et al., Ref. 13); (b) 0.1- μm technology (after Taur et al., Ref. 85.)

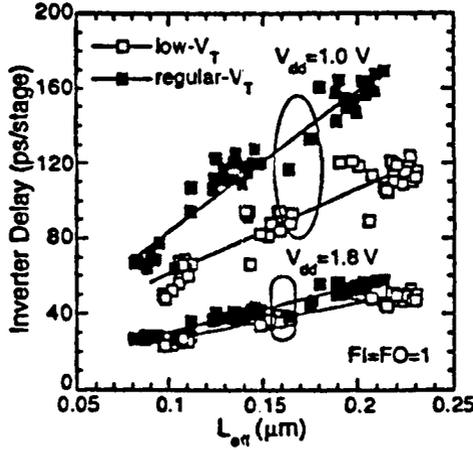


Figure 3.10 Inverter delay versus channel length for a 0.1- μm technology (after Su et al., Ref. 14.)

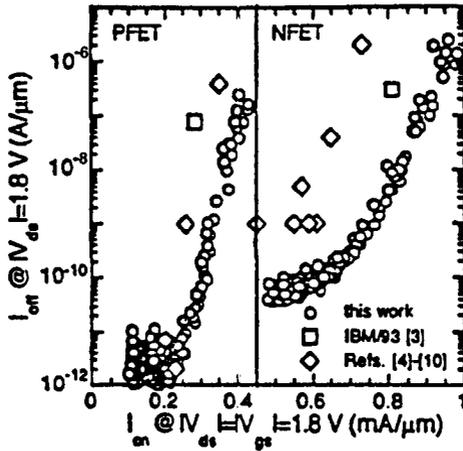


Figure 3.11 ON-current (I_{on}) versus OFF current (I_{off}) for a 0.1- μm technology (after Su et al., Ref. 14.)

described in Ref. 15:

$$\frac{Q'_B}{W} = \frac{qN_A W_m(L + L')}{2} \tag{3.71}$$

$$\frac{L + L'}{2L} = 1 - \frac{r_j}{L} \left(\sqrt{1 + \frac{2W_m}{r_j}} - 1 \right) \tag{3.72}$$

$$\Delta V_T = -\frac{qN_A W_m r_j}{2C_{ox} L} \left[\left(\sqrt{1 + \frac{2y_D}{r_j}} - 1 \right) + \left(\sqrt{1 + \frac{2y_S}{r_j}} - 1 \right) \right] \tag{3.73}$$

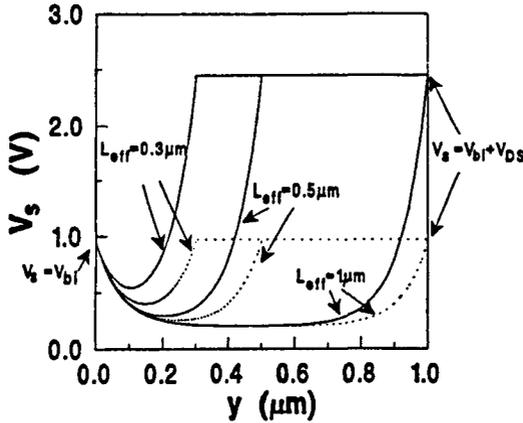


Figure 3.12 Calculated surface potential along the channel for different channel lengths. The device parameters are $t_{ox} = 10 \text{ nm}$, $N_A = 10^{16} \text{ cm}^{-3}$, $N_{S/D} = 10^{20} \text{ cm}^{-3}$, and $\eta = 1$, that is, $l = 0.1 \mu\text{m}$. The substrate bias is zero volts. The dashed line show the data for $V_{DS} = 0.05 \text{ V}$, and the solid lines show the data for $V_{DS} = 1.5 \text{ V}$, respectively (after Liu et al., Ref. 17.)

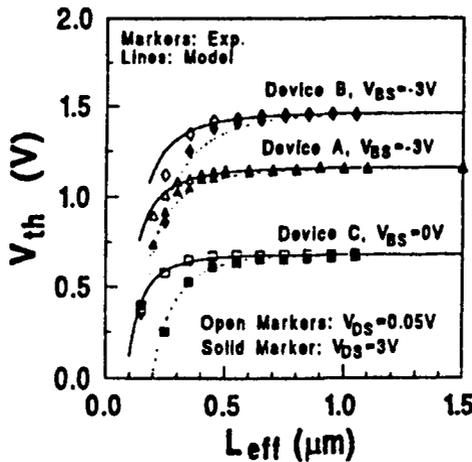
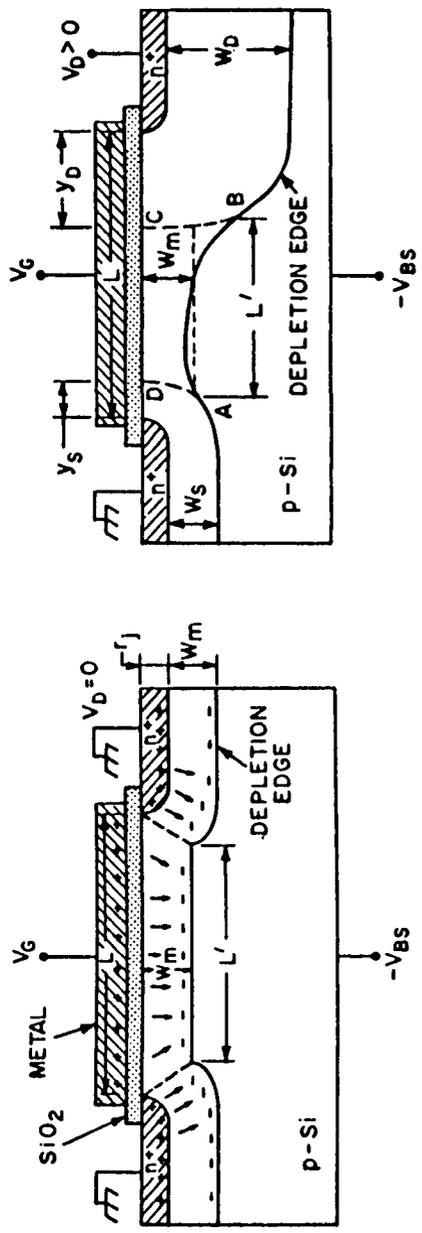
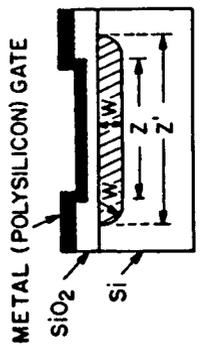


Figure 3.13 Experimental and calculated threshold voltage versus effective channel length for non-LDD (lightly doped drain) MOSFETs from different technologies. Device A: $t_{ox} = 5.5 \text{ nm}$, $N_A = 3.6 \times 10^{17} \text{ cm}^{-3}$, $x_j = 0.25 \mu\text{m}$, $l = 0.1 \mu\text{m}$. Device B: $t_{ox} = 8.6 \text{ nm}$, $N_A = 1.5 \times 10^{17} \text{ cm}^{-3}$, $x_j = 0.2 \mu\text{m}$, $l = 0.05 \mu\text{m}$. Device A: $t_{ox} = 15.6 \text{ nm}$, $N_A = 4.0 \times 10^{16} \text{ cm}^{-3}$, $x_j = 0.2 \mu\text{m}$, $l = 0.09 \mu\text{m}$ (after Liu et al., Ref. 17.)



(a)



(b)

Figure 3.14 Charge sharing in small MOSFETs: (a) short-channel effects; (b) narrow-width effects (after Sze, Ref. 3.)

where r_j is as defined in Figure 3.14. For narrow width MOSFETs, the effective depletion width is greater than the physical gate width because of the spreading of the depletion region beyond the physical gate width. The effective gate width can be expressed as

$$W_{\text{eff}} = W \left[1 + \frac{\pi W_m}{2 W} \right] \quad (3.74)$$

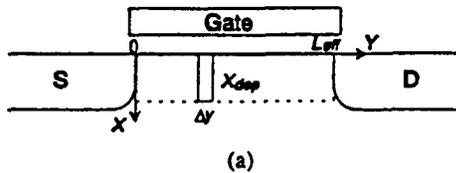
and the threshold voltage shift is

$$\Delta V_T = V_{FB} + 2\phi_{Fp} + \frac{\sqrt{2\varepsilon_s q N_A (2\phi_F + V_{SB})}}{C_{\text{ox}}} \left[1 + \frac{\pi W_m}{2 W} \right] \quad (3.75)$$

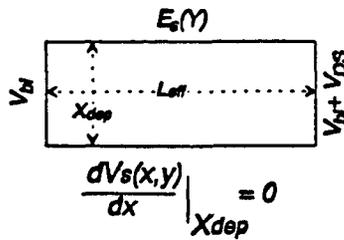
Quasi-Two-Dimensional Analysis

The charge-sharing model in Section 3.4.1, although simple and conceptually easy to understand, is unable to model the drain-induced barrier lowering¹⁶ (DIBL) or the dependence of the threshold voltage on the substrate bias. The quasi-two-dimensional analysis developed by Liu et al.¹⁷ solves the Poisson equation in the depletion region with suitable boundary conditions (see Fig. 3.15). Accounting for the electric fluxes that go in and come out of the rectangular box (Gaussian box) of height X_{dep} and length Δy in the channel depletion region and neglecting mobile carrier charge in the subthreshold region, we obtain

$$\varepsilon_s \frac{X_{\text{dep}}}{\eta} \frac{dE_s(y)}{dy} + \varepsilon_{\text{ox}} \frac{V_{GS} - V_{FB} - V_s(y)}{t_{\text{ox}}} = qN_A X_{\text{dep}} \quad (3.76)$$



(a)



(b)

Figure 3.15 Diagrams showing (a) the Gaussian box used in the quasi-two-dimensional analysis, and (b) the boundary conditions for solving Eq. 3.76. (after Liu et al., Ref. 17.)

where $E_s(y)$ is the lateral surface electric field, $V_s(y)$ is the channel potential at the Si/SiO₂ interface, N_A is the channel doping, and t_{ox} is the gate oxide thickness. The depletion-layer thickness, X_{dep} , is equal to $\sqrt{2\epsilon_s(\phi_s + V_{SB})/qN_A}$, $\phi_s = 2\phi_{FP}$ is the surface potential at the threshold of inversion, and η is a fitting parameter that depends on the drain voltage as a second-order effect. The solution to Eq. 3.76 under the boundary conditions of $V_s(0) = V_{bi}$ and $V_s(L) = V_{DS} + V_{bi}$ is

$$V_s(y) = V_{sL} + (V_{bi} + V_{DS} - V_{sL}) \frac{\sinh(y/l)}{\sinh(L/l)} + (V_{bi} - V_{sL}) \frac{\sinh[(L-y)/l]}{\sinh(L/l)} \quad (3.77)$$

In this equation, $V_{sL} = V_{GS} - V_{T0} + \phi_s$ represents the long-channel surface potential, and $V_{T0} = V_{FB} + \phi_s + qN_A X_{dep} t_{ox} / \epsilon_{ox}$ represents the long-channel threshold voltage. V_{bi} is the built-in potential between the source–substrate and drain–substrate junctions, and l is the characteristic length (see Section 3.4.2) defined as $l = \sqrt{\epsilon_s t_{ox} X_{dep} / \epsilon_{ox} \eta}$ and can be found from experiment for a given technology. In the subthreshold region, the shift of the surface potential minimum, compared to the long-channel case and to the low drain–substrate bias cases, can be equated to the shift of the threshold voltage. The surface potential minimum occurs at $y = y_0$, where $dV_s(y)/dy|_{y=y_0} = 0$. For $L \gg l$, Eq. 3.77 can be approximated as

$$V_s(y) = V_{sL} + (V_{bi} + V_{DS} - V_{sL})e^{(y-L)/l} + (V_{bi} - V_{sL})e^{-y/l} + (V_{bi} + V_{DS} - V_{sL})e^{-L/l} \quad (3.78)$$

After solving for the location y_0 of the surface potential minimum, the threshold voltage shift is given by¹⁷

$$\Delta V_T = 2(V_{bi} - \phi_s) + [V_{DS} + (V_{bi} - \phi_s)](1 - e^{-L/l}) + \frac{2\sqrt{(V_{bi} - \phi_s)^2 + (V_{bi} - \phi_s)[(V_{bi} - \phi_s) + V_{DS}]}(e^{L/l} - 1)}{4 \sinh^2(L/2l)} \quad (3.79)$$

For $L \gg l$, we obtain

$$\Delta V_T \simeq [3(V_{bi} - \phi_s) + V_{DS}]e^{-L/l} + 2\sqrt{(V_{bi} - \phi_s)(V_{bi} - \phi_s + V_{DS})}e^{-L/2l} \quad (3.80)$$

If the drain bias V_{DS} is low, Eq. 3.80 may be approximated as

$$\Delta V_T \simeq [2(V_{bi} - \phi_s) + V_{DS}](e^{-L/2l} + 2e^{-L/l}) \quad (3.81)$$

It is worth noting that the threshold voltage shift for low drain voltages is the sum of two exponential terms and the dependence of the threshold voltage shift on drain biases has the functional form $A\sqrt{V_{DS}} + B\sqrt{V_{DS}}$. This is illustrated in Figures 3.16 and 3.17, where the threshold voltage shift versus the effective channel length and drain voltages are plotted.

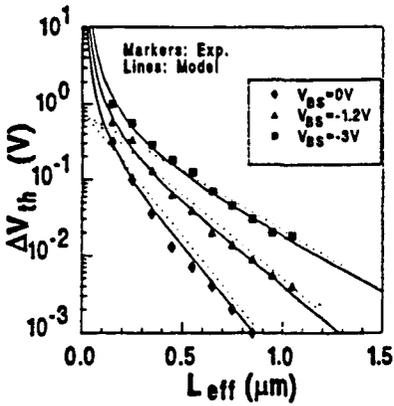


Figure 3.16 Threshold voltage shift versus effective channel length at $V_{DS} = 0.05$ V and different substrate biases for non-LDD devices. The solid lines are calculated results, and the dashed lines are lines best fitting the experimental data of $L_{\text{eff}} > 5l$. Note that all the dashed lines intersect at the same point of $2(V_{bi} - \phi_s) + V_{DS}$. (after Liu et al., Ref. 17.)

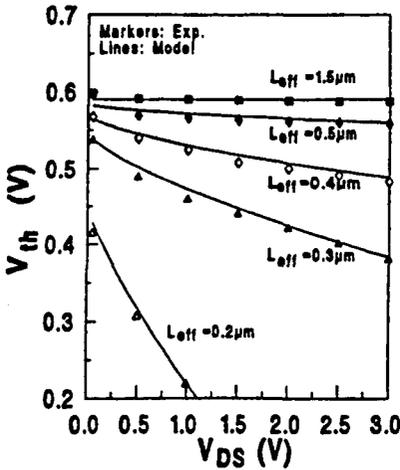


Figure 3.17 Threshold voltage shift versus drain voltage at $V_{SB} = 0$ with different effective channel lengths, demonstrating the effects of DIBL on the threshold voltage. The devices are the same as those in Fig. 3.16. (after Liu et al., Ref. 17.)

The characteristic length can be determined experimentally for a given technology by plotting the log of the threshold voltage shift versus the effective channel length for a low drain bias (see Eq. 3.81). The slope of the curve at large L is $-(2l \ln(10))^{-1}$ (see Fig. 3.18).

3.4.2 Characteristic Device Scale Length

Short-channel effects are 2D effects where the channel potential is no longer determined solely by the applied gate electric field in the direction normal to the Si/SiO₂ interface, but affected also by the proximity of the source-drain junctions and source-drain potentials. It is often desirable to have a characteristic length whereby one can compare the effective channel lengths to quantify the short-channel behavior. This characteristic length delineates long-channel device behavior from short-channel device behavior. This concept of the characteristic length, which was

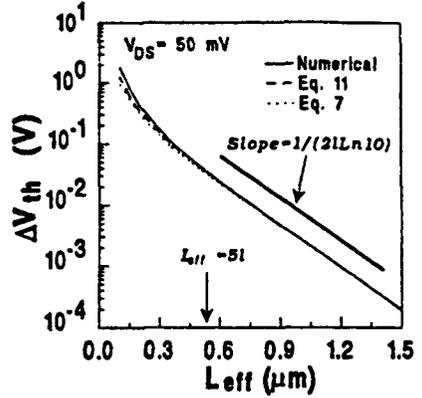


Figure 3.18 Threshold voltage shift versus effective channel lengths at $V_{DS} = 0.05$ V. The characteristic length l can be extracted from the slope of this plot at large L_{eff} (e.g., $L_{eff} \gg 5l$). (after Liu et al., Ref. 17.)

developed by Yan et al.,¹⁸ started from the groundwork of Young¹⁹ and Woo et al.²⁰ and was later extended by Liu et al.,¹⁷ Suzuki et al.,²¹ and Frank et al.²²

The characteristic length is applicable in the context of below-threshold behavior of the MOSFET where the energy barrier from the source to the channel controls the diffusion current or the punchthrough current. To illustrate the methodology of arriving at the characteristic length, we use a thin SOI MOSFET as an example. Other examples can be found in Refs. 17, 18, 21, and 22.

Consider the Poisson equation (Eq. 3.10) in two dimensions (rewritten as Eq. 3.82):

$$\frac{d^2\Phi}{dx^2} + \frac{d^2\Phi}{dy^2} = \frac{qN_A}{\epsilon_{Si}} \tag{3.82}$$

One can simplify the 2D Poisson equation into a one-dimensional (1D) Laplace equation along the direction of the channel by making some simplifying assumptions about the boundary conditions in the y direction (normal the Si/SiO₂ interface). Assuming that the potential, Φ in the direction normal to the Si/SiO₂ interface can be described by a parabolic function with coefficients c_0 , c_1 , and c_2 :

$$\Phi(x, y) = c_0(x) + c_1(x)y + c_2(x)y^2 \tag{3.83}$$

For SOI with a thick buried oxide, the electric field at the silicon–buried oxide interface approaches zero. The boundary conditions are therefore

$$\Phi(x, 0) = \Phi_f(x) = c_0(x) \tag{3.84}$$

$$\left. \frac{d\Phi(x, y)}{dy} \right|_{y=0} = \frac{\epsilon_{ox}}{\epsilon_{Si}} \frac{\Phi_f(x) - \Phi_{gs}}{t_{ox}} = c_1(x) \tag{3.85}$$

$$\left. \frac{d\Phi(x, y)}{dy} \right|_{y=t_{Si}} = \frac{\epsilon_{ox}}{\epsilon_{Si}} \frac{\Phi_{bs} - \Phi_b(x)}{t_{box}} = c_1(x) + 2t_{Si}c_2(x) \simeq 0 \tag{3.86}$$

where $\Phi_f(x)$, Φ_{gs} , $\Phi_b(x)$, and Φ_{bs} are potentials at the front gate oxide interface, the front gate, the back oxide interface, and the bulk substrate, respectively. Using these boundary conditions, Eq. 3.83 can be expressed as

$$\Phi(x, y) = \Phi_f(x) + \frac{\epsilon_{ox}}{\epsilon_{Si}} \frac{\Phi_f(x) - \Phi_{gs}}{t_{ox}} y - \frac{1}{2t_{Si}} \frac{\epsilon_{ox}}{\epsilon_{Si}} \frac{\Phi_f(x) - \Phi_{gs}}{t_{ox}} y^2 \quad (3.87)$$

The 2D Poisson equation of Eq. 3.82 can be transformed into a 1D Poisson equation using Eq. 3.87:

$$\frac{d^2 \Phi_f(x)}{dx^2} - \frac{\epsilon_{ox}}{\epsilon_{Si}} \frac{\Phi_f(x) - \Phi_{gs}}{t_{Si} t_{ox}} = \frac{qN_A}{\epsilon_{Si}} \quad (3.88)$$

Equation 3.88 can be further simplified into a 1D Laplace equation:

$$\frac{d^2 \phi(x)}{dx^2} - \frac{\phi(x)}{\lambda^2} = 0 \quad (3.89)$$

with the following substitution:

$$\lambda = \sqrt{\frac{\epsilon_{Si}}{\epsilon_{ox}} t_{Si} t_{ox}} \quad (3.90)$$

$$\phi(x) = \Phi_f(x) - \Phi_{gs} + \frac{qN_A}{\epsilon_{Si}} \lambda^2 \quad (3.91)$$

Note that Eq. 3.91 indicates that $\phi(x)$ and $\Phi_f(x)$ differ only by a position (x)-independent constant term and as far as obtaining relative potential (with respect to x) is concerned, the solution to $\phi(x)$ and $\Phi_f(x)$ are equivalent.

Equation 3.89 can be solved by using the two boundary conditions for potentials at the source and the drain:

$$\phi(x=0) = V_{bi} - \Phi_{gs} + \frac{qN_A}{\epsilon_{Si}} \lambda^2 = \phi_s \quad (3.92)$$

$$\phi(x=L) = V_{ds} + V_{bi} - \Phi_{gs} + \frac{qN_A}{\epsilon_{Si}} \lambda^2 = \phi_d \quad (3.93)$$

with the solution

$$\phi(x) = \frac{\phi_s [e^{(L-x)/\lambda} - e^{-(L-x)/\lambda}] + \phi_d [e^{x/\lambda} - e^{-x/\lambda}]}{e^{L/\lambda} - e^{-L/\lambda}} \quad (3.94)$$

The minimum potential at the Si/SiO₂ interface controls the punchthrough behavior. The solution to the potential minimum is given by Eq. 3.94 as

$$\phi_{\min} = 2\sqrt{\phi_s \phi_d} e^{-L/2\lambda} \quad (3.95)$$

TABLE 3.3 Characteristic Length Scales for Various MOSFET Device Structures

Structure	Characteristic Length Scale	
SOI	$\lambda = \sqrt{\frac{\epsilon_{Si}}{\epsilon_{ox}} t_{Si} t_{ox}}$	(Ref. 18)
Ground plane	$\lambda = \sqrt{\frac{\epsilon_{Si}}{2\epsilon_{ox}} \frac{t_{Si} t_{ox}}{\left[1 + \frac{\epsilon_{Si}}{\epsilon_{ox}} \frac{t_{ox}}{t_{Si}}\right]}}$	(Ref. 18)
	$0 = \epsilon_{Si} \tan(\pi t_{ox}/\lambda) + \epsilon_{ox} \tan(\pi t_{Si}/\lambda)$	(Ref. 22)
Double gate	$\lambda = \sqrt{\frac{\epsilon_{Si}}{2\epsilon_{ox}} t_{Si} t_{ox}}$	(Ref. 18)
	$\lambda = \sqrt{\frac{\epsilon_{Si}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox}}{4\epsilon_{Si}} \frac{t_{Si}}{t_{ox}}\right) t_{Si} t_{ox}}$	(Ref. 21)
	$1 = \frac{\epsilon_{Si}}{\epsilon_{ox}} \tan(\pi t_{ox}/\lambda) \tan(\pi t_{Si}/2\lambda)$	(Ref. 22)

Equation 3.91 indicates that the long-channel solution is obtained for $\phi_{\min} \rightarrow 0$ ($\Phi_f(x)$ is independent of x). For $L/2\lambda \gg 3$, ϕ_{\min} approaches zero. Therefore, the ratio of L to λ reflects the relative importance of short-channel behavior, and λ , therefore, can serve as a characteristic length scale for short-channel effects. More exact solutions for the general case where the dielectric constants of the gate insulator and the silicon are taken into account are given by Frank et al.²² Table 3.3 compares the characteristic length scales of various advanced MOSFET device structures.

From an analysis of the characteristic lengths, it is easy to notice that the short-channel behavior is determined largely by geometric factors such as the gate insulator thickness, the silicon channel thickness (or depletion depth), and the dielectric constants of the materials. The essence of short-channel device design (as far as electrostatics are concerned) is to control these parameters. The channel doping comes into play mainly through the dependence of the depletion depth on the channel doping. A MOSFET with a low channel doping can still have good short-channel behavior provided the channel is thin. This can be accomplished either by a heavily doped region below the channel (a “ground plane” structure) or physically confining the channel (e.g., thin SOI, double-gate MOSFET). This is discussed in more detail in Section 3.7.

3.5 TRANSPORT PROPERTIES

Figure 3.19 shows the measured carrier velocity versus electric field for Si, Ge, and GaAs.³ At low electric fields, the carrier drift velocity v_d is proportional to the

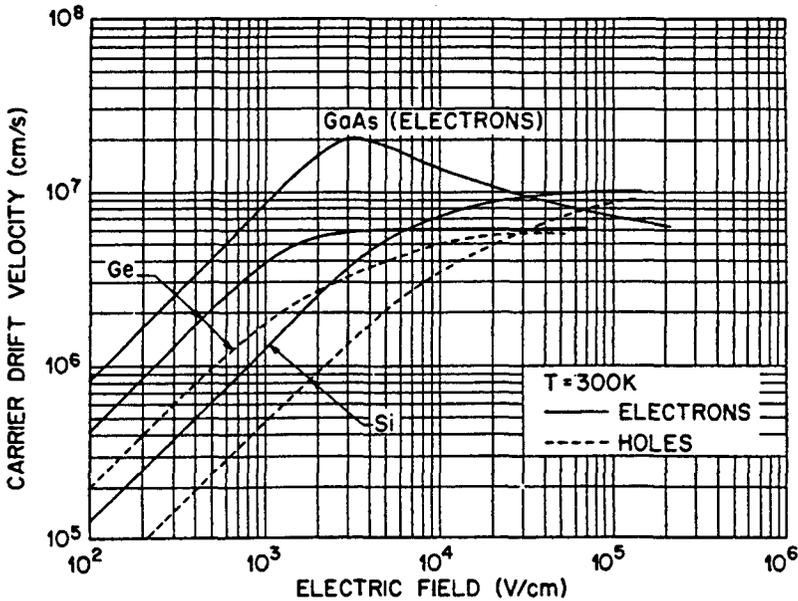


Figure 3.19 Measured carrier velocity versus electric field for high-purity Ge, Si, and GaAs. (after Sze, Ref 3.)

electric field E with the proportionality constant defined as the mobility μ :

$$v_d = \mu E \quad (3.96)$$

The mobility can be expressed in terms of the relaxation time between collisions with scattering centers (τ^{coll})

$$\mu = \frac{q\tau^{\text{coll}}}{m^*} \quad (3.97)$$

where m^* is the effective mass of the mobile carrier. At low longitudinal electric fields (fields along the direction of carrier transport), carrier transport can be described by the mobility. At high longitudinal fields, the carrier velocity may saturate. In addition, velocity overshoot may occur for very short channel MOSFETs where electric fields change rapidly and the carriers have no time to relax to equilibrium and gain energy. Low (longitudinal)-field mobility is treated in Section 3.5.1, and high-field transport is treated in Section 3.5.2.

3.5.1 Mobility

Scattering Mechanisms

At low longitudinal electric fields, the main scattering mechanism are³

1. Acoustic phonon scattering, $\mu_{\text{ac}} = [\sqrt{8\pi}q\hbar^4 C_{11}/3E_{\text{ds}}m^{*5/2}(kT)^{3/2}] \simeq (m^*)^{-5/2} T^{-3/2}$.

2. Ionized impurity (Coulomb) scattering, $\mu_I = [64\sqrt{\pi}\epsilon_s^2(2kT)^{3/2}/N_Iq^3m^{*1/2}] \cdot \{\ln [1 + (12\pi\epsilon_s kT/q^2 N_I^{1/3})^2]\}^{-1} \simeq (m^*)^{-1/2} N_I^{-1} T^{3/2}$
3. Surface roughness scattering²³⁻³⁰ due to the roughness of the Si/SiO₂ interface

Assuming that the scattering collisions are independent and the carrier has time to relax between collisions, the relaxation time in the presence of all scattering mechanisms can often be approximated by Matthiessen’s rule³¹

$$\begin{aligned} \frac{1}{\tau} &= \frac{1}{\tau_1} + \frac{1}{\tau_2} + \frac{1}{\tau_3} + \dots \\ \frac{1}{\mu} &= \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \dots \end{aligned} \tag{3.98}$$

Matthiessen’s rule breaks down when the relaxation times (τ) depends on the wavevector \vec{k} .³¹ In general

$$\begin{aligned} \frac{1}{\tau} &\geq \frac{1}{\tau_1} + \frac{1}{\tau_2} + \frac{1}{\tau_3} + \dots \\ \frac{1}{\mu} &\geq \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \dots \end{aligned} \tag{3.99}$$

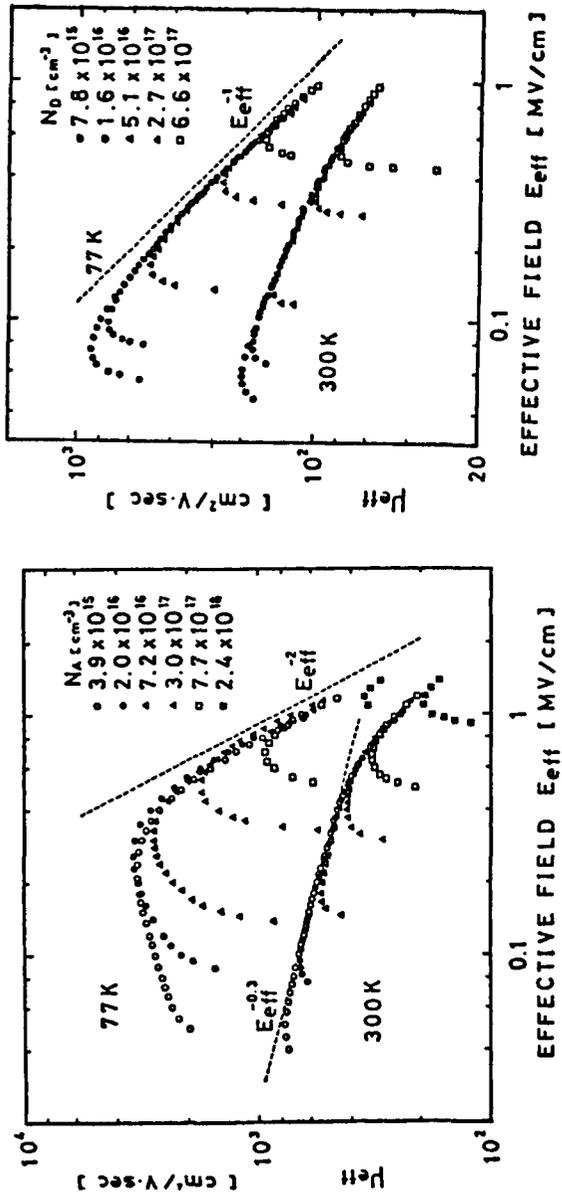
holds even when Matthiessen’s rule breaks down.³¹

In silicon MOSFETs, phonon scattering dominates at high temperatures and intermediate transverse (to the direction of carrier transport) electric fields; Coulomb (ionized impurity) scattering dominates at low temperatures, low transverse electric fields, and high doping; and surface roughness scattering dominates at low temperatures and high transverse fields.

Universal Mobility

It is often found experimentally³²⁻³⁵ (Fig. 3.20) that the measured effective mobility of silicon MOSFETs follows a “universal” curve independent of substrate doping and gate oxide thickness when plotted against the effective transverse electric field. Although these experiment results are conceptually simple, its general validity is not clear and the theoretical basis for such “universal” behavior is still lacking. The effective mobility is obtained by measuring the drain conductance

$$\begin{aligned} \mu_{\text{eff}} &= \frac{\left. \frac{\partial I_D}{\partial V_{DS}} \right|_{V_{DS}}}{\frac{W}{L} Q_n} \rightarrow 0 \\ &= \frac{g_{DS}}{\frac{W}{L} C_{\text{ox}} Q_n} \end{aligned} \tag{3.100}$$



(a)

(b)

Figure 3.20 Measured effective mobility versus effective field for (a) electrons and (b) holes for various substrate doping concentrations. (after Takagi et al., Ref. 35.)

The effective transverse electric field (called the “effective” field) is defined by

$$E_{\text{eff}} = \frac{\int E(x)n(x)dx}{\int n(x)dx} \quad (3.101)$$

The mobile carrier density per unit area is given by $Q_n = \int_0^{x_{\text{inv}}} n(x)dx$ and is often approximated by $Q_n \simeq C_{\text{ox}}(V_{GS} - V_T)$ in measurements. However, this approximation is poor near the threshold and a split C–V measurement³⁶ should be performed to obtain accurate measurements of the mobile carrier density. The effective field (E_{eff}) is often approximated in measurements as

$$E_{\text{eff}} = \frac{1}{\epsilon_s} \left(Q_b + \frac{1}{2} Q_n \right) \quad (3.102)$$

The approximation of this equation is exact for the case of a uniform channel doping if one assumes a Maxwell–Boltzmann distribution of the mobile carriers^{4,37,38} or that all the mobile carriers reside in the lowest subband with the triangular potential solution of the Schrödinger equation of the quantized inversion layer.³⁸

Many experiments have shown that universal mobility behavior is observed for holes only if the effective field is formulated as^{34,35}

$$E_{\text{eff}} = \frac{1}{\epsilon_s} (\zeta Q_b + \eta Q_n) \quad (3.103)$$

where $\zeta = 1$ and $\eta = \frac{1}{3}$. Some have even proposed a temperature-dependent ζ and η .³⁹ It is not clear from first principles why $\zeta = 1$ and $\eta = \frac{1}{3}$ have to be used for holes.⁴⁰ Krutsick and White³⁸ show that the values of ζ and η are dependent on the doping profiles of the channel region independent of whether the carriers are holes or electrons.

In Figure 3.20, where we show the measured effective mobility versus effective field for electrons and holes,³⁵ different scattering mechanisms dominate in three distinct regions:³⁵

1. At low effective fields and high doping, Coulomb scattering dominates. At high carrier concentrations (usually accompanied by high fields), Coulomb scattering is screened by the mobile carriers, and scattering due to Coulomb centers decreases. Figure 3.21 illustrates the measured mobility versus the inversion-layer electron areal density (N_s) as a function of the substrate doping (N_A). The effective mobility follows a N_A^{-1} and N_D^{-1} dependence.³⁵
2. At high effective fields, surface roughness scattering dominates. Effective mobility of electrons follows E_{eff}^{-1} at 300 K and follows E_{eff}^{-2} at 77 K.
3. In the intermediate effective field region, phonon scattering dominates. The effective mobility for electrons follows $E_{\text{eff}}^{-0.3}$ at 300 K⁴¹ and follows E_{eff}^{-1} at 77 K, while there is no single power-law trend for holes because no one mechanism dominates. For electrons in high substrate dopings, Coulomb

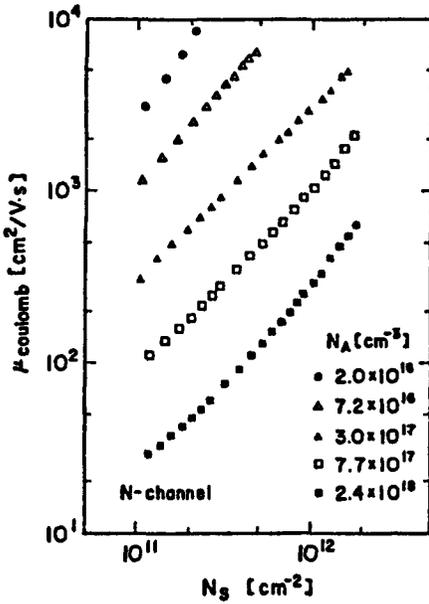


Figure 3.21 Measured effective mobility versus mobile carrier density with different substrate dopings. (after Takagi et al., Ref. 35.)

scattering persists into the high effective-field region and the phonon-scattering-dominated region disappears.

The universal mobility is also dependent on the crystalline orientation of the current transport direction.⁴² For SOI MOSFETs with a relatively thick silicon channel, universal mobility is manifested as long as the back channel electric field is taken into account in calculating the effective field.⁴³

Mobility Models

The mobility that enters into drain current equations such as Eq. 3.8 is obtained with Matthiessen's rule

$$\begin{aligned}
 \mu &= \frac{\mu_0}{1 + \frac{\mu_0}{\mu_{SR}}} \\
 &= \frac{\mu_0}{1 + \frac{E_{eff}}{E_{crit}}} \\
 &= \frac{\mu_0}{1 + \frac{\theta_s}{C_{ox}} |2Q_b + Q_n|}
 \end{aligned} \tag{3.104}$$

where μ_0 is the low-field mobility, μ_{SR} is the mobility due to surface roughness scattering, E_{crit} is the critical field for surface roughness-scattering, which depends

on the effective mass, the mean asperity height (Δ), and correlation length (L) of the assumed Gaussian-distributed surface undulation of the oxide–silicon interface;²³ θ_s characterizes the surface roughness scattering and is defined as $\theta_s = \epsilon_{\text{ox}}/2\epsilon_s t_{\text{ox}} E_{\text{crit}}$. The “low-field mobility” μ_0 contains the contributions from phonon scattering, Coulomb scattering from oxide charges, interface charges, and lattice ions. We can express μ_0 as

$$\mu_0 = \frac{\mu_{i,L}}{1 + \alpha(N_I + \int_0^{\phi_s(y)} D_{it}(\phi) d\phi)} \quad (3.105)$$

where N_I is the total number of all charges per unit area (sum of all trapped charges, Q_{ot} , fixed charges, Q_f , and interface charges, Q_{it}) in the plane of the Si/SiO₂ interface at zero band bending ($\phi_s = 0$), and is independent of ϕ_s . D_{it} is the density of interface states per unit area per eV (electron volt), and $\mu_{i,L}$ is the mobility, which includes phonon and impurity scattering in the silicon. Factor α is an empirical parameter that can also be derived from physical quantities.²⁵

Mobility Models for Numerical Simulation

Device simulations based on the discretization of the drift–diffusion equations^{44,45} often employ a mobility model that is dependent on local variables such as electric fields. Universal mobility models described in Section 3.5.1 are nonlocal models and cannot be directly applied to a device simulator based on discretization to a mesh.⁴⁰

An example of a local-field mobility model is the Lombardi model.⁴⁶ The mobility is modeled in the Lombardi model as

$$\frac{1}{\mu} = \frac{1}{\mu_{ac}} + \frac{1}{\mu_b} + \frac{1}{\mu_{sr}} \quad (3.106)$$

$$\frac{1}{\mu_{ac}}(E_{\perp}, T) = \left(B \frac{T}{E_{\perp}} + C \frac{1}{E_{\perp}^{1/3}} \right) \frac{1}{T} \quad (3.107)$$

$$\mu_b(N_A, T) = \mu_0 + \frac{\mu_{\text{max}}(T) - \mu_0}{1 + (N_A/C_r)^{\alpha}} - \frac{\mu_1}{1 + (C_s/N_A)^{\beta}} \quad (3.108)$$

$$\mu_{\text{max}}(T) = \mu_{\text{max}} \left(\frac{T}{300} \right)^{-\gamma} \quad (3.109)$$

$$\mu_{sr}(E_{\perp}) = \frac{\delta}{E_{\perp}^2} \quad (3.110)$$

where $E_{\perp} = |\vec{E} \times \vec{J}|/|\vec{J}|$. Table 3.4 lists the model parameters.

TABLE 3.4 Model Parameters of the Lombardi Mobility Model

		Electrons	Holes	Unit
μ_{ac}	B	4.75×10^7	9.93×10^7	cm/s
	C	$1.74 \times 10^5 \times N_A^{0.125}$	$8.84 \times 10^5 \times N_D^{0.0317}$	N_A, N_D in cm^{-3}
μ_b	μ_0	52.2	44.9	$\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$
	μ_{\max}	1417	470.5	$\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$
	μ_1	43.4	29.0	$\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$
	γ	2.5	2.2	—
	C_r	9.68×10^{16}	2.23×10^{17}	cm^{-3}
	C_s	3.43×10^{20}	6.10×10^{20}	cm^{-3}
	α	0.680	0.719	—
	β	2.00	2.0	—
μ_{sr}	p_c	—	9.23×10^{16}	cm^{-3}
	δ	5.82×10^{14}	2.05×10^{14}	V/s

Source: After Lombardi et al., Ref. 46.

3.5.2 Drift Velocity under High Fields

At high longitudinal electric fields, optical phonon scattering dominates and the drift velocity saturates.^{47–49} The most common semiempirical models for the velocity–field relationship are^{4,47,48,49,50}

- Model A:

$$v = \frac{\mu_{\text{eff}} E}{\left[1 + \left(\frac{\mu_{\text{eff}} E}{v_s} \right)^\beta \right]^{1/\beta}} \quad (3.111)$$

where $\beta = 2$ for electrons and $\beta < 2$ for holes⁴.

- Model B:

$$v = \frac{\mu_{\text{eff}} E}{1 + \left(\frac{\mu_{\text{eff}} E}{v_s} \right)} \quad (3.112)$$

- Model C:

$$v = \begin{cases} \frac{\mu_{\text{eff}} E}{1 + \left(\frac{\mu_{\text{eff}} E}{v_s} \right)} & \text{for } E < v_s / \mu_{\text{eff}} \\ v_{\text{sat}} & \text{for } E > v_s / \mu_{\text{eff}} \end{cases} \quad (3.113)$$

where $v_s = 2v_{\text{sat}}$. The saturation velocity (v_{sat}) in silicon is about 6×10^6 to 1×10^7 cm/s for electrons and about 4×10^6 to 8×10^6 cm/s for holes.

Velocity saturation causes the drain current to saturate at a lower drain voltage than described from first-order models^{4,50} (see Section 3.2.3, Eq.3.40). The drain saturation voltage in the presence of velocity saturation can be derived using the simpler piecewise-linear model (model C).

Substituting Eq. 3.113 into Eq. 3.8, the drain current is

$$I_D = WC_{\text{ox}}(V_G - V_T - V(y)) \frac{\mu_{\text{eff}} E(y)}{1 + \frac{\mu_{\text{eff}} E(y)}{v_s}} \quad (3.114)$$

and the electric field along the channel can be written as

$$E(y) = \frac{I_D}{[W\mu_{\text{eff}}C_{\text{ox}}(V_G - V_T - V(y))]v_s/\mu_{\text{eff}}} \quad (3.115)$$

Integrating Eq. 3.115 from $y = 0$ to $y = L$ with $V(y = 0) = 0$ and $V(y = L) = V_D$, we obtain

$$I_D = \frac{W}{L} \mu_{\text{eff}} C_{\text{ox}} (V_G - V_T - V_D/2) V_D \frac{1}{1 + \frac{\mu_{\text{eff}} V_D}{v_s L}} \quad (3.116)$$

The drain saturation voltage, $V_{D\text{sat}}$, is then obtained by substituting $V_{D\text{sat}}$ as $V(y)$ and v_s/μ_{eff} as $E(y)$ of Eq. 3.114 and combining with Eq. 3.116:

$$V_{D\text{sat}} = \frac{(v_s L / \mu_{\text{eff}})(V_G - V_T)}{(v_s L / \mu_{\text{eff}}) + (V_G - V_T)} \quad (3.117)$$

The velocity-saturation-limited drain saturation current can then be written as

$$I_{D\text{sat}} = WC_{\text{ox}} v_{\text{sat}} \frac{(V_G - V_T)^2}{(V_G - V_T) + v_s L / \mu_{\text{eff}}} \quad (3.118)$$

The velocity-saturation-limited drain saturation current is smaller than the carrier pinchoff-limited drain saturation current. The $I_{D\text{sat}}$ dependence on $(V_G - V_T)$ goes from $(V_G - V_T)^2$ to $(V_G - V_T)$ depending on the value of $v_s L / \mu_{\text{eff}}$ compared to $(V_G - V_T)$. Thus, a signature of short-channel, velocity-saturated behavior is the linear dependence of $I_{D\text{sat}}$ on $(V_G - V_T)$.

3.6 PARASITIC EFFECTS

The section focuses on several parasitic effects that are important for MOSFETs in the ULSI era. These non-ideal effects were ignored in the first-order models in Section 3.2, but are of utmost importance for device design.

3.6.1 Hot Carriers

If MOSFETs are scaled according to constant-field scaling as described in Section 3.3.1, the electric fields in the MOSFET remains the same as devices are scaled to a smaller geometry. However, because of other constraints (such as the desire to keep a higher power supply voltage in comparison to the threshold voltage to improve performance), the electric fields inside a MOSFET typically rise as devices are scaled. Carriers in the channel gain energy as they travel from the source to the drain and as the electric field increases, so does the energy gained by the carriers. These high-energy carriers are typically described by an “effective temperature”⁵¹ and thus they are called “hot carriers”.^{52,53}

Experimental Observations

An extensive literature exists for both experimental and theoretical studies of hot carriers.^{54,55} Here we summarize the salient features. As the channel carriers gain energy above the impact-ionization threshold energy, electron–hole pair generation occurs as a result of impact ionization.^{56–58} The holes (in n-channel MOSFET) can be collected at the substrate terminal as substrate current (I_{sub}). Some carriers gain enough energy and momentum toward the Si/SiO₂ interface due to a redirection scattering that they may surmount the Si/SiO₂ interface energy barrier and be injected into the gate oxide as gate current (I_G).⁵⁹ Currents through the gate insulator create interface traps, which in turn cause degradation of device performances⁶⁰ in the form of threshold voltage shifts (ΔV_T), and degradation of the transconductance (Δg_m), the subthreshold slope (ΔS), and the drain current drive (ΔI_D). These interface traps have a nonuniform distribution along the channel (depending on the location of the source of hot carriers), resulting in asymmetric device degradation when the source and the drain terminals are reversed.⁵³ The device degradation also depends on whether the hot-carrier stress is dc or ac as in a switching MOSFET in a circuit.⁶¹ Some hot carriers lose energy and emit photons (the energy-loss mechanisms are discussed in Refs. 62–68). These photons, depending on their wavelengths, may travel far inside the silicon substrate before being absorbed in the substrate and generate electron–hole pairs. The photon-generated minority carriers can be collected by a nearby junction (I_{coll}).

Figure 3.22 shows the typical measured gate current and substrate current. The peak of the substrate current typically occurs at $V_{GS} = V_{DS}/2$. The asymmetric bell-shaped curve is the result of two competing mechanisms. As the gate voltage is increased from below the threshold voltage, the supply of hot carriers is increased, resulting in an increase of the measured substrate current. At higher gate voltages, the normal electric field at the drain increases, thereby reducing the longitudinal electric field (E_m) and the carriers become less energetic, thus reducing the measured substrate current at high gate biases. A similar competing mechanism exists for the gate current. At low gate voltages, the supply of energetic channel carriers is low. At gate voltages above the drain voltage, the direction of the normal electric field becomes unfavorable for carrier injection into the gate insulator. Therefore, the gate current drops as the gate voltage is raised beyond the drain voltage.

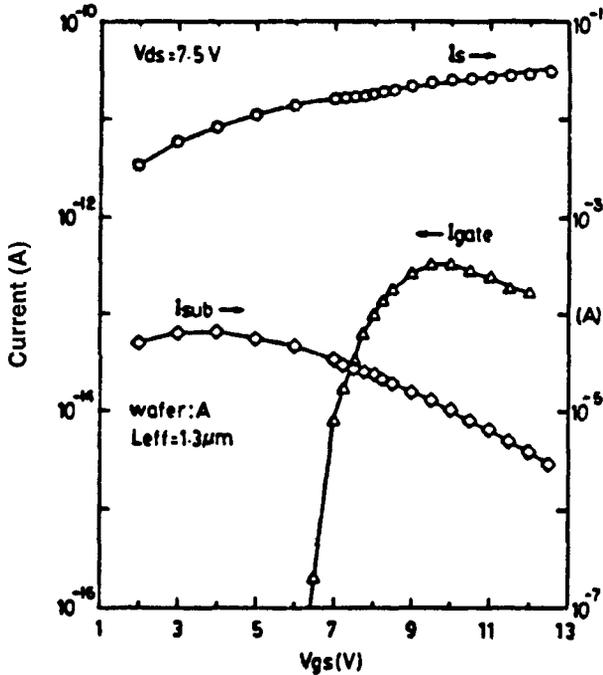


Figure 3.22 Measured source current (I_s), substrate current (I_{sub}), and gate current (I_{gate}) versus the gate-to-source voltage (V_{gs}) for a 1.3 μm device. (after Tam and Hu, Ref. 65.)

Figure 3.23 shows a typical I-V characteristics before and after hot-carrier stressing. The increase of V_T and S and decrease of g_m constitute device degradation. “Reverse” means that the source and drain terminals are reversed for testing after stressing. Figure 3.24 shows that the linear relationship among ΔV_T , Δg_m , and ΔS simplifies the characterization of degradations and suggests ΔN_{it} as the common cause for all three. Figure 3.25 shows the device lifetime after stress (lifetime is defined as $\Delta V_T = 10 \text{ mV}$) as a function of the substrate current for various device geometries and biases. Device lifetime and the substrate current follows a power-law relationship ($\tau \propto I_{sub}^\alpha$). This relationship simplifies the characterization of device lifetime as a function of device and bias parameters, and the substrate current is often used as a monitor of hot-carrier effects and device degradation. The proportionality constant, however, is technology-dependent.

Phenomenological Model

A phenomenological model⁵³ based on the “lucky electron” model^{52,69,70} is described here. More detailed physical models are described extensively in the literature (see, e.g., Ref. 55). The substrate current (I_{sub}), gate current (I_G), photon-generated minority current (I_{coll}), interface trap generation (ΔN_{it}), and the device lifetime (τ , defined as the time required to attain a degradation of device

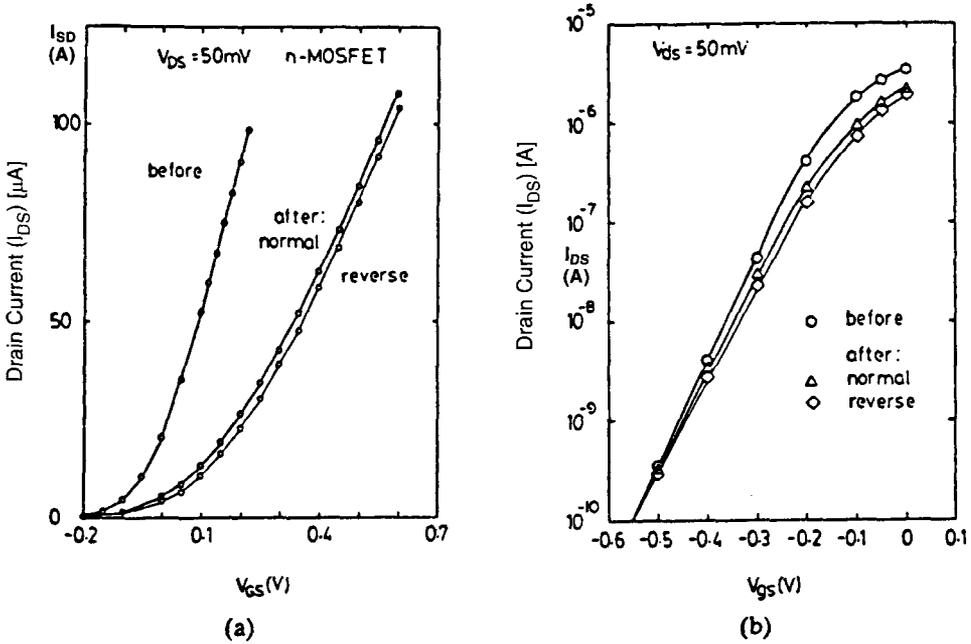


Figure 3.23 Typical I–V characteristics before (a) and after (b) severe hot-carrier stressing. Note the increase of V_T and S and the decrease of g_m . “Reverse” means that the source and drain terminals are reversed for testing after stressing. (after Hu et al., Ref. 53.)

performance such as threshold voltage shift) can be described by

$$I_{sub} = C_1 I_D e^{-\phi_i/q\lambda E_m} \tag{3.119}$$

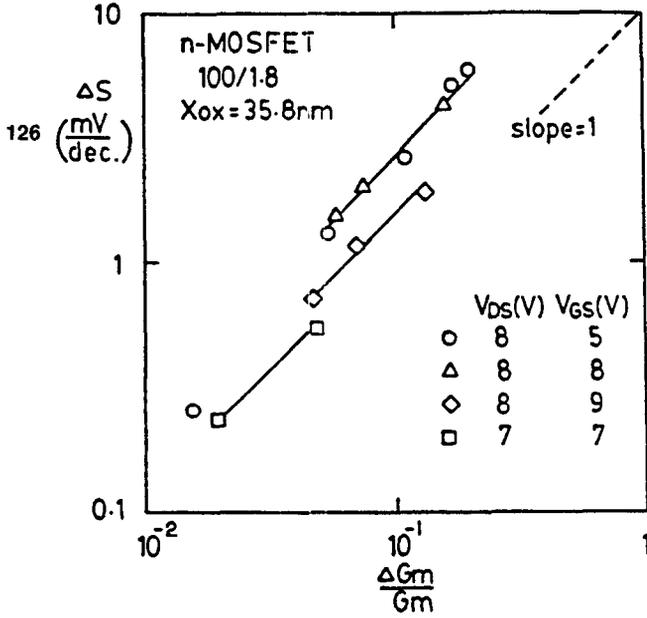
$$I_G = C_2 I_D e^{-\phi_b/q\lambda E_m} \tag{3.120}$$

$$I_{coll} = C_3 I_D e^{-\phi_{hv}/q\lambda E_m} \tag{3.121}$$

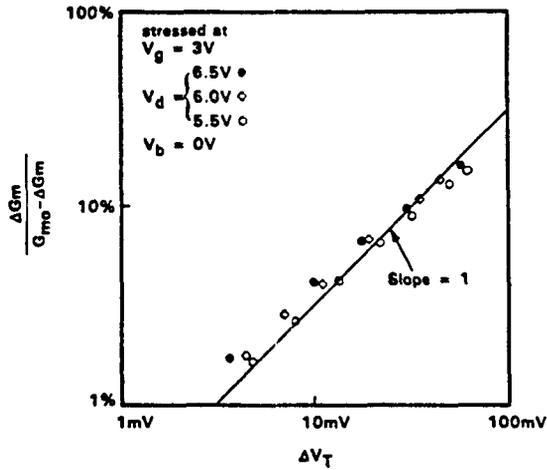
$$\Delta N_{it} = C_4 \left[t \frac{I_D}{W} e^{\phi_{it}/q\lambda E_m} \right]^n \tag{3.122}$$

$$\tau = C_5 \frac{I_D}{W} e^{\phi_{it}/q\lambda E_m} \tag{3.123}$$

where λ is the hot-carrier mean free path between scattering events, ϕ_i is the critical energy for impact ionization, ϕ_b is the critical energy to surmount the Si/SiO₂ interface energy barrier, ϕ_{hv} is the critical energy for photon emission, ϕ_{it} is the critical energy for interface trap generation, E_m is the maximum electric field along the direction of the drain current, and $C_1 - C_5$ are constants. $\phi_i \sim 1.3$ eV and



(a)



(b)

Figure 3.24 The linear relationship among ΔV_T , Δg_m , and ΔS simplifies the characterization of degradations and suggests interface trap generation as the common cause for all three. (after Hu et al., Ref. 53.)

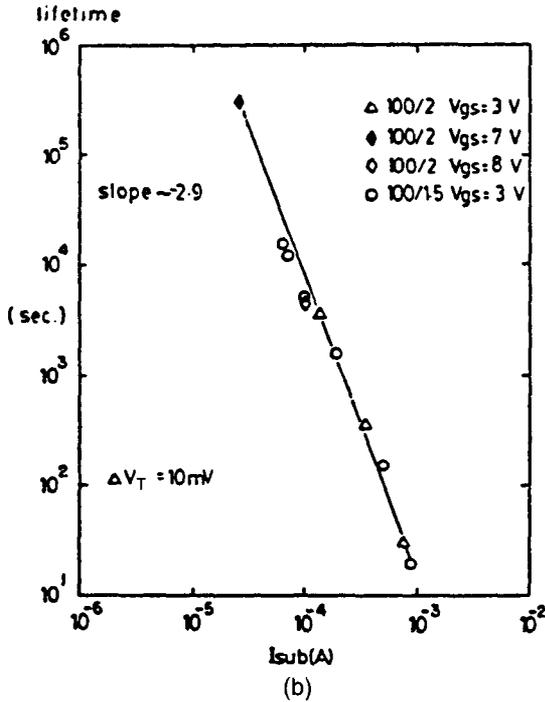
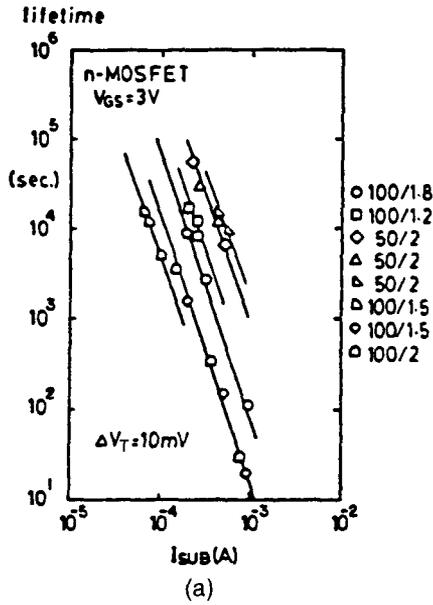


Figure 3.25 Plot of (a) device lifetime and the substrate current follows a power-law relationship ($\tau \propto I_{sub}^\alpha$); (b) the proportionality constant, which is technology-dependent. (after Hu et al., Ref. 53.)

depends on both the lateral and normal electric fields.⁷¹ ϕ_b is a function of the normal field in the gate oxide ($\phi_b = 3.2 - \beta\sqrt{E_{ox}} - \nu E_{ox}^{2/3}$, where E_{ox} is the normal field in the gate oxide, $\beta = 2.59 \times 10^{-4} (\text{V} \cdot \text{cm})^{1/2}$, $\nu \sim 4 \times 10^{-5} \text{V}^{1/3} \cdot \text{cm}^{2/3}$ (see Fig. 3.26)^{52,65} and is about 2–2.6 eV. Combining Eqs. 3.119 and 3.120, we obtain

$$\frac{I_G}{I_{sub}} = C_2 \left[\frac{I_{sub}}{C_1 I_D} \right]^{\phi_b/\phi_i} \simeq 4 \times 10^{-4} \left[\frac{I_{sub}}{I_D} \right]^{\phi_b/\phi_i} \quad (3.124)$$

Similarly, combining Eqs. 3.119 and 3.121, one obtains

$$\frac{I_{coll}}{I_D} \propto \left(\frac{I_{sub}}{I_D} \right)^{\phi_{hv}/\phi_i} \quad (3.125)$$

with $\phi_{hv}/\phi_i \sim 0.82$.⁶² The ratios ϕ_b/ϕ_i and ϕ_{hv}/ϕ_i can be obtained from the slope of the log–log plot of I_G/I_{sub} versus I_{sub}/I_D and I_{coll}/I_{sub} versus I_{sub}/I_D , respectively, as illustrated in Figures 3.27 and 3.28.

Energy Distribution Functions

The energy distribution of hot carriers is typically described by two approximations:⁵¹ (1) the lucky electron model and (2) the drifted Maxwellian model; both of these models are inadequate in most cases. The lucky electron model assumes weak scattering so that the electron reaches high energies during one ballistic flight in the electric field (E). Therefore, the value of the distribution

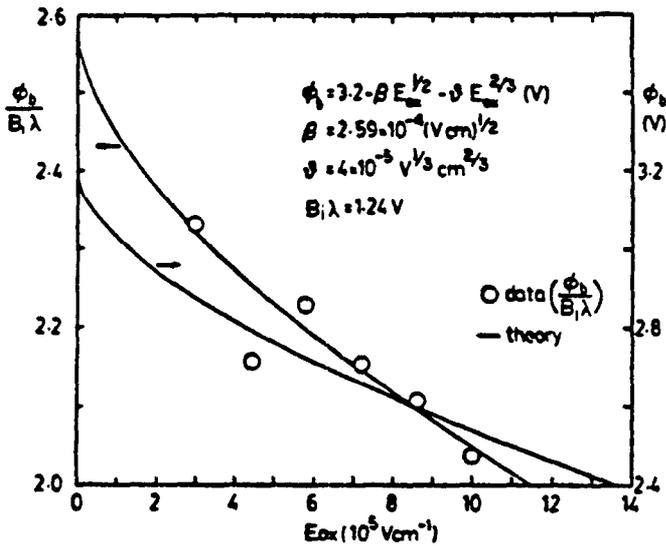


Figure 3.26 Dependence of the critical energy for surmounting the Si/SiO₂ energy barrier (ϕ_b) on the normal electric field (E_{ox}) (after Tam and Hu, Ref. 65.)

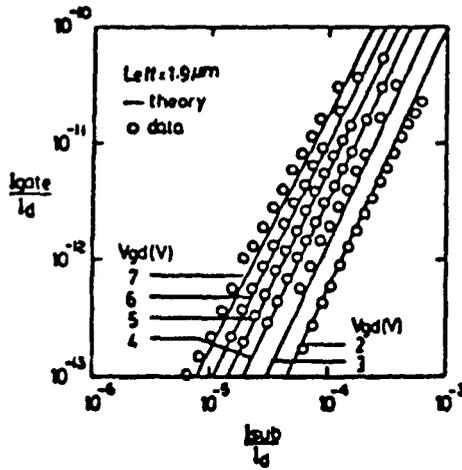


Figure 3.27 The correlation between the substrate current (I_{sub}) and the gate current (I_G) is universally independent of device parameters and bias. Each line is applicable for a specific oxide field, $E_{ox} = V_{GD}/t_{ox}$ ($t_{ox} = 82 \text{ nm}$ for the test device). Slope is $\phi_b(E_{ox})/\phi_i$, yielding $\phi_i = 1.3 \text{ eV}$. (after Hu et al., Ref. 53.)

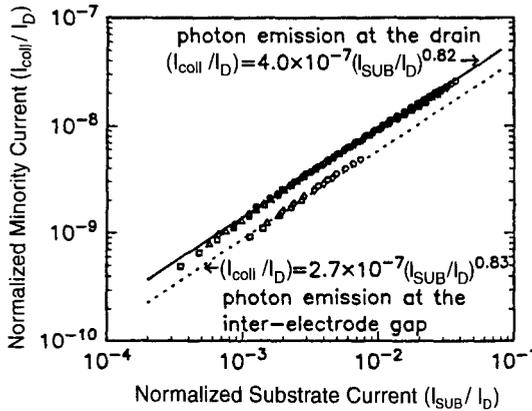


Figure 3.28 The correlation between the photon-generated minority current collected at a diode (I_{coll}) and the substrate current (I_{sub}). (after Wong, Ref. 62.)

function at a certain energy is given by the probability that an electron flies long enough without scattering to reach that energy. The energy dependence of the distribution function is given by

$$f_{LE}(\epsilon) \propto e^{-\epsilon/qE\lambda} \tag{3.126}$$

A drifted Maxwellian distribution function is the result of strong electron-electron scattering, the electron distribution is thermalized, and the acceleration in the

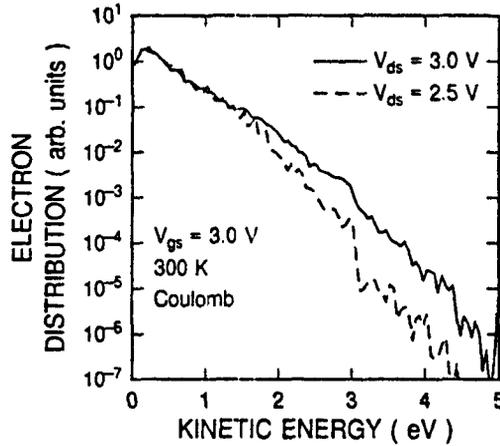


Figure 3.29 Electron energy distribution just inside the drain of a 0.15- μm n-MOSFET shows the presence of electrons above 3.2 eV at a drain bias of only 2.5 V. (after Fischetti et al., Ref. 55.)

electric field makes the temperature of the electrons T_e higher than the lattice temperature (T). The energy dependence is computed from a drifted distribution in momentum space:

$$f_{MW}(\varepsilon) \propto \frac{\sinh(\sqrt{2m\varepsilon}v_d/kT_e)}{\sqrt{2m\varepsilon}v_d/kT_e} e^{-\varepsilon/kT_e} \quad (3.127)$$

Figure 3.29 shows the electron energy distribution just inside the drain of a 0.15- μm n-MOSFET. It shows the presence of electrons above 3.2 eV even at a drain bias of only 2.5 V. The high-energy “tail” of the energy distribution function has been studied by Monte Carlo simulations.^{72,73} It is important to understand the observed experimental results of impact ionization by hot carriers at supply voltages below the “ionization threshold energy” (ϕ_i).

Drain Engineering

The most effective way to reduce hot-carrier degradation effects, apart from reducing the power supply voltage, is to use the lightly doped drain (LDD) structure.⁷⁴ The basic concept behind the LDD is to drop the drain voltage in a lightly doped drain region between the channel and the heavily doped drain, thereby reducing the maximum lateral electric field experienced by channel carriers.⁷⁵ Figure 3.30 shows the dependence of device lifetime and substrate current on the n^- implant dose in LDD devices. Although the substrate current, in general, correlates with the device lifetime, this correlation is not so simple in the presence of LDD because of structural differences such as the possibility of an underlap of the gate

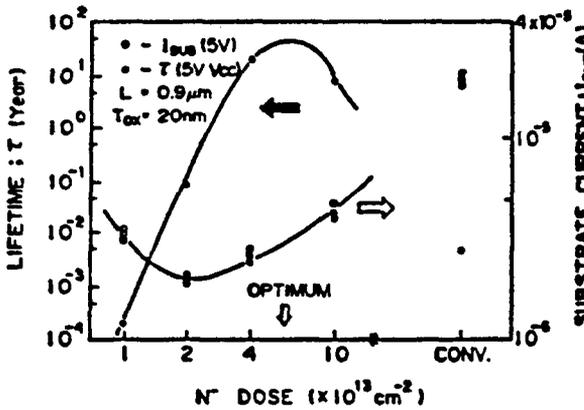


Figure 3.30 The dependence of device life time and substrate current on n^- implant dose in LDD devices. (after Toriumi, Ref. 54.)

and the drain region.^{53,54} The design of the n^- LDD implant dose must take into consideration the competing requirements to reduce (1) the carrier energy, (2) the series resistance, and (3) the gate-drain overlap capacitances.

3.6.2 Gate-Induced Drain Leakage (GIDL)

A subbreakdown leakage current is observable in the I_D/V_G characteristics of MOSFETs.⁷⁶⁻⁷⁹ Figure 3.31 illustrates this leakage current for MOSFETs biased in the OFF state. This leakage current [called *gate-induced drain leakage (GIDL)*] is attributed to band-to-band tunneling in the high-field regions in the silicon. While band-to-band tunneling is an effect to be avoided in MOSFET design to limit the transistor OFF current, it can be used to advantage as a low-voltage programming mechanism in non-volatile memory devices.^{80,81}

Band-to-Band Tunneling

Figure 3.32 shows the measured drain current characteristics for n-MOSFETs at different temperatures with the gate and substrate grounded. Region I is due to band-to-band tunneling in the high-field depletion region under the gate-drain overlap region. Figure 3.33 shows a schematic diagram of the high corner-field region and the associated band diagrams for n- and p-channel devices. For n-channel devices, electrons tunnel from the valence band to the conduction band when the applied gate-to-drain bias results in band bending beyond the silicon bandgap (similarly for holes in p-channel devices). The generated electrons travel laterally in the depleted drain and these carriers may be amplified by impact ionization if the carriers attain enough energy in the high-field region (as in a large drain bias). This impact-ionization-multiplied current contributes to the rise of drain current in region II of Figure 3.32. Depending on the doping concentration of the channel, the impact-

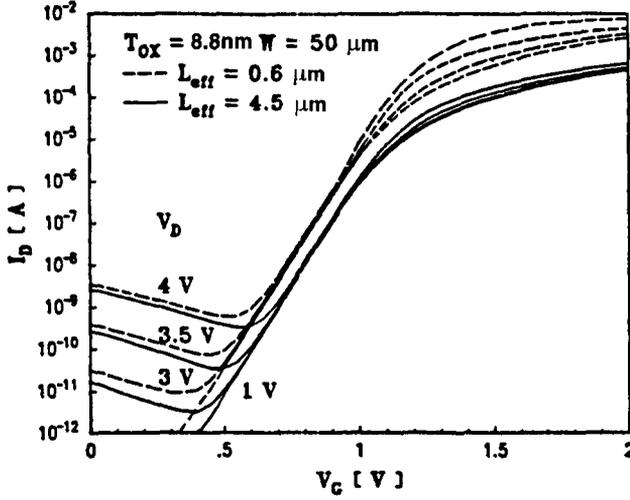


Figure 3.31 Subthreshold characteristics for two n-MOSFETs with $t_{ox} = 8.8 \text{ nm}$, $W = 50 \text{ }\mu\text{m}$, and $L = 4.5$ and $0.6 \text{ }\mu\text{m}$. Significant drain leakage currents can be observed when V_{DG} is high. (after Chen et al., Ref. 76.)

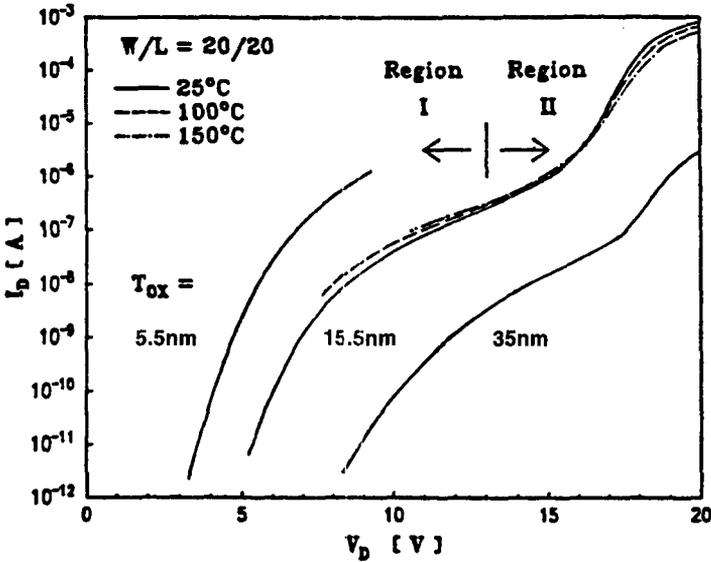


Figure 3.32 Drain current characteristics for n-MOSFETs at different temperatures with the gate grounded. A drain-voltage-independent leakage current, believed to be the thermal-generation current, has been subtracted for clarity. The drain current characteristics for devices with gate oxide thickness of 5.5 and 35 nm at room temperature are also included. (after Chen et al., Ref. 76.)

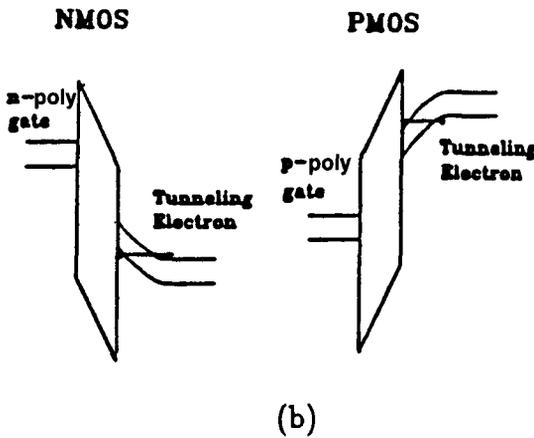
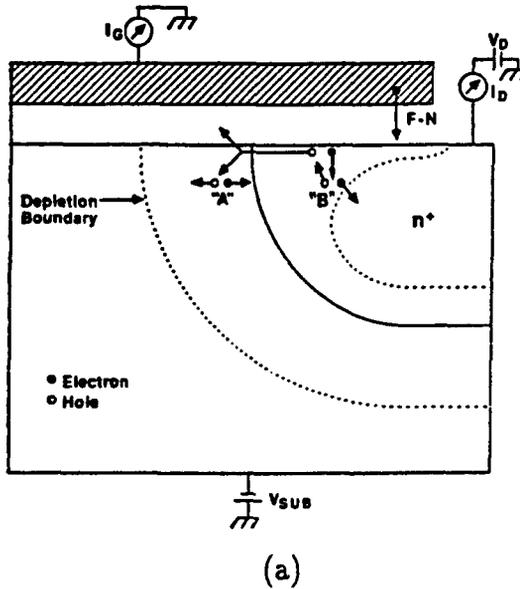


Figure 3.33 (a) A deep-depletion region is formed in the gate–drain overlap region. When channel doping is high, impact ionization is initiated by band-to-band tunneled holes in the lateral direction (process A). When channel doping is low, impact ionization is initiated by band-to-band tunneled electrons in the vertical direction (process B). (b) The energy-band diagram illustrates the band-to-band tunneling process in silicon in the gate–drain overlap region. Electron–hole pairs are generated by the tunneling of valence-band electrons into the conduction band and collected by the drain and the substrate separately. (after Chan et al., Ref. 77.)

ionization-multiplied band-to-band tunneling current can be initiated by either the normal electric field (for the case of low channel dopings, process B of Fig. 3.33) or the longitudinal electric field (for the case of high channel dopings, process A of Fig. 3.33) [79]. The impact-ionization-multiplied band-to-band tunneling current is greater when the band-to-band tunneled carrier is multiplied via process A, as shown in the experimental results in Figure 3.34.

Band-to-band tunneling can be used to generate hot carriers for injection into the gate insulator for programming of nonvolatile memory devices.^{80,81} Figure 3.35 illustrates the carrier injection process. When the electron (labeled A) band-to-band tunnels into the conduction band, the hole A' left behind travels across the high-field region and creates electron-hole pairs (B and B') via impact ionization. The generated electron B will drift back toward the Si/SiO₂ interface and may gain enough energy to travel across the SiO₂ barrier and become gate current (labeled C).

Band-to-Band Tunneling Model

The band-to-band tunneling current can be expressed as⁷⁶

$$\begin{aligned}
 I_D &= AE_s e^{-\frac{\pi m^* 1/2 e^3 V_g^2}{2\sqrt{2}qhE_s}} \\
 &= AE_s e^{-B/E_s}
 \end{aligned}
 \tag{3.128}$$

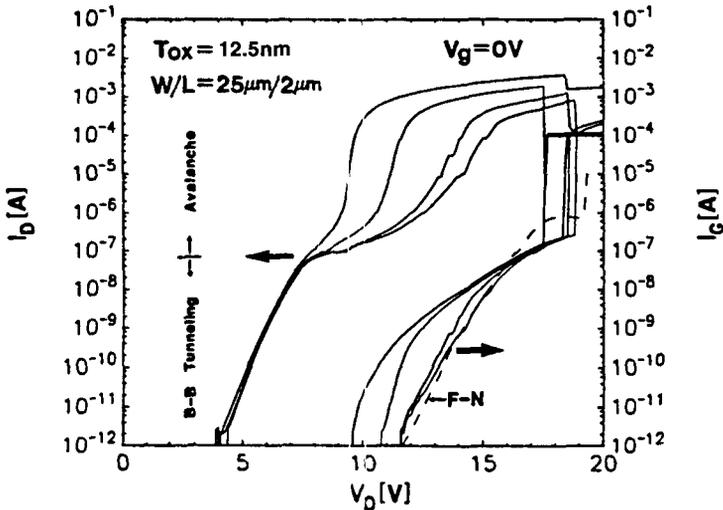


Figure 3.34 Measured drain and gate currents for n-channel MOSFETs with $t_{ox} = 12.5$ nm, and channel surface concentrations of $10, 5, 2,$ and $1 \times 10^{16} \text{ cm}^{-3}$ (from left to right). The dashed curve represents the normalized gate current (Fowler–Nordheim tunneling current) measured with $V_D = V_{sub}$ and the gate grounded. Drain I–Vs are clamped at high current levels due to a substrate resistance effect. (after Chang and Lien, Ref. 78.)

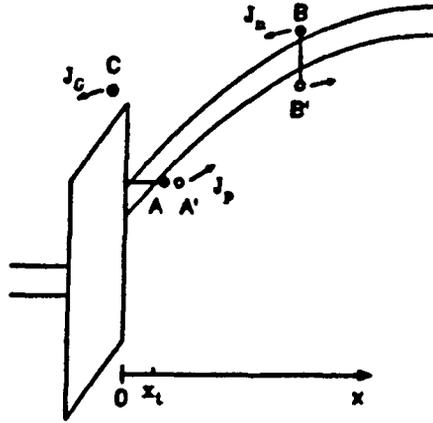


Figure 3.35 Energy-band diagram illustrating the mechanisms of the band-to-band tunneling-induced substrate hot-electron (BBISHE) injection. Electron A is a tunneling electron, A' the hole, and B and B' the electron-hole pair created by impact ionization. When the electron (labeled A) band-to-band tunnels into the conduction band, the hole A' left behind travels across the high-field region and creates electron-hole pairs (B and B') via impact ionization. The generated electron B will drift back toward the Si/SiO₂ interface and may gain enough energy to travel across the SiO₂ barrier and become gate current (labeled C). (after Chen and Teng, Ref. 81).

where $B = 21.3 \text{ MV/cm}$ with $m^* = 0.2 m_0$, and E_s the electric field in the silicon. For tunneling due to a normal electric field under the gate-drain overlap region, the electric field can be expressed as

$$E_s \simeq \frac{\epsilon_{\text{ox}}(V_{DG} - E_g)}{\epsilon_{\text{Si}}t_{\text{ox}}} \tag{3.129}$$

To model the gate current injection due to band-to-band tunneling-induced substrate hot-electron injection, the electron density at location B and the injection probability of the electron from B has to be calculated. Assuming that the additional holes created by impact ionization is negligible compared to the band-to-band tunneling current [$J_p(x) \simeq J_p(x_t) = J_{BB}$], the gate current density is

$$J_G = \int_{x_t}^{\infty} J_{BB} \times M(x) \times P(x) \frac{dx}{W} \tag{3.130}$$

The impact-ionization multiplication factor for a hole traveling from position x can

be expressed as⁸¹

$$M(x) = \frac{1}{1 - \int_x^{\infty} \alpha_p(x') dx'} - 1 \quad (3.131)$$

where α_p is the impact ionization coefficient for holes and can be expressed as

$$\alpha_p = \alpha_0 e^{-B_p \lambda_p / \varepsilon(x)} \quad (3.132)$$

where λ_p (= 4.6 nm at room temperature) is the mean free path for holes in silicon and $\varepsilon(x)$ is the average hole energy at position x . In a depletion region with constant doping concentration N_{sub} , $\varepsilon(x)$ can be expressed as⁸¹

$$\begin{aligned} \varepsilon(x) = E_{\text{max}} \lambda_p (1 - e^{-(x-x_t)/\lambda_p}) + \frac{qN_{\text{sub}}}{\varepsilon_s} \lambda_p^2 (1 - e^{-(x-x_t)/\lambda_p}) \\ - \frac{qN_{\text{sub}}}{\varepsilon_s} \lambda_p (x - x_t) e^{-(x-x_t)/\lambda_p} \end{aligned} \quad (3.133)$$

The substrate hot-electron injection probability⁵² can be generalized to calculate the probability at a position x (from the Si/SiO₂ interface)

$$P(x) = A e^{-x/\lambda_n} \quad (3.134)$$

where λ_n (= 9.1 nm at room temperature) is the mean free path of electrons in silicon.

3.6.3 Gate Capacitance Degradation

The gate capacitance is an importance measure of the transconductance and the current drive attainable by a MOSFET. For MOSFETs scaled to small dimensions, the gate capacitance is no longer simply the gate oxide capacitance, $C_{\text{ox}} = \varepsilon_0/t_{\text{ox}}$, but can best be described by an equivalent electrical thickness, $t_{\text{eq}} = \varepsilon_0/C_{\text{inv}}$, where C_{inv} is the gate capacitance at inversion. Figure 3.36 shows the calculated electrical equivalent oxide thickness (t_{eq}) versus physical oxide thickness (t_{ox}) curves for several combinations of substrate dopings and gate polysilicon dopings.⁸² The calculation compares the classic [Maxwell–Boltzmann distribution (MB) and Fermi–Dirac distribution (FD)] assumptions and the more accurate quantum-mechanical (QM) description. For thin physical oxide thickness, the discrepancy between physical oxide thickness and the electrical thickness is significant.

The gate capacitance is degraded by two effects:

1. Inversion-layer broadening due to quantum confinement of the 2D electron gas^{9,41}
2. Depletion of the polysilicon gate⁸³

These effects are described in the following two sections.

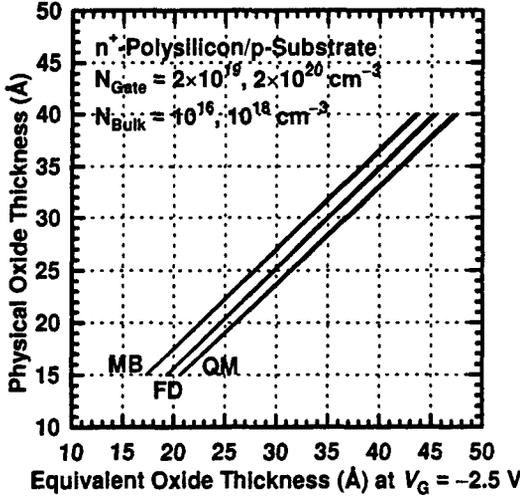


Figure 3.36 Calculated (QM) $t_{ox} - t_{eq}$ curves for n^+ -polysilicon/p-Si MOS devices. Two other groups are based on the classic model with Fermi–Dirac (FD) and Maxwell–Boltzmann (MB) statistics, respectively, for comparison. Each group contains four curves. (after Lo et al., Ref. 82.)

Inversion-Layer Capacitance

The small-signal transconductance of the MOSFET is given by⁷

$$\begin{aligned}
 g_m &= \left(\frac{W}{L}\right) \int_0^{V_{DS}} dV(y) \underbrace{\left\{ \frac{\mu_0 [1 + (2\theta_s/C_{ox})(Q_s - Q_n(C_s/C_i))]}{[1 + (\theta_s/C_{ox})(2Q_s - Q_n)]^2} \right\}}_{\mu_{FE}} \underbrace{\left\{ \frac{C_i C_{ox}}{C_{ox} + C_s + C_i} \right\}}_{C_{GC}} \\
 &= \left(\frac{W}{L}\right) \int_0^{V_{DS}} dV(y) \mu_{FE} C_{GC} \tag{3.135}
 \end{aligned}$$

The inversion-layer capacitance, $C_i = -(\partial Q_n / \partial \phi_s)$, becomes non-negligible compared to the gate oxide capacitance, C_{ox} , as gate oxides are scaled down below 10 nm or when the temperature is lowered. Figure 3.37 shows that the channel charge is reduced by quantum mechanical broadening of the inversion layer.⁴¹ The electron concentration peaks at the Si/SiO₂ interface in the classic 3D density-of-states approximation, whereas the electron concentration peaks away from the surface in the quantum-mechanical 2D density-of-states case.⁹

Assuming a triangular potential well, a 2D density of states for (100) silicon, and the inversion charge populating the lowest subband only (a good approximation at

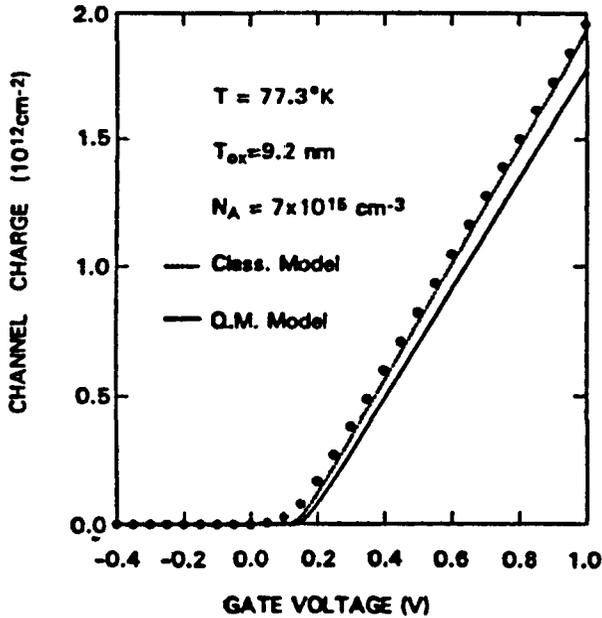


Figure 3.37 Channel charge versus gate voltage at 77 K. The solid line represents the quantum-mechanical model, while the dotted line represents the classical theory. (adapted from Bacarani and Wordeman, Ref. 41.)

low temperatures), the small-signal inversion layer capacitance is⁴¹

$$\begin{aligned}
 C_i &= -\frac{\partial Q_n}{\partial \phi_s} \\
 &= \frac{1}{\sqrt{2}} \frac{\epsilon_s \alpha F_{1/2}(u_s - u^*)}{L_D F(u_s, v)} \quad (3.136)
 \end{aligned}$$

where $u^* = (E_C - E_{Fn})/kT = (E_C - E_F)/kT + v$ is the normalized distance of the conduction-band edge in the silicon substrate from the quasi-Fermi level within the inversion layer, v is the normalized electron quasi-Fermi potential, $L_D = (\sqrt{\epsilon_s kT/q^2 N_A})$ is the Debye length, $\alpha = (2/\sqrt{\pi})N_C/N_A$, N_C is the effective density-of-states in the conduction band, $F(u, v) = \{ \frac{2}{3} \alpha [F_{3/2}(u - u^*) - F_{3/2}(u_i - u^*)] + u - u_i + L_D^2 u_i^2 / 2 \}^{1/2}$, $u_i = q\phi_i/kT$ is the normalized intrinsic potential, $u_i' = -qE_i/kT$ is the normalized electric field at $x = x_i$, and $F_{1/2}, F_{3/2}$ are Fermi integral of orders $\frac{1}{2}$ and $\frac{3}{2}$, respectively.

Note that the inversion-layer capacitance is not the same as ϵ_s/x_{av} , where x_{av} is the average distance of all the inversion charge. The small-signal, inversion-layer capacitance is given by the average distance at which the *incremental* charges are added as the gate voltage changes. When the inversion charge populates higher subbands (e.g., room temperature and high gate biases), the capacitance due to the density of states of the higher subbands has to be taken into account as well.

Polysilicon Gate Depletion

For MOSFETs with a heavily doped polysilicon gate, the polysilicon is either depleted or accumulated depending on the gate bias. In either case, the finite thickness of the depletion–accumulation region contributes to a reduction of the gate capacitance because the depletion–accumulation capacitance of the polysilicon gate is in series with the gate oxide capacitance.⁸³ This effect becomes significant as the gate oxide becomes thinner in the ULSI regime. It is common to employ an n^+ -gate polysilicon doping for n-channel MOSFETs and p^+ -gate polysilicon doping for p-channel MOSFETs;^{13,84} thus the polysilicon gate is depleted when the MOSFET channel is in inversion. The depletion capacitance of the polysilicon gate varies as the square root of the doping density of the polysilicon gate. Figure 3.38 shows the ratio of gate capacitance to gate oxide capacitance as a function of the doping of the polysilicon gate. The results highlight the observations that (1) gate capacitance is degraded for thin gate oxides due to polysilicon depletion effects and (2) it is extremely important to control the doping of the polysilicon gate to provide uniformity of the electrical gate oxide thickness. It is particularly difficult to achieve high doping for p^+ polysilicon gates because of the segregation of boron to the gate oxide and possible penetration of the boron into the channel during high-temperature activation anneals.

3.6.4 Gate Tunneling Current

As the physical gate oxide thickness is scaled to below 3 nm in ULSI devices, significant direct tunneling occurs under high gate biases between the MOSFET channel and the gate. Figure 3.39 shows the experimental and calculated gate tunneling current for various gate oxide thicknesses used in ULSI MOSFETs. The calculations were performed by solving, self-consistently, the Poisson equation and the Schrödinger equation for the 1D, gate oxide silicon system considering the

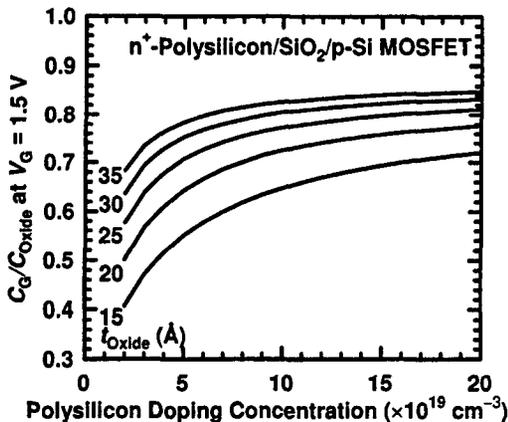


Figure 3.38 Calculated total capacitance to oxide capacitance ratio versus polysilicon doping concentration with oxide thickness as a parameter (1.5–3.5 nm). (after Lo et al., Ref. 82.)

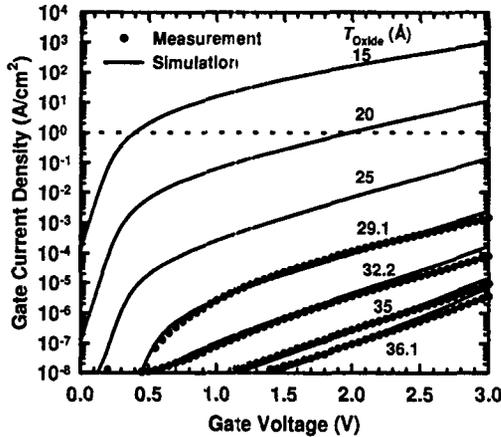


Figure 3.39 Measured and simulated gate tunneling current ($I_G - V_G$) characteristics under inversion conditions. The dotted line indicates the $1\text{-A}/\text{cm}^2$ limit for leakage current, which is usually taken as the limit for gate current for low standby-power dissipation.⁸⁵ (Figure after Lo et al., Ref. 82.)

polysilicon depletion effect in the gate material and 2D quantization effects in the silicon channel (see Section 3.6.3). The direct-tunneling current increases exponentially with decreasing gate oxide thickness. This gate tunneling current will impose a lower limit of standby-power dissipation and an upper limit of the number of active FETs on a chip for a given standby-power specification. For ULSI systems, a gate tunneling current of about $1\text{ A}/\text{cm}^2$ is usually considered acceptable for high-performance microprocessor applications.⁸⁵

3.6.5 Output Resistance

From first-order models described in Section 3.5.2, the drain current saturates (output resistance tends to infinity) when the drain voltage is larger than the drain saturation voltage. In practice, channel length modulation (CLM) will limit the output resistance ($r_o = (\partial I_D / \partial V_D)^{-1}$) to a finite value at the low-drain-voltage region, while drain-induced barrier lowering (DIBL) and substrate-current-induced body effect (SCBD) limit the output resistance at moderate and high-drain-voltage regions.⁸⁶⁻⁸⁹ A typical drain current and output resistance characteristics is illustrated in Figure 3.40, in which the regions where these effects dominate are marked. The finite output resistance has important ramifications for analog signal processing applications where the small-signal voltage gain ($g_m r_o$) is determined to a great extent by the output resistance.⁹⁰

The drain current in the presence of channel length modulation can be described in a simple model by

$$I_D = I_{D\text{sat}} \left[\frac{L}{L - \Delta L} \right] \quad \text{for } V_D > V_{D\text{sat}} \quad (3.137)$$

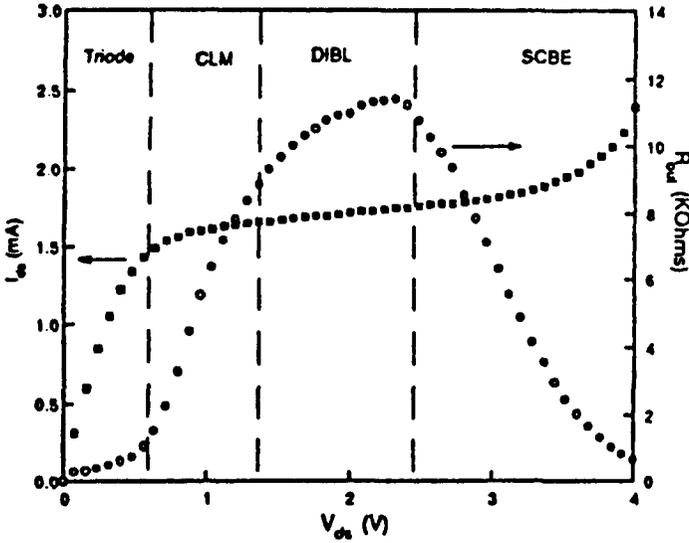


Figure 3.40 Typical drain current and output resistance. $W/L = 10/0.43$, $t_{ox} = 7.5$ nm. (after Huang et al., Ref. 86.)

where ΔL is the channel length shortening due to the depletion of mobile carriers (pinchoff) at the drain. It can be derived by ignoring the vertical electric field, assuming the mobile carrier density to be zero in the pinchoff region, and assuming a one-sided abrupt junction for the drain junction as

$$\Delta L = \left[\frac{2\epsilon_s}{qN_A} (V_D - V_{Dsat}) \right]^{1/2} \tag{3.138}$$

This expression overestimates ΔL . It also presents some unphysical results because of the pinchoff assumption (mobile charge is assumed to be zero in the pinchoff region). Ko⁴ analyzed the channel length modulation in more detail by including the mobile carriers and the effects of velocity saturation (see Section 3.5.2) and arrived at the following expression for ΔL

$$\Delta L = l \ln \frac{[(V_D - V_{Dsat})/l] + E_m}{v_s/\mu_{eff}} \tag{3.139}$$

$$l = \sqrt{\frac{\epsilon_s}{\epsilon_o} t_{ox} x_j} \tag{3.140}$$

$$E_m = \left[\frac{(V_D - V_{Dsat})^2}{l^2} + \left(\frac{v_s}{\mu_{eff}} \right)^2 \right]^{1/2} \tag{3.141}$$

and the drain current is

$$I_D = I_{D\text{sat}} \left[\frac{L}{L - \Delta L} \right] \left[\frac{V_{D\text{sat}} + v_s L / \mu_{\text{eff}}}{V_{D\text{sat}} + v_s (L - \Delta L) / \mu_{\text{eff}}} \right], \quad \text{for } V_D > V_{D\text{sat}} \quad (3.142)$$

The output resistance can then be computed from Eq. 3.142 as

$$r_o = \frac{1}{I_D} \frac{(V_D - V_{D\text{sat}})}{l} L + \frac{\mu_{\text{eff}} V_{D\text{sat}}}{v_s} \quad (3.143)$$

The output resistance in the presence of drain-induced barrier lowering and substrate-current-induced body effect were derived and compared with experimental data in Ref. 86.

3.7 EVOLUTION OF MOSFET DEVICE STRUCTURES

3.7.1 Historical Development

It is rather amazing that CMOS has had the same basic device structure through decades of development. The seemingly straightforward path of CMOS scaled toward 100 nm has in fact involved a tremendous amount of technological innovations not obvious in the device structure. Table 3.5 summarized major milestones in the structural and technological advances of CMOS.

The basic self-aligned polysilicon gate MOSFET was introduced in 1970 with ion implantation being used for the source and drain regions, and soon after for the channel as well. By the end of the decade, silicided polysilicon gate, self-aligned CMOS was fast becoming the industry standard. Dennard et al.⁵ proposed the scaling criteria for MOSFETs in the early 1970s, and by and large this has been adhered to, except for voltage, which has not been reduced as fast as stipulated (proportional to gate length) in the interests of voltage standardization and performance. Hot-carrier reliability issues were the main driving force behind reduced power supply voltages in sub-0.5- μm lithography.

The major difference, which set the submicrometer MOSFET apart from its predecessor, is the introduction of the sidewall spacer, enabled by reactive ion etching. For example, self-aligned silicide was introduced in the early 1980s, shallow trench isolation finally replaced all the variations of LOCOS,¹⁰⁶ and chemical-mechanical polishing (CMP)-based processes took over the backend metallization. There are also more subtle evolutions: substrate engineering became widely popular in the 0.25- μm regime,^{13,84} so was the source-drain engineering in 0.1- μm generations,^{102,107,108} although the idea was introduced much earlier in the form of lightly doped drain (LDD) to reduce hot-carrier-induced device degradation. Even the role of the spacer evolved with time; the spacer was first introduced to realize LDD. After the industry decided that it could freely scale the internal power

TABLE 3.5 Milestones in Bulk CMOS Technology^a

Year	Technology	Channel Length (μm)	Reduction to Practice	Reference
1960	MOSFET	—	Lab	91
1965	64-bit MOSFET SRAM	—	Lab	91
1966	Polysilicon gate self-aligned FET	—	Lab	92
1968	One transistor memory cell	—	Lab	93
1969	Ion-implanted channel	—	Lab	92
1970	CMOS Watch chip	—	Product	91
1971	Intel 4004 microprocessor	10	Product	94
1979	Silicided polysilicon gate	1	IEDM	95
1980	Sidewall spacer for S/D implant	1.5	IEDM	74
1982	Self-aligned silicide (salicide)	—	IEDM	96
1982	Trench isolation	—	IEDM	97
1983	Oxynitrides for gate dielectric	—	Lab	98
1985	Halo doping of source–drain junction	0.2	Lab	99
1986	N ⁺ -P ⁺ polysilicon gates	0.5	IEDM	100,101
1986	Retrograde channel doping	0.5	IEDM	100
1987	0.1- μm MOSFET	0.1	IEDM	102
1989	Chemical–mechanical polishing	—	IEDM	103
1992	Damascene Interconnect Technology	—	IEDM	104
1993	Copper interconnect	—	IEDM	105

^a This table reports the technological advances that have been included in today's CMOS production process, usually by their date of first appearance at IEDM, in the context of MOS VLSI devices.

Source: After Wong et al., Ref. 11.

supply voltage, LDD was abandoned, but the spacer stayed, and became a requirement for the salicide processes.

The 100-nm barrier was broken by Sai-Halasz et al.^{102,109} in 1987. By 1997, there have been many research articles describing sub-100-nm-gate-length MOSFETs.^{14, 107,108,110–117}

3.7.2 State-of-the-Art Bulk MOSFET

Several recent review articles have summarized the characteristics of state-of-the-art CMOS technology.^{10,85,118} Figure 3.41 illustrates most of the important features. The gates are fabricated with n- or p-type polysilicon so that both n-FETs and p-FETs are surface-channel devices, for maximum speed performance. The gates are topped with a metal silicide for lower gate series resistance, although the resistance is still higher than is desirable for maximum RF performance. Special lithographic techniques are used to pattern the gates with minimum dimensions that are 30–40% smaller than the general lithographic-feature size. The gate dielectric must be very thin, typically around 3 nm for the 0.1- μm gate length generation of technology.

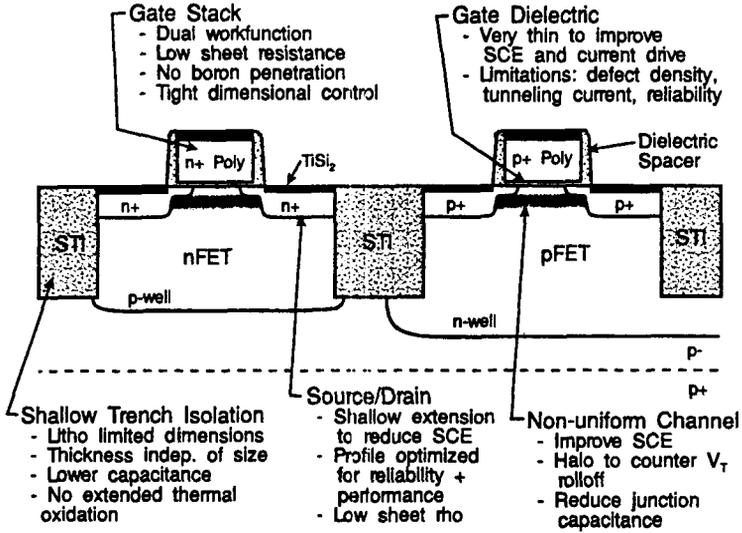


Figure 3.41 A cross-sectional diagram of typical state-of-the-art CMOS technology, indicating some of the more important features. (adapted from Davari et al., Ref. 10 and Wong et al., Ref. 11.)

Scaling requires such thin oxides to adequately limit short-channel effects, and to provide sufficient performance.

Shallow-trench isolation between devices involves etching trenches, filling them with deposited oxide, and polishing to planarize the devices. This process allows devices to be placed much closer together than was formerly possible with local oxidation of silicon, resulting in higher circuit density. The source and drain use shallow, moderately doped extensions under the gate edges and gate sidewalls. These are engineered to reduce short-channel effects and improve reliability with regard to hot electrons, and the deeper-contact implants allow low contact resistance.

Of great importance in achieving the shortest possible channel lengths is the engineering of the doping profiles in the channel region. Retrograde doping profiles can reduce the transverse electric field in the channel (improving mobility), and, at the same time, reduce two-dimensional effects by shielding the drain potential from the channel region. Superhalos⁸⁵ in the source and drain regions can be used to, at least partially, cancel 2D-induced threshold-voltage shifts (V_T), resulting in less V_T rolloff.

Figure 3.41 shows a cross section of a typical CMOS device. It does not show the wiring levels, but the wires are obviously very important in creating large integrated circuits and substantial technological progress is occurring there, too.^{119,120} By the year 2000, it is anticipated that most of the wiring will use copper, for its lower resistivity (40% lower than Al/Cu), lower processing cost and reduced electromigration.¹²¹ Lower-dielectric-constant insulators are also being considered to reduce wiring capacitance and improve speed. Materials such as spin-on-glass (FOX, Xerogel) and some polymers have dielectric constants ranging from 3 to

1.8^{120,122,123} and are actively being investigated for use as intermetal dielectrics. The use of a hierarchy of wiring sizes, from very fine wires at minimum lithographic dimension on the bottom of the wiring hierarchy to large “fat” wires on the top of the wiring hierarchy, is expected to prevent wiring from becoming a major bottleneck in future ICs.¹²⁴

3.7.3 Advanced MOSFET Device Structures

As CMOS is scaled into the nanometer regime, it is apparent that the use of dopants to control the conducting channel is reaching its limits. Device structural changes (to thin SOI, ground-plane FET, and double-gate FET) will alleviate some constraints and extend CMOS for two or three more generations but carries with them a different set of difficulties. This section provides an overview of these advanced MOSFET device structures.

Most of the advanced device structures involve use of the silicon-on-insulator (SOI) concept. SOI has many variants and one can discern an evolutionary path from very bulklike, to very venturesome structures. This path is illustrated in Table 3.6.

Partially depleted (PD) SOI¹²⁵ is very similar to bulk CMOS with the SOI island taking the place of the n or p tubs. The SOI is thick enough that the channel counter doping forms a conducting “body” under the FET channel whereas the source and drain implants usually penetrate to the back interface. Fully depleted (FD) SOI has a thinner and/or a more lightly doped SOI layer than PD SOI so that there is normally a negligible concentration of holes or electrons in an n- or p-channel FET. As illustrated in Fig. 3.42, FD SOI is predicted to have a much stronger drain-induced barrier lowering than PD SOI for all but the thinnest (< 40-nm) SOI layers. Both PD SOI and FD SOI will be treated in detail in another chapter.

The SOI structure can easily be extended to include a conducting layer underneath the silicon layer.^{126,127} This layer is called a *ground plane* or back-gate. The back-gate may also be a doped well in the substrate itself.¹²⁸ The back-gate serves two purposes:

1. It screen the channel from the bottom against penetration of the drain field into the source and thus facilitate scaling to shorter channel lengths. In a way it replaces the substrate of the heavy retrograde doping of the conventional bulk FET except that an insulating layer replaces the pn junction. This removes one of the limitations to scaling of the bulk FET, which is Zener breakdown of that junction. To provide effective screening the back-gate insulator should be very thin, within about twice the thickness of the front-gate oxide. The back-gate insulator increases the “body effect,” where the shift in threshold voltage resulting from the channel to back-gate voltage, and also increases the subthreshold slope factor due to the capacitive division effect, $C_g/(C_g + C_{sub})$, between the gate and the substrate.¹²⁹ Unlike the bulk substrate, however, the back-gate can be partitioned and locally connected to the source to avoid the body effect.
2. Another use for the back-gate is as a means of shifting the threshold voltage of the top-gate. The top-gate threshold voltage may be controlled over the range

TABLE 3.6 SOI Variants

Variant	Strengths	Weaknesses
Partially depleted SOI	Channel design bulklike V_T insensitive to box Interface	Very susceptible to floating body effects (but solutions are available) Same scaling constraints as bulk
Fully depleted SOI	Elimination of floating-body effects Elimination of punchthrough currents Elimination of drain-body tunneling	V_T sensitive to SOI thickness and back interface Back-channel potential may be influenced by drain voltage Difficulty of contacting thin SOI
Ground plane (GP)	Same as FD SOI GP shields channel from drain GP permits electrical control of V_T GP may be used as second gate	V_T sensitive to SOI thickness Difficulty of contacting thin SOI Degradation of subthreshold slope by close GP
Double gate (DG)	Maximum electrostatic control of channel and best scaling potential Best current drive and performance OR logic function within single device	Difficult to fabricate Misaligned top and bottom gates results in extra capacitance and loss of current drive V_T control difficult by conventional means
Stacked SOI (ST)	High functional density Shorter wires therefore higher performance and lower power	Fabrication complexity Difficult to cool

between strong accumulation and strong inversion of the back interface for the SOIAS structure.¹²⁶ For this purpose, the bottom insulator may be much thicker with a much smaller capacitive division ratio.

The double-gated (DG) FET is electrostatically much more robust than the standard single-gated MOSFET because the gates shield the channel from both sides suppressing penetration of the field from the gate, reducing short-channel effects.^{130–133} For conventional, single-gated, FETs the substrate acts as the bottom shield; yet this results in a tradeoff¹³⁴ between the degree of shielding and the reduction of the subthreshold slope, as discussed above. The double-gate FET does not have this limitation, and both gates are strongly coupled to the channel to increase transconductance. The relative scaling advantage of the DG FET is about 2 times the single-gate FET.¹³¹ The performance of the symmetrical version of the DG FET is further increased by higher channel mobility compared to a bulk FET since

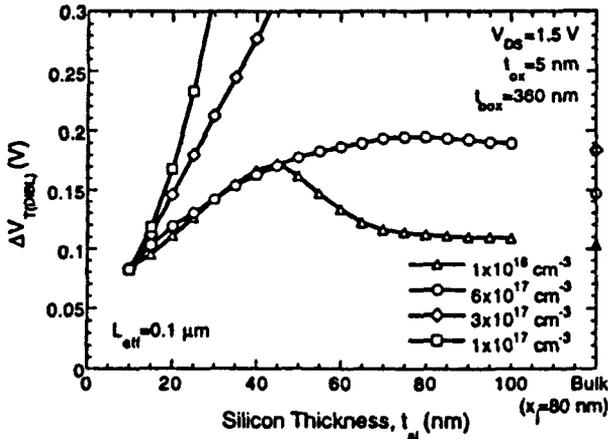


Figure 3.42 Threshold voltage shift due to drain-induced barrier lowering as a function of Si thickness for various channel dopings. (after Su et al., Ref. 154.)

the average electric field in the channel is lower. The lower electric field reduces interface roughness scattering according to the universal mobility model.^{34,35}

Different types of double-gate FET may also be classified according to the electrical function of the different layers and how they control the dimensionality and electric field configuration in the channel. Such a representation is given in Table 3.7.

The dynamic threshold MOS transistor (DTMOS)¹³⁵ is the poor man's double-gate FET where the doped body, in the context of PD SOI, acts as a back, pn junction gate. Its applications is limited by the low turnon voltage of the pn junction. The back gate may also be driven by an additional transistor¹³⁶ to maintain a high input impedance. The individual FETs have to be very narrow in order to propagate high-speed signals down the highly resistive back gate.¹³⁷ Failure to do so leads to deleterious effects of increased delay and, more seriously, increased power dissipation due to the delayed turnoff of the back-gate. For instance, for a body resistance of 20 K Ω /sq (typical of PD SOI), a total capacitance per unit width of 2 fF/ μ m, and a delay of 1 ns, the maximum width per FET finger is 2.8 μ m. Overall this approach may be useful for low-power, low-voltage circuits where these restrictions do not have a significant effect.

3.8 SUMMARY AND FUTURE TRENDS

Ever since the concept of MOSFET device scaling has been introduced, there have been predictions and speculations of when MOSFET scaling will reach a limit.^{138,139} Some of the limits are practical concerns related to the ability to produce and control material properties and geometric parameters, while some of the limits are fundamental with roots in device physics. Most of these predictions turned out to be

TABLE 3.7 The Double-Gate FET Electrical Family

Style	Description	Advantages	Disadvantages
SYMM	Intragap gates equally spaced from thin undoped channel	<ul style="list-style-type: none"> “Ideal” double-gate structure Separate control of both gates High channel mobility No dopant fluctuations 	<ul style="list-style-type: none"> Inflexible V_T control Channel doping leads to severe V_T fluctuations Quantum effects limit minimum channel thickness Space needed for extra gate contact
n^+/p^+	High-low gate workfunction combination to simulate midgap gate workfunction	<ul style="list-style-type: none"> Can use CMOS standard n^+/p^+ polysilicon gates Better electron confinement relaxes silicon thickness requirements 	<ul style="list-style-type: none"> Lower mobility than symmetrical structure Reduced control of gate farthest from channel
Wrap	<ul style="list-style-type: none"> Gate wraps around silicon beam Narrow beam plus undoped channel Wide beam plus doped channel 	<ul style="list-style-type: none"> Very compact structure (wide beam) Even better gate control than “ideal” structure (narrow beam) Variable device width (wide beam) 	<ul style="list-style-type: none"> Crystal orientation varies around channel (narrow beam) Interface states and V_T smearing Source/drain fanout difficult to fabricate Small current carrying capacity per channel; inflexible V_T control; enhanced quantum effects (narrow beam) V_T very sensitive to beam width; difficult doping profile control; wider structures have poorer scaling properties (wide beam)
DTMOS	Connected body acts as a back-junction FET	<ul style="list-style-type: none"> Closest to standard SOI CMOS implementation Flexible V_T control 	<ul style="list-style-type: none"> P/N junction turnon limits gate voltage High resistance of body gate

too conservative, in retrospect, in estimating the industry’s ability to innovate. Perceived barriers at the 1-, 0.25-, and 0.1- μm technology generations have successively been surmounted. In the late 1990s, especially after the publication of the widely referenced SIA *National Technology Roadmap for Semiconductors*,¹ there has been much discussion about being near the limit of CMOS device scaling due to fundamental physics.^{11,85,119,131,140–143}

Meindl¹⁴⁰ succinctly characterized the limits of CMOS scaling into a hierarchy of limits due to (1) physics, (2) materials, (3) device, (4) circuit, and (5) system. Many papers were published on this subject in the late 1990s,^{10,85,118,119,140,142–145} and the key limitations from a device physics point of view are summarized in this section.

The critical dimensions that needs to be engineered are the gate length (L_g), the gate oxide thickness (t_{ox}), the depletion depths under the gate (x_{depl}), the source–drain junction depth (x_j), and steepness of the source–drain junction. All these quantities must be scaled together.

3.8.1 Gate Length

The gate length is the smallest feature of the MOSFET patterned by lithography and etching. Optical lithography has been able to provide generations of feature-size reduction mainly through the reduction of the wavelength of light employed.¹⁴⁶ Employing light with wavelength shorter than 193 nm (for gate lithography in the 2003 generation) presents many difficulties, among which the availability of materials for the optical system is a major barrier. Currently there is much debate about how the lithographic requirements will be met beyond the 2001 timeframe, with X ray, extreme UV, and electron beam all being considered. Sublithographic feature size may be obtained by etching techniques or sidewall image-transfer techniques. However, these largely experimental techniques have never been proven in a manufacturing environment. The development of a reliable, manufacturable cost-effective lithographic technique is absolutely essential to continued progress in CMOS technology.

3.8.2 Gate Oxide Thickness

As indicated in the Table 3.2, the electrical thickness of the gate insulator must decrease with the channel length. Recent (at the time of writing) studies of tunneling through thin oxides^{82,147} have shown that silicon dioxide can potentially be thinned to slightly below 2 nm before the leakage current and the associated dissipation become so large as to be unacceptable (Fig. 3.39). For equivalent (electrical) SiO₂ thicknesses below 2 nm, thicker gate insulators with a higher dielectric constant than silicon dioxide are being considered to reduce the tunneling current through the gate insulator.²² However, reliability and insulator–semiconductor interface properties remain the most important concerns for such new materials.

3.8.3 Depletion Depths and Junction Depths

Depletion depths and junction depths have been controlled by ion implantation of appropriate dopants into selected regions and by limiting their movement during subsequent heat cycles. Very sharp profiles of the order of < 5 nm/decade would be needed in the sub-50-nm regime. For bulk MOSFET, the substrate doping must be raised with scaling to shield the drain field from penetrating into the source, thereby causing substantial 2D short-channel effects. The sharp junction profile, together

with the high substrate doping, will eventually lead to Zener breakdown at the drain junction.⁸⁵ By changing the device structure to use a very thin undoped silicon channel such as the double-gate MOSFET,^{131,130} precise and heavy dopant placement in the channel region is no longer necessary. However, the need for a steep lateral junction dopant profile remains.¹⁴⁸ The thin undoped silicon channel also alleviates the discrete dopant fluctuation problem.¹⁴⁹

Confining the channel to a thin silicon introduces two potential problems: (1) quantization modulated threshold voltage and (2) source–drain resistance due to depletion of the source–drain. When the channel becomes very thin, V_T will vary because of quantum shifts of the ground-state energy inversely proportional to the square of the Si thickness. In silicon this shift is smallest for a (1,0,0) channel. Assuming a simple particle-in-a-box model,^{131,150} the uncertainty of the threshold voltage (σ_{V_T}) is $\sigma_{V_T} = -(\hbar^2 \pi^2 / qm^* t_{Si}^2)(C_{tSi} / t_{Si})$.

For a 4-nm-thick silicon channel with a 20% channel thickness control ($\sigma_{tSi} / t_{Si} = 0.2$), the σ_{V_T} is 50 mV, which is too high (see Table 3.2). This is illustrated in Figure 3.43.¹⁴⁸ The “heavily doped” source–drain can easily be depleted by the gate to source–drain bias. For a equivalent gate oxide of 1 nm, $V_G - V_T = -0.4$ V, $V_D = 1$ V, the depleted silicon depth is ≈ 3 nm for a source–drain doped to 10^{20} cm^{-3} . A double-gate MOSFET therefore must have a channel of at least 6 nm thick or a source–drain doping greater than $1 \times 10^{20} \text{ cm}^{-3}$ to avoid current degradation due to source–drain depletion.

3.8.4 Random Fluctuation of Device Properties

Random fluctuation of device properties may ultimately limit the number of devices that can be intergrated on one chip. Fluctuations of device properties result in

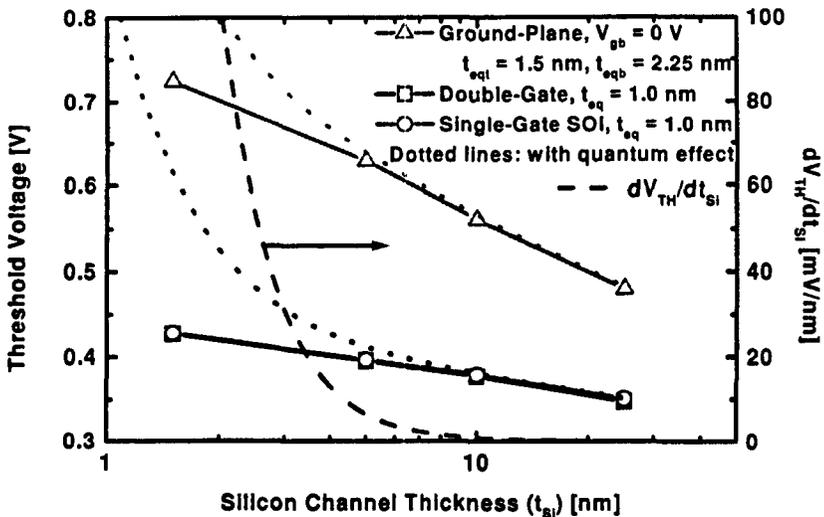


Figure 3.43 V_T shift in thin SOI due to quantum effect. (after Wong et al., Ref. 148.)

variations of the transistor current drive capabilities and propagation delays, leading to intolerable clock skews or malfunction of circuits that depends on matched or absolute values of device properties. The random fluctuation of the threshold voltage of a MOSFET due to the random fluctuation of the placement and number of dopant atoms is perhaps the most studied example associated with fundamental physics.^{138,149,151–153} Using a simple model of the threshold voltage fluctuation based on the fluctuation of the number of dopants under the gate, the standard deviation of the threshold voltage can be described by¹¹

$$\sigma_{V_T} = \left(\frac{q}{C_{\text{ox}}} \right) \sqrt{\frac{N_A W_{dm}}{3LW}} \left(1 - \frac{x_s}{W_{dm}} \right)^{3/2} \quad (144)$$

where C_{ox} is the gate oxide capacitance, L and W are the length and width of the MOSFET, W_{dm} is the maximum gate-induced depletion width, x_s is the width of the low-impurity (assumed undoped) region at the surface, and N_A is the doping concentration of the channel.⁸⁵ For a MOSFET with $W/L = 1$ and $L = 100$ nm, the standard deviation of the threshold voltage is 15 mV. Using this simple model, it is projected¹¹ that at least one transistor will have threshold voltages outside the specifications for proper circuit functioning [e.g., $\Delta V_T = 0.1(V_{DD} - V_T)$] after the year 2009 (0.07 μm) generation of the *National Technology Roadmap for Semiconductors*.¹

ACKNOWLEDGMENTS

The author would like to acknowledge the contributions of David J. Frank and Paul M. Solomon, who wrote a significant portion of the sections on device scaling, MOSFET device structures, and future trends in Ref. 11, which formed the basis of parts of this chapter.

REFERENCES

1. Semiconductor Industry Association (SIA), 4300 Stevens Creek Blvd., Suite 271, San Jose, CA 95129, *The National Technology Roadmap for Semiconductors* (NTRS), 1998.
2. R. Muller and T. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed., Wiley, New York, 1986.
3. S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1981.
4. P. Ko, "Approaches to Scaling," in N. Einspruch and G. Gildenblat, eds. *VLSI Electronics: Microstructure Science—Advances to MOS Device Physics*, Academic Press, 1989, Vol. 18, pp. 1–37.
5. R. Dennard et al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid State Circ.* **SC-9**, 256 (1974).
6. H. Pao and C. Sah, "Effects of Diffusion Current on Characteristics of Metal-Oxide (Insulator)-Semiconductor Transistors," *Solid-State Electron.* 927 (1966).

7. H.-S. Wong, M. White, T. Krutsick, and R. Booth, "Modeling of Transconductance Degradation and Extraction of Threshold Voltage in Thin Oxide MOSFET's," *Solid State Electron.* 953–968 (1987).
8. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design," *IEEE Trans. Electron Devices*, **ED-31** (4), 452 (1984).
9. F. Stern and W. E. Howard, "Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit," *Phys. Rev.* **163** (3) 816 (1967).
10. B. Davari, R. H. Dennard, and G. G. Snahidi, "CMOS Scaling, the Next Ten Years," *IEEE Proc.*, 595 (1995).
11. H.-S. P. Wong, D. Frank, P. M. Solomon, H.-J. Wann, and J. Welser, "Nanoscale CMOS," *IEEE Proc.* 537 (April 1999).
12. G. Moore, "Progress in Digital Integrated Electronics," *Int. Electron Devices Meeting*, 1975, pp. 11–13.
13. W.-H. Chang, B. Davari, M. R. Wordeman, Y. Taur, C. C.-H. Hsu, and M. D. Rodriguez, "A High-Performance 0.25- μm CMOS Technology: I—Design and Characterization," *IEEE Trans. Electron Devices* **39**, 959 (April 1992).
14. L. Su et al., "A High Performance 0.08 μm CMOS," *Symp. VLSI Technol.* 12 (1996).
15. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFET's," *Solid State Electron.* **17**, 1059 (1974).
16. R. Troutman, "VLSI Limitations Due to Two-Dimensional Field Effect," *IEEE Trans. Electron Devices*, 980 (1979).
17. Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. Ko, and Y. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFETs," *IEEE Trans. Electron Devices*, 86 (1993).
18. R. Yan, A. Ourmazd, and K. Lee, "Scaling the Si MOSFET: from Bulk to SOI to Bulk," *IEEE Trans. Electron Devices*, 1704 (1992).
19. K. Young, "Analysis of Conduction in Fully Depleted SOI MOSFET's," *IEEE Trans. Electron Devices*, 504 (1989).
20. J. Woo, K. W. Terrill, and P. Vasudev, "Two-Dimensional Analysis Modeling of Very Thin SOI MOSFET's," *IEEE Trans. Electron Devices*, 1999 (1990).
21. K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling Theory for Double-Gate SOI MOSFET's," *IEEE Trans. Electron Devices* **40**(12), 2326 (1993).
22. D. Frank, Y. Taur, and H.-S. P. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFET's," *IEEE Electron Device Lett.* 385 (1998).
23. Y. Cheng, "On the Scattering of Electrons in Magnetic and Electric Surface States by Surface Roughness," *Surf. Sci.* 663 (1971).
24. Y. Cheng, "Electron Mobility in an MOS Inversion Layer," *Conf. Solid-State Devices and Materials*, 1971, p. 173.
25. C. Sah, T. Ning, and L. Tschopp, *Surf. Sci.* 561 (1972).
26. Y. Cheng and E. Sullivan, "Scattering of Charge Carriers in Silicon Surface Layers," *J. Appl. Phys.* 923 (1973).
27. Y. Cheng, "Effect of Charge Inhomogeneities on Silicon Surface Mobility," *J. Appl. Phys.* 2425 (1973).

28. Y. Cheng and E. Sullivan, "Relative Importance of Phonon Scattering to Carrier Mobility in Si Surface Layer at Room Temperature," *J. Appl. Phys.* 3619–3625 (1973).
29. Y. Cheng and E. Sullivan, "On the Role of Scattering by Surface Roughness in Silicon Inversion Layer," *Surf. Sci.* 717 (1973).
30. Y. Matsumoto and Y. Uemura, "Scattering Mechanism and Low Temperature Mobility of MOS Inversion Layers," *Jpn. J. Appl. Phys.* (1974).
31. N. Aschroft and N. Mermin, *Solid State Physics*. Holt, Rienhart and Winston, 1976.
32. A. Sabnis and J. Clemens, "Characterization of the Electron Mobility in the Inverted (100) Si Surface," *Int. Electron Devices Meeting*, 1979, p. 18.
33. S. Sun and J. Plummer, "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces," *IEEE Trans. Electron Devices*, 1497 (1980).
34. J. Watt and J. Plummer, "Universal Mobility-Field Curves for Electrons and Holes in MOS Inversion Layers," *Symp. VLSI Technol.* 81 (1987).
35. S. Takagi, I. Iwase, and A. Toriumi, "On the Universality of Inversion-layer Mobility in n- and p-Channel MOSFET's," *Int. Electron Devices Meeting*, 1988, p. 398.
36. C. Sodini, T. Ekstedt, and J. Moll, "Charge Accumulation and Mobility in Thin Dielectric MOS Transistors," *Solid State Electron.* 831 (1982).
37. R. Booth, M. White, H.-S. Wong, and T. Krutsick, "The Effect of Channel Implant on MOS Transistor Characterization," *IEEE Trans. Electron Devices* 2501–2509 (1987).
38. T. Krutsick and M. White, "Considerations of Doping Profiles in MOS-FET Mobility Modeling," *IEEE Trans. Electron Devices* 1153 (1988).
39. C.-L. Huang and G. Goldenblat, "Measurements and Modeling of the n-Channel MOSFET Inversion Layer Mobility and Device Characteristics in the Temperature Range 60–300 K," *IEEE Trans. Electron Devices* 1289–1300, (1990).
40. H.-S. Wong " 'Universal' Effective Mobility of Empirical Local Mobility Models for n- and p-Channel Silicon MOSFET's," *Solid State Electron.* 179–188 (1993).
41. G. Bacarani and M. Wordeman, "Transconductance Degradation in Thin-Oxide MOSFET's," *IEEE Trans. Electron Devices* 1295–1304 (1983).
42. S. Takagi, M. Iwase, and A. Toriumi, "Effects of Surface Orientation on the Universality of Inversion Layer Mobility in Si MOSFETs," *Conf. Solid-State Devices and Materials*, 1990.
43. M. Sherony, L. Su, J. Chung, and D. Antoniadis, "SOI MOSFET Effective Channel Mobility," *IEEE Electron Device Lett.* 276–278 (1994).
44. E. Buturla, J. Johnson, S. Furkay, and P. Cottrell, "A New 3-D Device Simulation Formulation," in *NASCODE VI: 6th Int. Conf. Numerical Analysis of Semiconductor Devices and Integrated Circuits* (Dublin), Boole Press, 1989, pp. 291–296.
45. S. Selberherr, W. Hansch, M. Seavey, and J. Slotboom, "The Evolution of the MINIMOS Mobility Model," *Archiv fur Elektronik und Ubertragungstechnik*, 161 (1990).
46. C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, "A Physically Based Mobility Model for Numerical Simulation of Nonplanar Devices," *IEEE Trans. Computer-Aided Design* 1164 (1988).
47. S. Schwarz and S. Russek, "Semi-Empirical Equations for Electron Velocity in Silicon; Part I—Bulk," *IEEE Trans. Electron Devices* 1629 (1983).

48. S. Schwarz and S. Russek, "Semi-Empirical Equations for Electron Velocity in Silicon; Part II—MOS Inversion Layer," *IEEE Trans. Electron Devices*, 1634 (1983).
49. K. Thornber, "Relation of Drift Velocity to Low-Field Mobility and High-Field Saturation Velocity," *J. Appl. Phys.* 21:27 (1980).
50. C. Sodini, P. Ko, and J. Moll, "The Effect of High Fields on MOS Device and Circuit Performance," *IEEE Trans. Electron Devices* 1386 (1984).
51. T. Vogelsang and W. Hansch, "The Electron High-Energy Distribution Function: A Comparison of Analytical Models with Monte Carlo Calculations," *J. Appl. Phys.* **69**, (6), 3592 (1991).
52. T. Ning, C. Osburn, and H. Yu, "Emission Probability of Hot Electrons from Silicon into Silicon Dioxide," *J. Appl. Phys.* **48**(1), 286 (1977).
53. C. Hu, S. Tam, F. Hsu, P. Ko, T. Chan, and K. Terrill, "Hot-Electron-Induced MOSFET Degradation—Model, Monitor, and Improvement," *IEEE Trans. Electron Devices* **ED-32**, (2), 375 (1985).
54. A. Toriumi, "Experimental Study of Hot Carriers in Small Size Si-MOSFETs," *Solid State Electron.* **32**, (12), 1519–1525 (1989).
55. M. Fischetti, S. Laux, and E. Crabbé, "Understanding Hot-Electron Transport in Silicon Devices: Is There a Short Cut?" *J. Appl. Phys.* 1058 (1995).
56. M. Fischetti and S. Laux, "Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices Including Band-Structure and Space-Charge Effects," *Phys. Rev. B* **B38**, 9721 (1988).
57. N. Sano, M. Tomizawa, and A. Yoshii, "Monte Carlo Analysis of Ionization Threshold in Silicon," *Appl. Phys. Lett.* 653 (1990).
58. N. Sano, M. Tomizawa, and A. Yoshii, "Electron Transport and Impact Ionization in Si," *Phys. Rev. B* 12122 (1990).
59. S. Tam, P. Ko, and C. Hu, "Lucky-Electron Model of Channel Hot-Electron Injection in MOSFET's," *IEEE Trans. Electron Devices* **ED-31** (9), 1116 (1984).
60. M.-S. Liang, C. Chang, W. Yang, C. Hu, and R. W. Brodersen, "Hot Carriers Induced Degradation in Thin Gate Oxide MOSFETs," *Int. Electron. Devices Meeting*, 1983, p. 186.
61. W. Weber, C. Werner, and G. Dorda, *IEEE Electron. Device Lett.* 518 (1984).
62. H.-S. Wong, "Experimental Verification of the Mechanism of Hot-Carrier-Induced Photon Emission in n-MOSFET's with an Overlapping CCD Gate Structure," *IEEE Electron Device Lett.* **EDL-13**, 389 (Aug. 1992).
63. J. Bude, N. Sano, and A. Yoshii, "Hot-Carrier Luminescence in Si," *Phys. Rev. B* 5848 (1992).
64. L. Selmi, H.-S. Wong, E. Sangiorgi, M. Lanzoni, and M. Manfredi, "Investigation of Hot Electron Luminescence in Silicon by Means of Dual-Gate MOSFET's," *Int. Electron Devices Meeting*, 1993, p. 531.
65. S. Tam and C. Hu, "Hot-Electron-Induced Photon and Photocarrier Generation in Silicon MOSFET's," *IEEE Trans. Electron Devices* **ED-31**(9), 1264 (1984).
66. P. Childs, R. Stuart, and W. Eccleston, "Evidence of Optical Generation of Minority Carriers from Saturated MOS Transistors," *Solid State Electron.* **27**(7), 685 (1983).

67. A. Toriumi, M. Yoshimi, M. Iwase, Y. Akiyama, and K. Taniguchi, "A Study of Photon Emission from n-Channel MOSFET's," *IEEE Trans. Electron Devices* **ED-34** (7), 1501 (1987).
68. T. Tsuchiya and S. Nakajima, "Emission Mechanism and Bias-Dependent Emission Efficiency of Photon Induced by Drain Avalanche in Si MOSFET's," *IEEE Trans. Electron Devices* **ED-32** (2), 405 (1985).
69. W. Shockley, "Problems Related to p-n Junctions in Silicon," *Solid State Electron.* 35-67 (1961).
70. J. Verwey, R. Kramer, and B. de Maagt, "Mean Free Path of Hot Electrons at the Surface of Boron-Doped Silicon," *J. Appl. Phys.* **46** (6), 2612 (1975).
71. H.-S. Wong, "Gate Current Injection in MOSFET's with a Split-Gate (Virtual Drain) Structure," *IEEE Electron Device Lett.* 262 (1993).
72. M. Fischetti and S. Laux, "Monte Carlo Study of Sub-Band-Gap Impact Ionization in Small Silicon Field Effect Transistors," *Int. Electron Devices Meeting*, 1995, p. 305.
73. A. Abramo, C. Fiegna, and F. Venturi, "Hot Carrier Effects in Short MOSFETs at Low Applied Voltages," *Int. Electron Devices Meeting*, 1995, p. 301.
74. S. Ogura, P. Tsang, W. Walker, D. Critchlow, and J. Shepard, "Elimination of Hot Electron Gate Current by Lightly Doped Drain-Source Structure," *Int. Electron Devices Meeting*, 1981, p. 651.
75. K. Mayaram, K. Lee, and C. Hu, "A Model for the Electric Field in Lightly Doped Drain Structure," *IEEE Trans. Electron Devices* 1509 (1987).
76. J. Chen, T. Chan, I. Chen, P. Ko, and C. Hu, "Subbreakdown Drain Leakage Current in MOSFET," *IEEE Electron Device Lett.* 515 (1987).
77. T. Chan, J. Chen, P. Ko, and C. Hu, "The Impact of Gate-Induced Drain Leakage Current on MOSFET Scaling," *Int. Electron Devices Meeting*, 1987, p. 718.
78. C. Chang and J. Lien, "Corner-Field Induced Drain Leakage in Thin Oxide MOSFETs," *Int. Electron Devices Meeting*, 1987, p. 714.
79. C. Chang, S. Haddad, B. Swaminathan, and J. Lien, "Drain-Avalanche and Hole-Trapping Induced Gate Leakage in Thin-Oxide MOS Devices," *IEEE Electron Device Lett.* 588 (1988).
80. I. Chen, C. Kaya, and J. Paterson, "Band-to-Band Tunneling Induced Substrate Hot-Electron (BBISHE) Injection: A New Programming Mechanism for Nonvolatile Memory Devices," *Int. Electron Devices Meeting*, 1989, p. 263.
81. I. Chen and C. Teng, "A Quantitative Physical Model for the Band-to-Band Tunneling-Induced Substrate Hot Electron Injection in MOS Devices," *IEEE Trans. Electron Devices*, 1646 (1992).
82. S.-H. Lo, D. Buchanan, and Y. Taur, "Modeling and Characterization of n⁺- and p⁺-Polysilicon-Gated Ultra Thin Oxides (21-26A)," *Symp. VLSI Technol.* 149-150 (1997).
83. C. Wong et al., "Doping of n⁺ and p⁺ Polysilicon in a Dual-Gate CMOS Process," *Int. Electron Devices Meeting*, 1988, p. 238.
84. B. Davari, W.-H. Chang, K. E. Petrillo, C. Y. Wong, D. Moy, Y. Taur, M. R. Wordeman, J. Y. C. Sun, and C. C.-H. Hsu, "A High-Performance 0.25- μ m CMOS Technology: II—Technology," *IEEE Trans. Electron Devices* **39**, 967 (April 1992).

85. Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S.-H. Lo, G. Sai-Halasz, R. Viswanathan, H.-J. C. Wann, S. Wind, and H.-S. Wong, "CMOS Scaling into the Nanometer Regime," *IEEE Proc.* 1997, p. 486.
86. J. Huang, Z. Liu, M. Jeng, P. Ko, and C. Hu, "A Physical Model for MOSFET Output Resistance," *Int. Electron Devices Meeting*, 1992, p. 569.
87. H.-S. Chen, C. Teng, J. Zhao, L. Moberly, and R. Lahri, "Analog Characteristics of Drain Engineered Submicron MOSFET's for Mixed-Signal Applications," *Solid State Electron.* 1857 (1995).
88. J. Chung, K. N. Quader, C. Sodini, P. Ko, and C. Hu, "The Effects of Hot-Electron Degradation on Analog MOSFET Performance," *Int. Electron Devices Meeting*, 1990, p. 553.
89. L. Su, J. Yasaitis, and D. Antoniadis, "A High-Performance Scalable Submicron MOSFET for Mixed Analog/Digital Applications," *Int. Electron Devices Meeting*, 1991, p. 367.
90. C. Sodini, S. Wong, and P. Ko, "A Framework to Evaluate Technology and Device Design Enhancements for MOS Integrated Circuits," *IEEE J. Solid State Circ.* 24(1), 118 (1989).
91. C. Sah, "Evolution of the MOS transistor—from Conception to VLSI," *IEEE Proc.* 1280–1326 (1988).
92. R. B. Fair, "History of Some Early Developments in Ion-Implantation Technology Leading to Silicon Transistor Manufacturing," *Proc. IEEE* 86, 111 (1998).
93. R. Dennard, "Field Effect Transistor Memory," U.S. Patent 3,387,286, filed July 14, 1968.
94. P. Bondy, "Moore's Law Governs the Silicon Revolution," *Proc. IEEE* 86, 78 (1998).
95. B. Crowder and S. Zirinsky, "Metal Silicide Interconnection Technology—a Future Perspective," *IEEE Trans. Electron Devices*, 26, 369 (1979).
96. C. Lau, Y. See, D. Scott, J. Bridges, S. Perna, and R. Davies, "Titanium Disilicide Self-Aligned Source/Drain + Gate Technology," *Int. Electron Devices Meeting*, 1982, p. 714.
97. R. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *Int. Electron Devices Meeting*, 1982, p. 237.
98. S. Wong, C. Sodini, T. Eckstedt, H. Grinolds, K. Jackson, and S. Kwan, "Low Pressure Nitrided Oxide as a Thin Gate Dielectric for MOSFETs," *J. Electrochem. Soc.* 130, 1139 (1983).
99. C. Codella and S. Ogura, "Halo Doping Effects in Submicron LDD Device Design," *Int. Electron Devices Meeting*, 1985, p. 230.
100. J. - C. Sun, Y. Taur, R. Dennard, S. Klepner, and L. Wang, "0.5 μ m-Channel CMOS Technology Optimized for Liquid-Nitrogen-Temperature Operation," *Int. Electron Devices Meeting*, 1986, p. 236.
101. S. Hillenius, R. Liu, G. Georgiou, D. W. R. L. Field, A. Kornblit, D. Boulin, R. Johnston, and W. Lynch, "A Symmetric Submicron CMOS Technology," *Int. Electron Devices Meeting*, 1986, p. 252.
102. G. Sai-Halasz et al., "Experimental Technology and Characterization of Self-Aligned 0.1 μ m Gate-Length Low-Temperature Operation NMOS Devices," *Int. Electron Devices Meeting*, 1987, p. 397.

103. B. Davari, C. Koburger, R. Schulz, J.D. Warnock, Y. Taur, W. Schwittek, J. K. DeBrosse, M. L. Kerbaugh, and J. Mauer, "A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polish (CMP)," *Int. Electron Devices Meeting*, 1989, p. 61.
104. F. White, W. Hill, S. Eslinger, E. Payne, W. Cote, B. Chen, and K. Johnson, "Damascene Stud Local Interconnect in CMOS Technology," *Int. Electron Devices Meeting*, 1992, p. 301.
105. J. Paraszczak, D. Edelstein, S. Cohen, E. Babich, and J. Hummel, "High Performance Dielectrics and Processes for ULSI Interconnection Technologies," *Int. Electron Devices Meeting*, 1993, p. 261.
106. B. Davari, C. Koburger, T. Furukawa, Y. Taur, W. Noble, A. Megdanis, J. Warnock, and J. Mauer, "A Variable-Size Shallow Trench Isolation (STI) Technology with Diffused Sidewall Doping for Submicron CMOS," *Int. Electron Devices Meeting*, 1988, p. 92.
107. Y. Taur et al., "High Transconductance 0.1 μm pMOSFET," *Int. Electron Devices Meeting*, 1992, pp. 901–904.
108. Y. Taur et al., "High Performance 0.1 μm CMOS Devices with 1.5 V Power Supply," *Int. Electron Devices Meeting*, 1993, pp. 127–130.
109. G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, and E. Ganin, "High Transconductance and Velocity Overshoot in NMOS Devices at the 0.1 μm -Gate-Length Level," *IEEE Electron Device Lett.* **EDL-9**, 464 (1988).
110. R. Yan et al., "High-Performance 0.1 μm Room Temperature Si MOSFETs," *Symp. VLSI Technol.* 86 (1992).
111. Y. Mii et al., "High Performance 0.1 μm nMOSFET's with 10 ps/Stage Delay (85 K) at 1.5 V Power Supply," *Symp. VLSI Technol.* 91–92 (1993).
112. M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, and H. Iwai, "Sub-50 nm Gate Length n-MOSFETs with 10 nm Phosphorus Source and Drain Junctions," *Int. Electron Devices Meeting*, 1993, p. 119.
113. Y. Mii et al., "An Ultra-Low Power 0.1 μm CMOS," *Symp. VLSI Technol.* 9–10 (1994).
114. H. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "Tunneling Gate Oxide Approach to Ultra-High Current Drive in Small-Geometry MOSFET's," *Int. Electron Devices Meeting*, 1994, p. 593.
115. C. Wann et al., "High-Performance 0.07 μm CMOS with 9.5 ps Gate Delay and 150 GHz f_T ," *IEEE Electron Device Lett.* **18**, 625 (Dec. 1997).
116. F. Assaderaghi et al., "A 7.9/5.5 psec Room/Low Temperature SOI CMOS," *Int. Electron Devices Meeting*, 1997, p. 415.
117. G. Timp et al., "Low Leakage, Ultra-Thin, Gate Oxides for Extremely High Performance Sub-100 nm nMOSFETs," *Int. Electron Devices Meeting*, 1997, p. 930.
118. S. Asai and Y. Wada, "Technology Challenges for Integration Near and Below 0.1 μm ," *IEEE Proc.* 505 (1997).
119. M. T. Bohr, "Interconnect Scaling—the Real Limiter to High Performance ULSI," *Int. Electron Devices Meeting*, 1995, p. 241.
120. S. Sun, "Process Technologies for Advanced Metallization and Interconnect Systems," *Int. Electron Devices Meeting*, 1997, p. 765.

121. D. Edelstein et al., "Full Copper Wiring in a Sub-0.25 μm CMOS ULSI Technology," *Int. Electron Devices Meeting*, 1997, p. 773.
122. Zielinski et al., "Damascene Integration of Copper and Ultra-Low-k Xerogel for High Performance Interconnects," *Int. Electron Devices Meeting*, 1997, p. 936.
123. M. Matsuura, I. Tottori, K. Goto, K. Maekawa, and M. Hirayama, "A Highly Reliable Self-planarizing Low-k Intermetal Dielectric for Subquarter Micron Interconnects," *Int. Electron Devices Meeting*, 1997, p. 785.
124. G. Sai-Halasz, "Performance Trends in High-End Processors," *IEEE Proc.* **83**, 20 (1995).
125. G. Shahidi, C. Anderson, B. Chappel, T. Chappel, J. Comfort, B. Davari, R. Dennard, R. Franch, P. McFarland, J. Neely, T. Ning, M. Polcari, and J. Warnock, "A Room Temperature 0.1 μm CMOS on SOI," *IEEE Trans. Electron Devices* **12**, 2405 (1994).
126. I. Yang, C. Vieri, A. Chandrakasan, and D. Antoniadis, "Back-Gated CMOS on SOI for Dynamic Threshold Voltage Control," *IEEE Trans. Electron Devices* **44**, 822 (1997).
127. M. Horiuchi, T. Teshima, K. Tokumasu, and K. Yamaguchi, "High-Current, Small Parasitic Capacitance MOSFET on a Poly-Si Interlayered (PSI) SOI Wafer," *Symp. VLSI Technol.* 33 (1995).
128. T. Kachi, T. Kaga, S. Wakahara, and D. Hisamoto, "Variable Threshold-Voltage SOI CMOSFETs with Implanted Back-Gate Electrodes for Power-Managed Low-Power and High-Speed Sub-1- ν ULSIs," *Symp. VLSI Technol.* 124 (1996).
129. C. Wann, R. Tu, B. Yu, C. Hu, K. Noda, T. Tanaka, M. Yoshida, and K. Hui, "A Comparative Study of Advanced MOSFET Structures," *Symp. VLSI Technol.* 32 (1996).
130. C. Fiegna, H. Iwai, T. Wada, T. Saito, E. Sangiorgi, and B. Ricco, "A New Scaling Methodology for the 0.1–0.025 μm MOSFET," *Symp. VLSI Technol.* 33 (1992).
131. D. Frank, S. Laux, and M. Fischetti, "Monte Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Far Can Si Go?" *Int. Electron Devices Meeting*, 1992, p. 553.
132. H.-S. Wong, D. Frank, Y. Taur, and J. Stork, "Design and Performance Considerations for Sub-0.1 μm Double-Gate SOI MOSFET's," *Int. Electron Devices Meeting*, 1994, p. 747.
133. H.-S. Wong, K. Chan, and Y. Taur, "Self-Aligned (Top and Bottom) Double-Gate MOSFET with a 25 nm Thick Silicon Channel," *Int. Electron Devices Meeting*, 1997, p. 427.
134. E. Nowak, J. Johnson, D. Hoyniak, and J. Thygesen, "Fundamental MOSFET Short-Channel- V_t /Saturation Current/Body Effect Trade-off," *Int. Electron Devices Meeting*, 1994.
135. F. Assaderaghi, S. Park, D. Sinitzky, J. Bokor, P.-K. Ko, and C. Hu, "A Dynamic Threshold Voltage MOSFET (DTMOS) for Low Voltage Operation," *IEEE Electron Device Lett.* **15**, 510 (1994).
136. I.-Y. Chung, Y.-J. Park, and H.-S. Min, "A New SOI Inverter Using Dynamic Threshold for Low-Power Applications," *IEEE Electron Device Lett.* **18**, 248 (1997).
137. C. Wann, F. Assaderaghi, R. Dennard, C. Hu, G. Shahidi, and Y. Taur, "Channel Profile Optimization and Device Design for Low-Power High-Performance Dynamic-Threshold MOSFET," *Int. Electron Devices Meeting*, 1996, p. 113.

138. B. Hoeneisen and C. A. Mead, "Fundamental Limitations in Microelectronics—I MOS Technology," *Solid State Electron.* **15**, 819 (1972).
139. R. H. Dennard, "Technology Challenges for Ultrasmall Silicon MOSFET's," *J. Vac. Sci. Technol.* **19**, (3) 537 (1981).
140. J. Meindl, "Low Power Microelectronics—Retrospect and Prospect," *IEEE Proc.* **83**, 619 (April 1995).
141. J. Stork, *IEEE Proc.* 607 (1995).
142. C. Mead, "Scaling of MOS Technology to Submicrometer Feature Sizes," *J. VLSI Signal Process.* 9–25 (1994).
143. Y. Taur and E. Nowak, "CMOS Devices Below 0.1 μm : How High Will Performance Go?" *Int. Electron Devices Meeting*, 1997, p. 215.
144. J. Meindl et al., *Int. Solid State Circuits Conf.* 124 (1993).
145. J. Meindl, V. De, D. Wills, J. Eble, X. Tang, J. Davis, B. Austin, and A. Bhavnagarwala, "The Impact of Stochastic Dopant and Interconnect Distributions on Gigascale Integration," *Int. Solid State Circ. Conf.* 232 (1997).
146. T. Brunner, "Pushing the Limits of Lithography for IC Production," *Int. Electron Devices Meeting*, 1997, p. 9.
147. F. Rana, S. Tiwari, and D. Buchanan, "Self-Consistent Modeling of Accumulation Layers and Tunneling Currents through Very Thin Oxides," *Appl. Phys. Lett.* **69**(8), 1104 (1996).
148. H.-S. Wong, D. Frank, and P. Solomon, "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFET's at the 25 nm Channel Length Generation," *Int. Electron Devices Meeting*, 1998, p. 407.
149. H.-S. Wong and Y. Taur, "Three-Dimensional 'Atomistic' Simulation of Discrete Microscopic Random Dopant Distributions Effects in Sub-0.1 μm MOSFET's," *Int. Electron Devices Meeting*, 1993, pp. 705–708.
150. C. Kittel, *Introduction to Solid State Physics*, Wiley, New York, 1956, Chapter 11, p. 283.
151. R. W. Keyes, "Effect of Randomness in the Distribution of Impurity Ions on FET Thresholds in Integrated Electronics," *IEEE J. Solid State Circ.* 245 (1975).
152. T. Mizuno, J. Okamura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuations Using on 8 k MOSFET's Array," *VLSI Symp.* 41 (1993).
153. V. De, X. Tang, and J. Meindl, "Random MOSFET Parameter Fluctuation Limits to Gigascale Integration (GSI)," *VLSI Symp.* 198 (1996).
154. L. T. Su, H. Hu, J. B. Jacobs, M. Sherony, A. Wei, and D. A. Antoniadis, "Tradeoffs of Current Drive vs. Short-Channel Effect in Deep-Submicrometer Bulk and SOI MOSFET's," *Int. Electron Devices Meeting*, 1994, p. 649.

PROBLEMS

- 3.1 Assume that in the year 1999, the gate length L is 0.25 μm , gate oxide thickness is 5 nm, substrate doping density is $2 \times 10^{17} \text{ cm}^{-3}$, and the scaling factor α is 0.7 every 2 years. What would the gate length, gate oxide thickness, and substrate doping density be by the year 2013 according to the scaling rule?

- 3.2 If the scaling factor is α/ϵ , how does the effective field E_{eff} scale?
- 3.3 A n-MOSFET has an intrinsic transconductance of 3000 mS/mm. The extrinsic series resistance of the source–drain is $300 \Omega \cdot \mu\text{m}$. Calculate the measured extrinsic transconductance of this MOSFET.
- 3.4 Solve the Poisson equation for a cylindrical MOSFET where the silicon substrate is a cylinder and the gate material and the gate oxide surrounds the cylindrical silicon substrate.
- 3.5 Calculate the natural scaling length of a cylindrical MOSFET where the silicon substrate is a cylinder and the gate material and the gate oxide surround the cylindrical silicon substrate.
- 3.6 Calculate the relaxation time between scattering for a carrier mobility of $250 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$.
- 3.7 Assume that the polysilicon gate is doped to a doping density of $1 \times 10^{20} \text{ cm}^{-3}$. What is the gate capacitance at a bias of 1.2 V for a gate oxide of 2.5 nm and a substrate doping of $1 \times 10^{18} \text{ cm}^{-3}$?
- 3.8 Calculate the shift in threshold voltage for a silicon channel of 3 nm using a simple particle-in-a-box assumption.
- 3.9 Calculate the uncertainty of threshold voltage for a 10% variation of the thickness of the silicon channel for a double-gate MOSFET with a 1.5 nm thickness. Use a simple particle-in-a-box assumption.
- 3.10 Calculate the threshold voltage fluctuation due to discrete random dopants for a properly scaled MOSFET in the year 2003 assuming “uniform doping” and compare this with the NTRS-allowed threshold voltage tolerance.

Device Miniaturization and Simulation

SANJAY BANERJEE and BEN STREETMAN

The University of Texas at Austin
Austin, TX

4.1 INTRODUCTION

In this chapter, we review the various modeling and simulation techniques used for semiconductor devices, with primary emphasis on those that are useful for ULSI (MOSFETs and BJTs).^{1,2} We examine these simulation methods at different levels of complexity and accuracy. First we discuss a fundamentally quantum mechanical picture of what is a complicated, many-body electron transport problem. We shall see what the various approximations are that enable one to reduce this problem to a single-electron Schrödinger equation, with an “effective” mass to quantum-mechanically account for the complicated interactions of the electrons with the periodic lattice potential. These electrons with an effective mass, different from the true mass, will be shown to respond to the external potential according to Newtonian dynamics, except for occasional scattering mechanisms. This forms the base of what is called *semiclassical electron dynamics*.³ The carrier concentrations are generally determined on the basis of a parabolic density of states and the quantum-statistical Fermi–Dirac (FD), or classical Maxwell–Boltzmann (MB) distribution function. One can then simulate devices in the semiclassical regime using a single-particle picture with statistical Monte Carlo (MC) techniques. Alternatively, one can form a collective picture of carrier transport based on the evolution of the distribution function according to semiclassical dynamics, which is the Boltzmann transport equation (BTE). Next we will show, with examples drawn from MOSFETs and BJTs, how this treatment simplifies to the conventional device simulation formalism under certain conditions.^{1,2} Conventional modeling is based on the self-consistent solution of the Poisson equation with the carrier (electron and hole) continuity equations and the drift diffusion equations for current transport. What will become clear from that discussion is how the approximations implicit in conventional device

modeling break down under certain conditions of device miniaturization and high fields. We will conclude this chapter by briefly examining how even the more fundamental semi-classical dynamics picture and the BTE equation break down on further device scaling, which puts us in the realm of quantum transport.

4.2 BAND STRUCTURE

4.2.1 Single-Electron Effective-Mass Equation

The motion of electrons in a semiconductor crystal is clearly a complicated, many-body problem where the electrons are subjected to a variety of forces. The exact solution involves solving the Schrödinger equation for the entire crystal. As we recall from quantum mechanics, this requires determining the Hamiltonian, or the sum of the kinetic-energy and potential-energy operators, for all the ions and electrons in the semiconductor device.³ This is written below, followed by an explanation of the different terms and the various approximations made to convert this complicated equation to a more tractable, single-electron, effective-mass equation.

$$\left\{ -\frac{\hbar^2}{2M} \sum_{i=1}^{N_{\text{Avo}}} \nabla_i^2 + \sum_R \frac{(Ze)^2}{|R|} - \frac{\hbar^2}{2m} \sum_{i=1}^{N_{\text{Avo}}} \nabla_i^2 - \sum_{i=1}^{N_{\text{Avo}}} \sum_R \frac{Ze^2}{|r_i - R|} + \frac{1}{2} \sum_{ij} \frac{e^2}{r_{ij}} - eV(\vec{r}, t) - eV_{e,\alpha}(\vec{r}, t) \right\} \Psi = E\Psi \quad (4.1)$$

Here, the first two terms refer, respectively, to the kinetic-energy operator and the potential-energy operator of the ions in the crystal with a mass M , a charge Ze , and a periodic spacing of R corresponding to the relevant Bravais lattice [Face-centered cubic (FCC) lattice in the case of Si]. The third term refers to the kinetic energy of the electrons with a mass m . The remaining terms pertain to the various potential-energy components for the electrons, which, respectively, involve the interactions of the electrons with the periodic lattice potential, the electron–electron repulsive potential (without self-interactions), and finally the sum of random scattering potentials, $V(\vec{r}, t)$, as well as the externally applied potential. The various terms are summed over all the ions, and all the valence electrons in the system (indices i and j), which roughly corresponds to an Avogadro number, N_{Avo} , of terms. Here, E is the total energy of the system and

$$\Psi(\vec{r}) = \Psi(r_1 s_1, \dots, r_N s_N) \quad (4.2a)$$

is the multielectron wavefunction for an N -electron system in terms of the spatial coordinates r and the spin coordinates s . Since electrons are fermions subject to the Pauli exclusion principle, the wavefunction has to be suitably antisymmetrized by constructing a Slater determinant in terms of the individual electron

wavefunctions:

$$N\text{-fermion wavefunction} = \begin{vmatrix} \psi_1(r_1 s_1) & \cdots & \psi_1(r_N s_N) \\ \vdots & & \vdots \\ \psi_N(r_1 s_1) & & \psi_N(r_N s_N) \end{vmatrix} \quad (4.2b)$$

We next discuss the various simplifications of the preceding multielectron Schrödinger equation which leads to the single-electron effective-mass equation and forms the basis of most transport models in semiconductor devices.^{4,5} The most simplistic assumption is that the electrons move in straight-line trajectories in the crystal in a *perfect* array of *fixed* ions (the static ion approximation), where the ions have no role except for occasional collisions with electrons, which changes their trajectories drastically. This is the basis of the Drude free-electron model.³ The electrons collide between t and $t + dt$ with a probability $= dt/\tau$, where τ is the relaxation time, which in the Drude picture was erroneously attributed to the collisions of the electrons with the static, periodic array of ions. This is the *relaxation-time approximation* (RTA). The static ion approximation is clearly not valid because lattice ions execute simple harmonic oscillations about their equilibrium positions.

In spite of these lattice vibrations or phonons, however, one can ignore the kinetic-energy term of the nuclei in the Hamiltonian. Since the electrons are much lighter than the ions, they move much faster. Hence, using the adiabatic or Born–Oppenheimer approximation, the electrons reach the ground-state energy corresponding to the instantaneous location of the ions. As such, the ion location or spacing, R , becomes a parameter and the total electron energy, E , can be expressed as a function of the lattice spacing as $E(R)$. It may be noted that one cannot a priori assume that the kinetic energy of the ions is negligible compared to the kinetic energy of the electrons because, although the ions travel much slower, they are a lot more massive than the electrons. We can only drop the kinetic-energy term of the ions in the spirit of the Born–Oppenheimer approximation.

A much more glaring deficiency of the Drude free-electron model than the static ion approximation, is its failure to correctly (quantum-mechanically) account for the periodic lattice potential. Since the periodicity of the lattice potential is on a shorter distance scale than the De Broglie wavelength of electrons (about 12 nm in Si), one absolutely cannot treat the electrons as Newtonian point objects. The wave nature of electrons is fully manifested in terms of the response to the periodic lattice potential, and requires the solution of the Schrödinger equation. This is a very important point philosophically because it implies that there is *no* scattering of the electron wave propagation by the perfectly periodic array of ions in a lattice, unlike in the Drude picture. Any scattering is due to the *random* component of the potential, $V(\vec{r}, t)$. We shall see later how this scattering is handled quantum mechanically using time-dependent perturbation theory and the Fermi “golden rule.” We shall then see that the RTA is still valid, but the τ is not due to electrons scattering off the periodic array of ions but from deviations from periodicity due to phonons or static defects such as ionized impurities.

The detailed electron–electron interactions are approximated using the independent electron approximation and the concept of screening. Because of four factors, the coulombic repulsion of the valence electrons, an “effective” repulsion due to so-called exchange correlations of fermions, attraction between the ion core and core electrons, and finally the high kinetic energy of the valence electrons in the vicinity of the ion cores, the net result is that one can generally ignore the many-body aspects of the Hamiltonian, H , and instead consider the one-electron Schrödinger equation with an effective “screened” periodic lattice potential.

$$H\psi_0 = \left[-\frac{\hbar^2 \nabla^2}{2m} - eV_{\text{lat}}(r) - eV(\vec{r}, t) - eV_{\text{ext}}(\vec{r}, t) \right] \psi_0 = E\psi_0 \quad (4.3)$$

Here ψ_0 is the one-electron wavefunction. Solving this equation, first without the scattering and external potentials, gives us the one-electron bandstructure.⁴ The information of the electron–lattice interaction is thus subsumed in the bandstructure, $E(k)$, and the “effective” mass of the electron. Once the bandstructure is computed, one no longer needs to retain the periodic lattice potential explicitly in the Hamiltonian. Simply replacing the true free electron mass, m , by the effective mass, m^* , one approximately accounts for the periodic lattice potential:

$$H\psi = \left[-\frac{\hbar^2 \nabla^2}{2m^*} - eV(\vec{r}, t) - eV_{\text{ext}}(\vec{r}, t) \right] \psi = (E - E_c)\psi \quad (4.4)$$

Here, the electron energies, E , are measured with respect to the band-edge energies, E_c , and the true wavefunction ψ_0 is replaced by the envelope function, ψ .⁶ This is the one electron effective-mass equation in the envelope function approximation, which forms the theoretical underpinning of semiclassical electron transport. It can be solved for any externally applied potential with a slow spatial variation (on a much longer distance scale than the De Broglie wavelength of electrons) assuming the electrons to be Newtonian point masses, albeit with an effective mass different from the true free-electron mass. In fact, this is the reason the model is called *semiclassical dynamics*. Part of the treatment involving the lattice potential and the scattering potential is quantum-mechanical, while the part involving the external potential is classical Newtonian mechanics.³ The rationale behind this dichotomy is the same as why light has to be treated as waves if the dimensions of the obstacles are comparable to the wavelength of light, while if the dimensions are much larger, geometric ray optics suffices. Before we continue with a detailed description of semiclassical dynamics and quantum theory of scattering, we digress briefly with a discussion of bandstructure.

4.2.2 Dispersion Relationship

The single-electron bandstructure, $E(k)$, is the dispersion relationship for electrons determined by the solution of the Schrödinger equation^{3,4}

$$H\psi = E\psi \quad (4.5a)$$

where the classical Hamiltonian is

$$H = \text{KE} + \text{PE} = \frac{p^2}{2m} + V(r) \quad (4.5b)$$

where KE is kinetic energy and PE is potential energy.

We start our discussion by first examining the trivial case of a free electron moving in a constant (zero) potential in one dimension. The corresponding time-independent Schrödinger equation is

$$\frac{-\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V(x)\psi = E\psi \quad (4.6a)$$

$$\downarrow \\ V = \text{constant} = 0$$

$$\frac{d^2\psi}{dx^2} + k^2\psi = 0 \quad k = \sqrt{\frac{2mE}{\hbar^2}} \quad (4.6b)$$

Putting in the time dependence, the two stationary (i. e., constant-energy) solutions of the above are

$$\begin{aligned} \psi(x, t) &= A_+ e^{+i(kx - \omega t)} && \rightarrow +x \text{ direction} \\ &+ A_- e^{+i(-kx - \omega t)} && \rightarrow -x \text{ direction} \end{aligned} \quad (4.7)$$

where we use a normalization volume, V , for the wavefunction so that

$$A_+ = \frac{1}{\sqrt{V}} \rightarrow 0 \quad \text{as} \quad V \rightarrow \infty.$$

We get two sets of plane waves, of which we consider only the positive x direction. Using this solution, we can compute expected values of the momentum and energy using the usual recipe in quantum mechanics.

$$\langle p \rangle = \left\langle \psi \left| \frac{\hbar}{i} \frac{\partial}{\partial x} \right| \psi \right\rangle = \hbar k = \frac{h}{\lambda} \quad (4.8a)$$

where the De Broglie $\lambda = h/p$

$$\langle E \rangle = \left\langle \psi \left| \frac{-\hbar}{i} \frac{\partial}{\partial t} \right| \psi \right\rangle = \frac{\hbar^2 k^2}{2m} = \frac{p^2}{2m} = \hbar\omega \quad (\text{Planck relationship}) \quad (4.8b)$$

This is just the kinetic energy for the free particle. This gives the parabolic “bandstructure,” $E(k)$, of an electron in free space. The corresponding equienergy surfaces are perfectly spherical, centered at $k=0$. Any potential energy (which, in

general, is going to be spatially varying) has to be added to this:

$$E = \frac{p^2}{2m} + V(r) = \frac{\hbar^2 k^2}{2m} + V(r) \quad (4.9)$$

In “simplified” band diagrams, the band-edge value, $E(k)$ at $k = 0$, tracks the potential energy as a function of real space, r . The slopes of these simplified band diagrams, thus, reflect the electric field at various points in space.

The dynamics of these electrons follow from Hamilton’s equations of motion:

$$\dot{p} = -\frac{\partial H}{\partial r}; \quad \dot{r} = \frac{\partial H}{\partial p} \quad (4.10a)$$

From the first equation, we get for $V(r) = \text{constant}$

$$\dot{p} = 0$$

Otherwise

$$\dot{p} = -\frac{\partial V}{\partial r} = F = \hbar \dot{k} \quad (4.10b)$$

From the second equation, we find

$$\dot{r} = v = \frac{p}{m} = \frac{\hbar k}{m} = \frac{1}{\hbar} \frac{dE}{dk} \quad (4.10c)$$

Actually, if we consider plane-wave eigenstates as above, they have the same probability density function at all points in space. Hence, to treat transport problems, one needs to somehow localize these electrons by creating wavepackets, by summing over a range of k vectors as follows:

$$\psi(r, t) = \sum_{k'} g(k') \exp \left[i \vec{k}' \cdot \vec{r} - \frac{\hbar k'^2 t}{2m} \right] \quad (4.11)$$

The “uncertainty” in k that one thereby introduces is related to the uncertainty in the electron location by the Heisenberg relationship. The group velocity of these wavepackets for free electrons (or, as seen below, for Bloch electrons in a periodic crystal potential) is given by

$$\dot{\vec{r}} = \vec{v} = \frac{\partial \omega}{\partial \vec{k}} = \frac{1}{\hbar} \frac{\partial E}{\partial \vec{k}} \quad \text{or} \quad \frac{1}{\hbar} \vec{\nabla}_k E(k) \quad (4.12)$$

using the Planck relationship. The electron velocity is clearly perpendicular to the equienergy, $E(k)$, surfaces. We see that an electron in the stationary state $E(k)$ has a

nonvanishing velocity that continues *forever* in a *perfectly* periodic crystal potential. There can be no scattering of the propagating electron wave by the fixed, periodic array of ions, as in the Drude model, because we have already accounted for this crystal potential in the band structure. Only *deviations* from periodicity can cause scattering.

It is reassuring to see that we get the Newton's laws of motion for the simple case (Eq. 4.10). What makes these results very useful and interesting is that they are applicable with minor modifications to semiclassical dynamics for electrons moving in a crystal.³

Note that

$$m = \frac{\hbar^2}{\left(\frac{d^2 E}{dk^2}\right)} \quad (4.13)$$

In other words, the mass is inversely related to the curvature of the bandstructure. If Eq. 4.13 is applied to a realistic bandstructure of a real crystal, it gives us the "effective" mass of the electron.

Next, we briefly review the solution of the Schrödinger equation in a one-dimensional (1D) potential well or "box" of finite width, with infinite walls, to point out the qualitative differences of the solution with the free-electron case. This is important to appreciate what happens in a crystal. The Schrödinger equation (Eq. 4.6b) looks the same as before. The difference is in the boundary conditions imposed at the ends of the box where the wavefunction is forced to vanish at $x = 0$ and $x = a$ because of the infinite potential.

$$\psi_n(x) = A_n \sin k_n x \quad (4.14)$$

where $k_n = n\pi/a$ ($n = \pm 1, \pm 2, \dots$) and

$$E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{\hbar^2 \pi^2 n^2}{2ma^2} \quad (4.15)$$

The quantum numbers n and the corresponding eigenenergies are determined by imposing the requirement that the wavefunction be square-integrable. The key difference from the previous free-electron case is that for the former, *all* values of k were allowed and we got a continuous $E(k)$ relationship for plane-wave eigenstates. This is basically because the Hamiltonian is translationally invariant for any arbitrary translation, because $V(r)$ is assumed to be constant. On the other hand, breaking the translational invariance of H in real space for a potential well now allows only *discrete* values of k and $E(k)$.³

In a crystal, we have the potential profile shown schematically in Figure 4.1. The electrons are confined in a "finite" potential box whose width corresponds to the size of the finite crystal, and whose potential barriers are related to the workfunction (i.e., difference between the Fermi level and the local vacuum level). For conduction

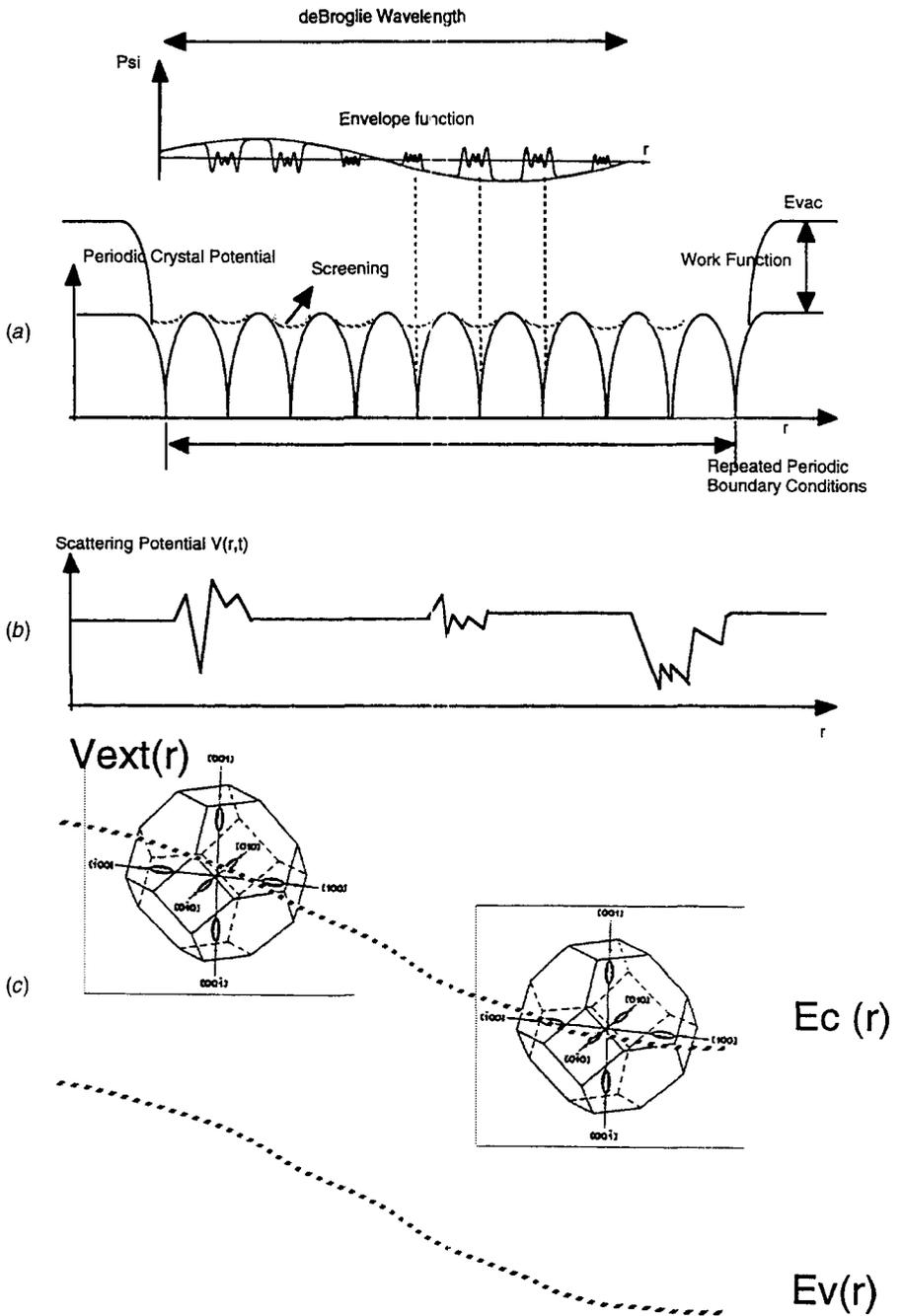


Figure 4.1 The (a) periodic, (b) scattering, and (c) external potentials (V_{ext}) considered in semiclassical electron dynamics, along with the corresponding wavefunctions and envelope functions for a certain De Broglie wavelength. The k -space description in terms of bandstructure is superposed to show the dynamics of the electron.

electrons in a semiconductor, the barriers are actually given by the electron affinity, or the difference between the conduction band edge and the vacuum level. The bottom of this potential “box” is not constant, as in the case considered before, but rather has another periodic potential corresponding to the screened periodic lattice potential:

$$V(r + \vec{R}) = V(r) \quad (4.16a)$$

The finite crystal is converted to an “infinite” crystal through the mathematical artifice of the Born–Von Karman periodic boundary condition, which states that *any* function, including, for example, the wavefunction should have the same value on either face of the finite crystal. The rationale behind this is that it is mathematically simpler to describe electron dynamics and bandstructure in an infinite solid, and as long as we are interested in “bulk” properties, as opposed to surface effects, the essence of the problem has not been changed:

$$\psi(r) = \psi(r + L) = \psi(r + Na) \quad (4.16b)$$

In other words, we are considering a particle in a finite box of length $L (= Na)$ where we have N lattice points in the finite crystal along that particular direction, with a superimposed periodic modulating function corresponding to the screened ion core potential. Here a and R refer to the lattice constant and Bravais lattice vector, respectively, and are clearly related to each other.

Therefore, we have two levels of periodicity:³

1. Periodicity of the primitive cell, a or R , which is tantamount to quantization in a box of size a . Correspondingly we get a set of eigenvalues (particle-in-a-box states), corresponding to different band indices n or reciprocal lattice vectors, \vec{K} .
2. Born–Von Karman periodic boundary condition of the crystal over distance L , which corresponds to quantization in a box of size L . This gives rise to a set of eigenvalues corresponding to different, discrete values of wavevector k . Typically L is so large, that the k vectors are very closely spaced and we get an almost continuous $E(k)$.

It is well known that any periodic function can be described more simply and elegantly in terms of its corresponding Fourier series, in the *conjugate* space. For example, if we have a sampled periodic signal in the time domain t , we get a discrete Fourier transform (DFT) in the conjugate frequency domain where the highest frequency is the so-called Nyquist frequency given by the inverse of the temporal sampling rate. The spacing between the frequency components in the DFT spectrum is inversely related to the time period of the signal:

$$f(t) \xleftrightarrow{\text{DFT relation}} \tilde{f}(\omega)$$

On the other hand, when we look at any kind of wave propagation [electron De Broglie waves, elastic waves (phonons), or electromagnetic waves with wavelengths comparable to the lattice spacing (X ray)] in the discrete, periodic medium of the crystal, we get a Fourier transform relationship in r space and k space, as follows. The crystal medium can be considered to be *discrete* because the atoms occupy discrete locations in space, just like a *sampled* time-dependent signal.

$$f(\vec{r}) \longleftrightarrow \tilde{f}(\vec{k})$$

Any function which has the periodicity of the Bravais lattice, such as the lattice potential or the periodic part of the Bloch function, will have a Fourier expansion as follows:

$$u_k(\vec{r}) = u_k(\vec{r} + \vec{R}) \quad (4.17a)$$

$$u_k(\vec{r}) = \sum_{\vec{K}} A_{\vec{K}} e^{i\vec{K} \cdot \vec{r}} \quad (4.17b)$$

where

$$A_{\vec{K}} = \frac{1}{v} \int_v u_k(\vec{r}) e^{-i\vec{K} \cdot \vec{r}} d\vec{r} \quad (4.17c)$$

and v = volume of primitive cell.

Here, the Fourier components are in terms of the reciprocal lattice vectors, \vec{K} .

$$\vec{K} = 2\pi \left[\frac{\ell_x}{a_1} \hat{x} + \frac{\ell_y}{a_2} \hat{y} + \frac{\ell_z}{a_3} \hat{z} \right] \quad (4.18)$$

Here, the ℓ 's are integers, and the a 's are the primitive vectors in real space for the Bravais direct lattice.

For a direct lattice, $\{\vec{R}\}$, the reciprocal lattice $\{\vec{K}\}$ is defined formally as

$$e^{i\vec{K} \cdot \vec{R}} = 1 \quad (4.19)$$

Direct lattice: primitive vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$

Reciprocal lattice: primitive vectors $\vec{b}_1, \vec{b}_2, \vec{b}_3$, where

$$\vec{b}_1 = 2\pi \frac{\vec{a}_2 \times \vec{a}_3}{\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)} \quad (4.20a)$$

$$\vec{b}_2 = 2\pi \frac{\vec{a}_3 \times \vec{a}_1}{\vec{a}_2 \cdot (\vec{a}_3 \times \vec{a}_1)} \quad (4.20b)$$

$$\vec{b}_3 = 2\pi \frac{\vec{a}_1 \times \vec{a}_2}{\vec{a}_3 \cdot (\vec{a}_1 \times \vec{a}_2)} \quad (4.20c)$$

If we construct the reciprocal lattice primitive vectors using this equation, we have

$$\vec{b}_j \cdot \vec{a}_i = 2\pi\delta_{ij}. \quad (4.21)$$

This, then, satisfies the formal definition of the reciprocal lattice (Eq. 4.19), which compactly describes the periodicity of the *direct* lattice. Once we determine the reciprocal lattice, we can construct a special primitive cell, the Wigner–Seitz cell by choosing any K point and drawing perpendicular bisectors to neighboring K points in reciprocal space (Bragg planes). This gives us the Brillouin zone, whose special physical significance is that it has k -space values closer to the origin than to any other \vec{K} point. As such, it turns out that it contains all relevant k -space information. The \vec{K} corresponds to the highest, meaningful k vector, and is the analog of the Nyquist frequency in a DFT.

For a Si face-centered cubic (FCC) direct Bravais lattice with a lattice constant a in real space r , we get

$$\vec{b}_1 = \frac{4\pi}{a} \frac{1}{2} (\hat{y} + \hat{z} - \hat{x}) \quad (4.22a)$$

$$\vec{b}_2 = \frac{4\pi}{a} \frac{1}{2} (\hat{z} + \hat{x} - \hat{y}) \quad (4.22b)$$

$$\vec{b}_3 = \frac{4\pi}{a} \frac{1}{2} (\hat{x} + \hat{y} - \hat{z}) \quad (4.22c)$$

In other words, we get a body-centered cubic (BCC) reciprocal lattice in reciprocal space, k , with side $4\pi/a$.

The volume of a cell in a reciprocal lattice is

$$V = \frac{(2\pi)^3}{(\text{volume of primitive cell in direct lattice})} \quad (4.23)$$

Just as any function with the periodicity of the direct lattice R can be expanded in terms of K , any function with the Born–Von Karman periodic boundary condition over distance L can be Fourier-expanded as follows:

$$\Psi_k(r+L) = \Psi_k(r) \quad (4.24a)$$

$$\Psi_k(r) = \sum_k C_k e^{i\vec{k}\cdot\vec{r}} \quad (4.24b)$$

where

$$C_k = \frac{1}{V} \int_V \Psi_k(r) e^{-i\vec{k}\cdot\vec{r}} d\vec{r} \quad (4.24c)$$

and V = volume of crystal.

The periodic boundary condition gives the allowed values of \vec{k} as follows:

$$\vec{k} = \frac{h_1}{N_1} \vec{b}_1 + \frac{h_2}{N_2} \vec{b}_2 + \frac{h_3}{N_3} \vec{b}_3 \quad (4.25)$$

Here, the h 's are integer quantities from 1 to N . Hence, the volume per k point

$$= \frac{\vec{b}_1 \cdot (\vec{b}_2 \times \vec{b}_3)}{N_1 N_2 N_3} = \left(\frac{2\pi}{N_1 a_1} \right) \left(\frac{2\pi}{N_2 a_2} \right) \left(\frac{2\pi}{N_3 a_3} \right) = \frac{(2\pi)^3}{L^3} \quad (4.26)$$

In other words, combining this equation with Eq. 4.23, we see that the number of allowed quantum states is given by the the total number of lattice points in the finite crystal, N , and it thus scales with the crystal volume. This equation (Eq. 4.26) is also extremely useful because it gives us the so-called phase-space density in k space, and as shown later in our discussion of carrier statistics, enables us to compute the density of states in k space.

4.3 SEMICLASSICAL ELECTRON DYNAMICS

For the electron wavefunction in free space, we saw earlier that

$$\psi = \frac{1}{\sqrt{V}} e^{i\vec{k} \cdot \vec{r}} \quad \text{where} \quad H = \frac{p^2}{2m} \quad (4.27)$$

Now, for electrons in a periodic crystal potential where

$$H = \frac{p^2}{2m} + V(\vec{r}) \quad \text{and} \quad V(\vec{r} + \vec{R}) = V(\vec{r}) \quad (4.28)$$

the eigenstates are no longer plane waves, but are given by the Bloch–Floquet wavefunctions:

$$\psi \sim e^{i\vec{k} \cdot \vec{r}} U_{\vec{k}}(\vec{r}) \quad (4.29)$$

where $U_{\vec{k}}(\vec{r} + \vec{R}) = U_{\vec{k}}(\vec{r})$.

In other words, we get plane waves modulated by a function with the same periodicity as the lattice potential for these Bloch electrons subject to semiclassical electron dynamics. The properties of these electrons bear a remarkable resemblance to those of the free electrons considered earlier.

Thus, to summarize our discussion of semiclassical electron dynamics, we categorize the forces or potentials acting on the valence/conduction electrons as follows (see Fig. 4.1):

1. *Periodic crystal potential.* This is treated quantum-mechanically for bandstructure, $E(k)$, and m^* because the De Broglie wavelength of electrons is greater than the periodicity of the lattice potential.
2. *Random scattering potential due to impurities and/or defects* [$V(r)$], *or phonons* [$V(r,t)$]. This is treated quantum-mechanically by time-dependent perturbation theory and the Fermi golden rule.
3. *External driving potential due to applied voltages or EM waves:* This is treated classically by Newtonian mechanics using the effective mass, assuming spatial variations of external potential are on longer distance scales than the De Broglie wavelength. Otherwise, a quantum treatment is required *even* for the external forces.

In exact analogy with the free-electron case (Eqs. 4.10b, 4.12), the two governing equations of these semiclassical Bloch electrons are^{3,7} reproduced here as Eqs. (4.30a) and (4.30b):

$$\vec{v} = \frac{1}{\hbar} \vec{\nabla}_k E(k) \quad (4.30a)$$

This implies that the group velocity of an electron wavepacket is perpendicular to the equienergy surfaces in the 3D band diagram, and that the electron wave travels forever in a perfectly periodic lattice potential without scattering. Scattering is caused only by deviations from periodicity, as discussed in detail in a later section.

$$\hbar \dot{\vec{k}} = -e(\vec{\varepsilon} + \vec{v} \times \vec{B}) \quad (4.30b)$$

This indicates that the time evolution of the momentum is given by the Lorentz force. Since we are not considering the total force on the right-hand side, but just the external force, the electron momentum here clearly cannot be the true momentum. It is called the *quasimomentum* or *crystal momentum*. As described below, these two governing equations are solved self-consistently, and sequentially with the Poisson equation in Monte Carlo simulations using a single-particle approach, or a collective approach as with the BTE.

For such Bloch electrons, with the modulated plane-wave functions, the $E(k)$ relationship is no longer perfectly parabolic and the equienergy surfaces are not exactly spherical. However, invoking the effective mass theorem, they are approximately so near the band edges. The details of the effective mass are described next.

We start by considering the time derivative of the electron velocity given by semiclassical dynamics, in order to determine the acceleration of the electrons.

$$\frac{d\vec{v}}{dt} = \frac{1}{\hbar} \frac{d}{dt} \nabla_k E(k) = \frac{1}{\hbar} \nabla_k \left[\nabla_k E \cdot \frac{d\vec{k}}{dt} \right] \quad (4.31a)$$

or

$$\vec{a} = \frac{1}{\hbar^2} \nabla_k [\nabla_k E(k) \cdot \vec{F}] \quad (4.31b)$$

We see that the acceleration, in general, is in a direction different from the direction of external force, \vec{F} :

$$a_i = \frac{1}{\hbar^2} \sum_j \frac{\partial^2 E}{\partial k_i \partial k_j} F_j \quad \text{for } i, j = 1, 2, 3, \dots \quad (4.31c)$$

This is not a repudiation of Newtonian mechanics per se. The electrons accelerate in a direction given by the total force (external plus internal periodic lattice forces), where the internal periodic potential information has been encapsulated in the bandstructure, $E(k)$. This, therefore, enables us to define an inverse effective mass tensor as

$$\frac{1}{m_{ij}} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_i \partial k_j} \quad (\text{curvature of } E(k)) \quad (4.32)$$

We can diagonalize this (3×3) inverse mass tensor. If there is an extremum of $E(k)$ at $k = k_0$, we can perform a Taylor series expansion of $E(k)$ at $k = k_0$:

$$E = E(k_0) \pm \frac{\hbar^2}{2} \sum_i \frac{(k_i - k_{0i})^2}{m_i} \quad (4.33a)$$

$$E = E(k_0) \pm \frac{\hbar^2}{2} \left(\frac{(k_1 - k_{01})^2}{m_1^*} + \frac{(k_2 - k_{02})^2}{m_2^*} + \frac{(k_3 - k_{03})^2}{m_3^*} \right) \quad (4.33b)$$

For the conduction band in Si, $k_0 \neq 0$, but is approximately at 80% toward the X point in the Brillouin zone. There are, therefore, six equivalent ellipsoidal $E(k)$ equienergy surfaces:

$$E(k) = \frac{\hbar^2}{2} \left(\frac{(k_l)^2}{m_l} + \frac{2(k_t)^2}{m_t} \right) \quad (4.33c)$$

Here, m_l^* = longitudinal effective mass and m_t^* = transverse effective mass.

For the valence band in Si, $k_0 = 0$. The hole equienergy surfaces are warped spheres,⁵ so that, approximately

$$m_1^* = m_2^* = m_3^* = m^*$$

Now, however, we have two sets of degenerate valence bands centered at the gamma point, a heavy-hole band with an effective mass $m^* = m_{\text{hh}}^*$, and a light-hole band

with a mass m_{lh}^* . There is also a splitoff band, separated in energy from the heavy- and light-hole bands by the spin-orbit coupling.⁵ These band-curvature-tensor effective masses have to be combined in physically meaningful ways to determine the density-of-states effective mass (geometric mean) and the conductivity effective mass (harmonic mean), as discussed below.

4.4 EQUILIBRIUM STATISTICS

Now that we have seen how individual carriers move in a semiconductor, we need to determine how many carriers there are in the various bands to find the collective behavior of the ensemble. The overall recipe is first described, followed by detailed calculations. Clearly, the total number of electrons in a band will depend on the number of quantum states available, or the density of states (DOS), which we briefly touched earlier, as well as the probability with which these available states are occupied, which is given by the Fermi-Dirac (FD) distribution function:^{1,2}

$$n = \int_{E_c}^{\infty} \underbrace{g(E)}_{\text{DOS}} \underbrace{f(E)}_{\text{FD Prob}} dE \quad (4.34)$$

1. We first count states in k space; this is the fundamentally correct way to count quantum states.
2. Next, we can use the bandstructure $E(k)$ to count states in E space. The rationale for doing so is that the FD probabilities depend on electron energies E , rather than directly on k . For example, we know that near band extrema, we have a parabolic $E(k)$ relationship, which leads to a parabolic density of states in three dimensions.
3. We next multiply the density of states by the Fermi-Dirac distribution function. For electrons in band n in the repeated zone band scheme, where n is the index for K (e.g., valence band, conduction band,) due to quantization inside the primitive cell, we get

$$f[E_n(k)] d\vec{r} \frac{d\vec{k}}{4\pi^3} = \left(\frac{1}{4\pi^3} \right) \frac{d\vec{r} d\vec{k}}{\exp\left(\frac{E_n(k) - E_F}{kT}\right) + 1} \quad (4.35)$$

This is the number of electrons in $d\vec{r} d\vec{k}$ because of FD statistics in the n th band. Here, we have accounted for spin degeneracy of the electrons by using an additional factor of 2 in the phase space density.

4. Finally, to get the total number of electrons in a band, we integrate from E_c to $E_{c, \text{top}} \cong \infty$

With the overall calculation scheme in mind, we next look at the details. First, we determine the DOS in various cases. We have already seen that the number of states

in a k space of Δk (for three dimensions)

$$= \left\{ \frac{L^3}{(2\pi)^3} \Delta k \right\} \times (2) \text{ spin} \quad (4.36a)$$

Per unit volume, for three dimensions, the number of states

$$= \frac{2}{(2\pi)^3} (\Delta k) \quad (4.36b)$$

In general, for N dimensions, per unit volume, we can generalize this expression as

$$\text{Number of states} = \frac{2}{(2\pi)^N} (\Delta k) \quad (4.36c)$$

As mentioned above, we can then transform from k space to E space using the $E(k)$ relationship:

$$g(E) \Delta E = \frac{2}{(2\pi)^N} (\Delta k) \quad (4.36d)$$

The simplest bandstructure is parabolic

$$E(k) = \frac{\hbar^2 k^2}{2m^*} \quad (4.37)$$

$$k = \sqrt{\frac{2m^*E}{\hbar^2}}; \quad dk = \left\{ \sqrt{\frac{m^*}{2\hbar}} \right\} \frac{1}{\sqrt{E}} dE \quad (4.38)$$

In general, the DOS in N dimensions ($N = 1, 2, 3$) then becomes

$N = 1$ (1D quantum “wire”):

$$\Delta k = 2(dk); \quad dk = \left\{ \sqrt{\frac{m^*}{2\hbar}} \right\} \frac{1}{\sqrt{E}} dE \quad (4.39a)$$

$$g(E)dE = \frac{2}{(2\pi)^1} (2dk) = \frac{\sqrt{2m^*}}{\pi\hbar\sqrt{E}} dE \quad (4.39b)$$

$N = 2$ (2D electron or hole gas, as in a MOSFET inversion layer):

$$\Delta k = (2\pi k)dk; \quad k dk = \frac{m^*}{\hbar^2} dE \quad (4.40a)$$

$$g(E)dE = \frac{2}{(2\pi)^2} (2\pi k)dk = \frac{m^*}{\pi\hbar^2} dE \quad (4.40b)$$

$N = 3$ (3D bulk):

$$\Delta k = 4\pi k^2 dk \quad (4.41a)$$

$$g(E)dE = \frac{2}{(2\pi)^3} 4\pi k^2 dk = \frac{\sqrt{2}}{\pi^2} \left(\frac{m^*}{\hbar^2}\right)^{3/2} E^{1/2} dE \quad (4.41b)$$

We see that for current generation ULSI devices, the most important cases are the 3D parabolic DOS, and the 2D constant DOS. The 2-DEG in an inversion layer is discussed below, where we shall see that by combining the constant 2D DOS with the bulk 3D DOS, we get a so-called “staircase” DOS with very important ramifications.⁷ Future, more esoteric ULSI devices may involve quantum wires or zero-dimensional “dots” which are characterized by singularities in the DOS. All of the above is true, with the caveat that the $E(k)$ relationship is parabolic. While that is generally a good assumption near the band edges (that are populated), there are definite nonparabolicities in the bandstructure at higher energies, with concomitant effects on the DOS. This will be very important later when we look at carrier scattering rates, especially at high fields in small devices. The carrier energies can be so high that the nonparabolic DOS is probed in terms of the states available for electrons to scatter to. The DOS for any arbitrary $E(k)$ can be determined as follows.

The total volume in reciprocal space between E and $(E + dE)$ is a surface integral

$$= \int_s ds dk_n \quad (4.42)$$

$$\text{Phase space density} = \underbrace{2}_{SPIN} \left(\frac{1}{2\pi}\right)^3 V \quad (4.43)$$

$$g(E)dE = \frac{V}{4\pi^3} \int_s ds dk_n \quad (4.44a)$$

$$dE = \nabla_k E \cdot \vec{dk} = |\nabla_k E| dk_n \quad (4.44b)$$

$$g(E)dE = \frac{V}{4\pi^3} \int \frac{ds}{|\nabla_k E(k)|} dE \quad (4.44c)$$

The DOS, $g(E)$, is no longer smooth, but has Van Hove singularities when $|\nabla_k E(k)| = 0$.

4.4.1 Fermi-Dirac Statistics

Next, we focus on Fermi–Dirac statistics. We will not go through the details of the calculations, but merely point out the physical assumptions involved.^{2,7} The distribution function is determined by calculating the number of distinct ways we can put n_k indistinguishable electrons in g_k states at an energy level E_k , subject to the

Pauli exclusion principle:

$$w_k = \frac{(g_k)(g_k - 1)(g_k - \overline{n_k - 1})}{n_k!} = \frac{g_k!}{(g_k - n_k)!n_k!} \quad (4.45a)$$

For N levels in a band, the number of distinct ways we can put in the various electrons gives us the so-called multiplicity function:

$$W_b = \prod_k W_k =: \prod_k \frac{g_k!}{(g_k - n_k)!n_k!} \quad (4.45b)$$

To the question “What is the most probable distribution of the n_k electrons in the various E_k levels (degeneracy of g_k in level E_k)?” the statistical-mechanical answer is “In thermal equilibrium, the distribution that is most disordered (i.e., has the maximum entropy), or that can occur in the largest number of ways is the most probable.”

We, therefore, have to maximize W_b with respect to n_k . We assume here that the probability of occupancy of each allowed (degenerate) quantum state is the same. We also assume that the total number of electrons in the band is fixed:

$$\sum_k n_k = \text{constant} \Rightarrow \sum_k dn_k = 0 \quad (4.46a)$$

and that the total energy in the band is constant:

$$E_{\text{tot}} = \sum_k E_k n_k = \text{constant} \Rightarrow \sum_k E_k dn_k = 0 \quad (4.46b)$$

We invoke the Boltzmann definition of entropy

$$S = k \ln W \quad (4.47)$$

to get the Fermi–Dirac distribution function:

$$f(E_k) = \frac{1}{\exp\left[\frac{E_k - E_F}{kT}\right] + 1} \quad (4.48a)$$

For the limit of high energies, $E \gg E_F$, we obtain

$$f(E) \cong \exp\left[-\frac{E_F - E}{kT}\right] \quad (4.48b)$$

This is the classical Maxwell–Boltzmann limit of the Fermi–Dirac distribution function, where the Fermi level E_F has a clear thermodynamic interpretation. The Euler relation in thermodynamics relates the total internal energy in the semiconductor to the intensive variables, temperature (T), pressure (P), chemical

potential (μ), and electrostatic potential (ψ), as well as extensive or system-size-dependent variables, entropy (S) volume (V), number of electrons (n_s), and charge ($Q = -en_s$):

$$E_{\text{total}} = TS - PV + \mu n_s + \psi Q = TS - PV + n_s(\mu - e\psi) \quad (4.49a)$$

This can be rewritten in terms of the Helmholtz free energy:

$$F = E_{\text{total}} - TS = -PV + n_s(\mu - e\psi) \quad (4.49b)$$

Finally, the Fermi level is recognized to be the change of the Helmholtz free energy or the electrochemical potential when an electron is added to or removed from the device. This implies that gradients in the Fermi level cause carrier flow due to the electrical forces or the gradients in the electrostatic potential (drift), plus chemical forces due to gradients in the concentration (diffusion). The Fermi level is flat in equilibrium.

$$\left. \frac{\partial F}{\partial n_s} \right|_{T,V} = (\mu - e\psi) = E_F = \xi \quad (4.50)$$

If the semiconductor is in equilibrium (i.e., no sources of excitation such as bias, or electromagnetic radiation) and the material is uniform in terms of doping and composition, $E_F = \mu$ (chemical potential). If we have a nonuniform material (doping or composition), we have a built-in electrostatic potential added to the chemical potential, so that the Fermi level is now $=\xi$ (electrochemical potential).

Once we have the probabilities of electron occupancy, the probability of hole occupancy becomes

$$1 - f(E) = \frac{1}{\exp\left(\frac{E_F - E}{kT}\right) + 1} \quad (4.51)$$

Now, we are in a position to compute the carrier densities in the conduction and valence bands:

$$\begin{aligned} n &= \int_{E_c}^{E_{\text{ctop}}} g(E) f(E) dE = \int_{E_c}^{\infty} \left\{ \left(\sqrt{2\pi} \left(\frac{2m^*}{h^2} \right)^{3/2} (E - E_c)^{1/2} \right) \left(\frac{1}{1 + \exp\frac{E - E_F}{kT}} \right) \right\} dE \\ &= 4\pi \left(\frac{2m^* kT}{h^2} \right)^{3/2} \int_{E_c}^{\infty} \left\{ \left(\frac{\left(\frac{E - E_c}{kT} \right)^{1/2}}{\exp\frac{[E - E_c - (E_c - E_F)]}{kT} + 1} \right) \right\} d \frac{(E - E_c)}{kT} \\ &= 4\pi \left(\frac{2m^* kT}{h^2} \right)^{3/2} \int_0^{\infty} \frac{\xi^{1/2} d\xi}{e^{\xi - \eta} + 1} = 2 \left(\frac{2\pi m^* kT}{h^2} \right)^{3/2} \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{\xi^{1/2} d\xi}{e^{\xi - \eta} + 1} \quad (4.52) \end{aligned}$$

Here, the integral is a special case of the Fermi–Dirac integral of order $j(= \frac{1}{2})$ as defined below.

$$F_j(\eta) = \frac{1}{\Gamma(j+1)} \int_0^\infty \frac{\xi^j d\xi}{e^{\xi-\eta} + 1} \quad (4.53a)$$

where the gamma functions,

$$\begin{aligned} \Gamma(j+1) &= j\Gamma(j); & \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi} \\ \Gamma(j) &= \int_0^\infty t^{j-1} e^{-t} dt, \quad j > 0 & \Gamma(j+1) &= j! \text{ for integer "j"} \\ F_{1/2}(\eta) &= \frac{1}{\Gamma\left(\frac{1}{2}+1\right)} \int_0^\infty \frac{\xi^{1/2} d\xi}{e^{\xi-\eta} + 1} \end{aligned} \quad (4.53b)$$

Although only the case of $j = \frac{1}{2}$ is relevant for computing carrier densities, the higher-order integrals come in handy when computing other physical quantities of interest such as the average kinetic energy of the conduction electrons.

We can finally write down how many electrons and holes we have in the semiconductor under nondegenerate conditions, where the Fermi level is several kT away from the band edge, such that the Maxwell–Boltzmann classical distribution function is valid:^{1,2}

$$n_0 = 2 \left(\frac{2\pi m_{de}^* kT}{h^2} \right)^{3/2} e^{-(E_c - E_F)/kT} \quad (4.54a)$$

$$n_0 = N_c e^{-(E_c - E_F)/kT} = N_c f(E_c) \quad (4.54b)$$

where N_c is the “effective” DOS or the weighted sum of the actual DOS, mapped to the conduction band edge, which is multiplied by the probability of occupancy at E_c .

Similarly, for holes

$$p_0 = N_v [1 - f(E_v)] \quad (4.54c)$$

$$p_0 = N_v e^{-(E_F - E_v)/kT} \quad (4.54d)$$

In undoped or intrinsic material $E_F \equiv E_i$:

$$n_i = N_c e^{-(E_c - E_i)/kT} = p_i = N_v e^{-(E_i - E_v)/kT} \quad (4.54e)$$

On the other hand, in doped or extrinsic material

$$n_0 = N_c e^{-(E_c - E_F)/kT} e^{(E_i - E_i)/kT} \quad (4.54f)$$

$$n_0 = N_c e^{-(E_c - E_i)/kT} e^{(E_F - E_i)/kT} \quad (4.54g)$$

$$n_0 = n_i e^{(E_F - E_i)/kT} \quad (4.54h)$$

$$p_0 = n_i e^{(E_i - E_F)/kT} \quad (4.54i)$$

These expressions can still be used in nonequilibrium, except that the Fermi level, which is an equilibrium thermodynamic concept, has to be replaced by separate quasi-Fermi levels or Imrefs for electrons and holes:^{1,2}

$$\begin{aligned} n_0 &= n_i e^{(E_F - E_i)/kT} \\ \Rightarrow n &= n_i e^{(E_{Fn} - E_i)/kT} \end{aligned} \quad (4.54j)$$

$$\begin{aligned} p_0 &= n_i e^{(E_i - E_F)/kT} \\ \Rightarrow p &= n_i e^{(E_i - E_{Fp})/kT} \end{aligned} \quad (4.54k)$$

The separation of the quasi-Fermi levels is dependent on how far from equilibrium we have driven the system, specifically, the applied bias. The concept of Imrefs can be justified by the concept of local equilibration of electrons in the conduction band, and of holes in the valence band on a much shorter time scale (dielectric relaxation time) than the time scale for equilibration between bands (recombination lifetimes).

For the sake of completeness, we conclude this discussion by contrasting the Fermi–Dirac distribution for fermions with half-integer spin such as electrons, with the Bose–Einstein distribution for integer spin particles of interest to us such as phonons or photons. The key difference in Eq. (4.55) is that the negative sign is applicable in the denominator for bosons, unlike the positive sign for fermions. We will use the Bose–Einstein distribution later when we perform electron–phonon scattering calculations:

$$f(E) = \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) \pm 1} \quad (4.55)$$

For the purpose of counting available states in a band, we have to determine the density-of-states effective mass as

$$g_c(E) = 4\pi \left(\frac{m_{de}^*}{\hbar^2}\right)^{3/2} (E - E_c)^{1/2} \quad (4.56a)$$

Hence, the density-of-states effective mass is the geometric mean of the band-curvature effective masses.

$$m^* = (m_1^* m_2^* m_3^*)^{1/3} \quad (4.56b)$$

Here, we must keep track of the number of equivalent minima and whether they are at the Brillouin zone edge, in which case they are only counted in part because the minima are shared with adjacent Brillouin zones.⁴

In the Si conduction band, for electrons

$$(m_{de}^*)^{3/2} = N_{\text{eq}}(m_l^* m_t^{*2})^{1/2} \quad (4.56c)$$

$$m_{de}^* = \zeta^{2/3} (m_l^* m_t^{*2})^{1/3} \quad (4.56d)$$

For the holes in the valence band in Si, on the other hand, since we have a degenerate heavy-hole band and a light-hole band at the Γ point ($k = 0$):⁵

$$g_v(E) = 4\pi \left(\frac{m_{dh}^*}{h^2} \right)^{3/2} (E_v - E)^{1/2} = 4\pi \left[\left(\frac{m_{hh}^*}{h^2} \right)^{3/2} + \left(\frac{m_{lh}^*}{h^2} \right)^{3/2} \right] (E_v - E)^{1/2} \quad (4.57a)$$

$$m_{dh}^* = [(m_{hh}^*)^{3/2} + (m_{lh}^*)^{3/2}]^{2/3} \quad (4.57b)$$

We ignore the splitoff band here because it is at a higher (hole) energy than the heavy- and light-hole bands.

If, instead, we are concerned with electron-hole pairs as individual entities as during transport issues that we consider next, we have to define the conductivity effective mass, m^* , as the harmonic mean of the band curvature effective masses.

$$m^* = 3 \left(\frac{1}{m_1^*} + \frac{1}{m_2^*} + \frac{1}{m_3^*} \right)^{-1} \quad (4.58)$$

This is because the mobility of carriers is inversely proportional to the effective mass

$$\mu = \frac{e\langle\tau\rangle}{m^*} \quad (4.59)$$

4.5 SCATTERING THEORY

As was mentioned in a previous section, semiclassical electron dynamics predicts that there is no scattering by the periodic lattice potential. Scattering is only caused by deviations from periodicity, generally a weak perturbation. One therefore uses time-dependent perturbation theory to handle the problem. We briefly sketch that formalism first, and then apply the resulting expression, the Fermi golden rule, to scattering mechanisms that are relevant for Si-based ULSI devices.⁷ We ask ourselves the following question: “What is the probability, $S(\vec{k}, \vec{k}')$, that an electron initially in an eigenstate $|\vec{k}\rangle$ of some unperturbed Hamiltonian, H_0 will scatter to a state, $|\vec{k}'\rangle$, after interaction with a perturbing (scattering) Hamiltonian that can be time-dependent, $V(\vec{r}, t)$ (e.g., phonons), or time-independent, $V(\vec{r})$ (e.g., ionized

impurities)?” To answer this question, we need to solve the time-dependent Schrödinger equation:

$$[H_0 + V(r, t)]\psi(r, t) = i\hbar \frac{\partial \psi(r, t)}{\partial t} \quad (4.60)$$

where

$$H_0 \psi_k^0 = E(k) \psi_k^0 \quad \text{and} \quad \psi_k^0(r, t) = \psi_k^0(r) e^{[-iE(k)t/\hbar]} \quad (4.61)$$

Since these eigenstates are the solutions of a Hamiltonian, they form a complete orthonormal set, and any function, including the perturbed eigenstates, can be expanded in terms of these basis states:

$$\psi(r, t) = \sum_k C_k(t) \psi_k^0(r) e^{[-iE(k)t/\hbar]} \quad (4.62)$$

We launch the electron with an initial wavevector, k . Hence, the initial condition becomes,

$$C_k(t=0) = 1, \quad C_{k'}(t=0) = 0, \quad k' \neq k$$

To find $C_{k'}(t)$, we need to substitute $\psi(r, t)$ in terms of the expansion (Eq. 4.62) into the Schrödinger equation (Eq. 4.60). The probability that the electron emerges in a new state k' after interaction with the perturbing Hamiltonian for a time t , and the corresponding transition rate are, respectively

$$p(k = k') = \lim_{t \rightarrow \infty} |C_{k'}(t)|^2 \quad (4.63a)$$

$$S(\vec{k}, \vec{k}') = \lim_{t \rightarrow \infty} \frac{|C_{k'}(t)|^2}{t} \quad (4.63b)$$

This gives us the so-called Fermi golden rule, whose utility is emphasized by its name:

$$S(k, k') = \frac{2\pi}{\hbar} |\langle k' | V(r) | k \rangle|^2 \cdot \{ \delta[(E(k') - E(k) - \hbar\omega)] + \delta[(E(k') - E(k) + \hbar\omega)] \} \quad (4.64a)$$

Let us point out the physical significance of the various terms:

1. The δ functions involving the carrier energies account for conservation of energy in the scattering process, where the first term signifies, for example, phonon absorption, and the second term, phonon emission [with phonon energy, $E(q) = \hbar\omega$]. Here q is the phonon wavevector, which is related to the

phonon frequency through the dispersion relationship. The $E(k')$ and $E(k)$ are well defined only if the lifetimes in these states are long. For high scattering rates, and consequently short lifetimes, the Heisenberg uncertainty relationship introduces a Lorentzian lineshape (collisional) broadening:

$$\Delta E \cong \frac{\hbar}{\Delta t}$$

Although we will talk about phonons as the prototypical time-dependent perturbation in this chapter, this formalism is quite universal, and is applicable to optical problems involving electron-photon interactions also.

2. If the perturbation is time-independent (e.g., ionized dopants or crystal defects), we have what is called a *static scatterer*, rather than a *dynamic scatterer* (as for phonons). Then the scattering is elastic, and there is no change in the carrier energy (only the carrier momentum) after scattering:

$$S(k, k') = \frac{2\pi}{\hbar} |\langle k' | V | k \rangle|^2 [\delta(E(k') - E(k))] \quad (4.64b)$$

This is a very important point that we will come back to in the final section on quantum transport in mesoscopic systems. We shall see then that static scatterers cannot randomize the phase of electrons, unlike dynamic scatterers and, thus, the phase interference in a device is retained over longer distances at lower temperatures, where the occupancy of the phonon “bath” is smaller.

3. We next shift our attention from the temporal variation of the scattering potential to the spatial variation, $V(r)$. First, we Fourier-decompose the perturbation Hamiltonian in the spatial domain:

$$V(r) = \sum_q V_q e^{i\vec{q}\cdot\vec{r}} \quad (4.65)$$

Then, we compute the matrix element of the perturbing Hamiltonian in terms of these components:

$$\langle k' | V | k \rangle = \sum_q \frac{V_q}{\text{Vol}} \int_{\text{Vol}} e^{i(k-k'+q)\cdot\vec{r}} d\vec{r} = \sum_q V_q \delta(k - k' + q) \quad (4.66)$$

It may be noted that the δ function takes care of momentum conservation, just as we earlier accounted for energy conservation during the collision:

$$\vec{k}' - \vec{k} = \vec{q}$$

It may also be noted that the matrix element is nothing but the Fourier transform of $V(r)$.

$$\langle k' | V | k \rangle = V_{k'-k}$$

If $V(r) = \text{constant} = V_0$, then

$$\langle k' | V(r) | k \rangle = V_0 \delta(k - k') \quad (4.67)$$

This treatment is valid as long as the Fourier components, $V_q \ll$ Fourier components of the crystal potential (i.e., the effective-mass theorem is valid).⁶ Also, although we have illustrated the technique using plane-wave eigenstates, as mentioned before, we actually need Bloch states. The main difference is that the periodic part of the Bloch function introduces another term in the overlap integral, as shown below:

$$\begin{aligned} \langle k' | V(r) | k \rangle &= I(k, k') \int e^{-i\vec{k}' \cdot r} V(r) e^{i\vec{k} \cdot r} \\ I(k, k') &= \int u_{k'}^*(r) u_k(r) dr \quad (\text{overlap integral}) \end{aligned} \quad (4.68)$$

We shall, however, ignore such mathematical “details” for the rest of the discussion.

Once the formalism of the Fermi golden rule is appreciated, it is easy, in principle, to understand the detailed computations for different scattering mechanisms. The difficult part is to get the physics right in terms of the perturbing, scattering Hamiltonian. We illustrate the calculations using the scattering mechanisms that are most relevant for Si ULSI devices.

4.5.1 Ionized Impurity Scattering

An ionized dopant impurity (acceptor or donor) is a point charge that, by itself, has a coulombic potential:

$$V_{ap}(r) = -\frac{e}{4\pi \epsilon_0 r} \quad (4.69a)$$

This can be Fourier transformed to yield

$$V_{ap}^q = -\frac{e}{\epsilon_0 q^2} \quad (4.69b)$$

The point-fixed dopant charge causes redistribution of the mobile majority carriers in the immediate vicinity, and causes “screening” of the bare coulombic potential over a Debye screening length, L_D . The true potential then becomes

$$V_{tr}(r) = \frac{e}{4\pi \epsilon_r} e^{-r/L_D} \quad (4.69c)$$

This screening is, thus, basically a result of the polarization of the semiconductor dielectric medium, which can be succinctly described in terms of a wavevector-dependent dielectric “constant”:

$$\epsilon = \epsilon_0 \epsilon_r \left(1 + \frac{q_s^2}{q^2 \epsilon_r} \right), \quad q_s = \frac{2\pi}{L_D} \quad (4.69d)$$

This gives the Fourier components of the screened coulombic potential:

$$V_{tr}^q = \frac{e}{\epsilon_0 \epsilon_r (q^2 + q_s^2 / \epsilon_r)} \quad (4.69e)$$

These are the matrix elements of the scattering (perturbing) Hamiltonian:

$$\langle k' | H_i | k \rangle = \frac{1}{\text{Vol}} \int_{\text{vol}} d\vec{r} (e V_{tr}) e^{i(k-k') \cdot r} \quad (4.70a)$$

$$M_{kk'} = \frac{1}{\text{Vol}} \frac{e^2}{\epsilon_0 \epsilon_r (q^2 + q_s^2 / \epsilon_r)} \quad (4.70b)$$

The scattering rate from the Fermi golden rule is then given by

$$S(k, k') = \frac{2\pi}{\hbar} \frac{(\text{Vol})^2 e^4}{\epsilon_0^2 \epsilon_r^2 (q^2 + q_s^2 / \epsilon_r)^2} \cdot \delta(E(k) - E(k')) \quad (4.70c)$$

Notice that the energy of the carriers is conserved because ionized impurity scattering from ions (which are much heavier than the carriers) is elastic. Also momentum is conserved.

$$\Rightarrow \vec{q} = \vec{k} - \vec{k}', \text{ where } |\vec{k}| = |\vec{k}'|.$$

$$q^2 = |\vec{k} - \vec{k}'|^2 = k^2 - k'^2 - 2kk' \cos\theta = 4k^2 \sin^2\left(\frac{\theta}{2}\right)$$

where $\theta =$ angle between \vec{k} and \vec{k}' .

For an impurity concentration per unit volume of N_I

$$S(k, k') = \frac{2\pi}{\hbar} \frac{e^4 N_I \delta(E(k) - E(k'))}{(\text{Vol})(\epsilon_0 \epsilon_r)^2 \left(4k^2 \sin^2\left(\frac{\theta}{2}\right) + q_s^2 / \epsilon_r\right)^2} \quad (4.70d)$$

Since ionized impurity scattering is elastic, $|\vec{k}| = |\vec{k}'|$. From the denominator in (Eq. 4.70d), we see that the scattering is anisotropic and that it favors forward scattering ($\theta = 0$). This is the expression for the Brooks–Herring model for screened ionized-impurity scattering.⁸

4.5.2 Average Relaxation Times

Once we have a scattering rate, we want to calculate appropriate averages over the entire ensemble of electrons. The key issues in this averaging procedure are described next. The total scattering rate is obtained by summing over all the states

that the electron can scatter to.

$$\frac{1}{\tau} = \sum_{k'} S(k, k') \quad (4.71a)$$

However, this by itself is not very physically meaningful because all scattering events are clearly not the same in terms of how much change they involve in physical quantities such as the momentum (k) or energy [$E(k)$] of the electrons. Hence, in Eq. (4.71a) the sum must obviously be weighted by the relative change of the physical quantity that we are interested in. For example, if we are interested in the relaxation of the electron momenta, we weight the Eq. (4.71a) sum by the relative change of $k = [1 - (k' \cos \theta)]/k$. Performing the integration in k -space using spherical coordinates, we calculate the momentum relaxation time as

$$\begin{aligned} \frac{1}{\tau_m} &= \sum_{k'} S(\vec{k}, \vec{k}') \left(1 - \frac{k' \cos \theta}{k}\right) \\ &= \frac{\text{Vol}}{(2\pi)^3} \int_0^{2\pi} \int_{-1}^1 \int_0^\infty S(k, k') (1 - \cos \theta) k'^2 dk' d(\cos \theta) d\phi \end{aligned} \quad (4.71b)$$

Note that here we do not use a factor of 2 due to electron spin because there is no spin flip during ionized impurity scattering. The Brooks–Herring expression for momentum relaxation time for heavily screened ionized impurities then becomes

$$\tau_m(k) = \left\{ \frac{16\sqrt{2}m^*\pi(\epsilon_0\epsilon_r)^2}{N_I e^4} \left[\ln(1 + \gamma^2) - \frac{\gamma^2}{1 + \gamma^2} \right] E^{\frac{3}{2}}(k) \right\}$$

where

$$\gamma^2 = (2L_D)^2 k^2 = 8m^* E(k) \frac{L_D^2}{\hbar^2} = \frac{32\pi^2 m^* E(k)}{\hbar^2 q_s^2} \quad (4.71c)$$

We see from this expression that as $E(k)$ or temperature increases, the momentum relaxation time goes up, because there is less deflection of carriers by impurities.

Similarly, the energy relaxation time is defined as

$$\frac{1}{\tau_E} = \sum_{k'} S(\vec{k}, \vec{k}') \left(1 - \frac{E(k')}{E(k)}\right) \quad (4.72)$$

However, since ionized impurity scattering is elastic, the relative change of the carrier energy is zero and, therefore, $\tau_E = \infty$. The same sort of averaging will be done for all the other scattering mechanisms.

Note, however, that even after summing over all the final states, k' , the expression for the momentum relaxation time is still a function of the wavevector, k . To get ensemble averages for *all* the electrons, we will need information about the distribution function of the electrons, $f(k)$, to obtain $\langle \tau_m \rangle$ or the ensemble momentum relaxation time, which will enter into the mobility expression for drift-diffusion-type models. We address these ensemble averages in the next section after we look at the solution of the BTE.

4.5.3 Conwell–Weisskopf Model

While the Brooks–Herring model works well for heavily screened ionized impurities, the expression for scattering rate is not valid for the lightly screened case. This is basically because the impact parameter, or the distance of closest approach between the electron and the ionized impurity can take any value from zero to infinity in the Brooks–Herring model.⁹ However, for the unscreened case, it makes more physical sense to account for not just isolated ionized impurities, but for the effect of neighboring impurities also.

Hence, we cut off the maximum impact parameter at $b_{\max} = 1/2(N_I)^{1/3}$, which is half the average spacing between the ionized impurities, at an impurity concentration of N_I . This gives us the Conwell–Weisskopf model:

$$\tau_m(k) = \frac{16\pi\sqrt{2m^*}(\epsilon_0\epsilon_r)^2}{N_I e^4} [\ln(1 + \gamma_{cw}^2)]^{-1} E(k)$$

where

$$\gamma_{cw} = 2 \left(\frac{4\pi(\epsilon_0\epsilon_r)b_{\max}}{q^2} \right) E(k) \quad (4.73)$$

4.5.4 Carrier–Carrier Scattering

For carrier concentrations above 10^{17} cm^{-3} (e.g., in the inversion layer of MOSFETs), the binary collisions between mobile carriers is very important. The model is very similar to that for ionized impurity scattering, except that now we have scattering between two comparable masses, and we have to keep track of two sets of wavevectors. The scattering is subject to momentum and energy conservation:

$$\begin{aligned} \vec{k}_1 + \vec{k}_2 &:= \vec{k}'_1 + \vec{k}'_2 \\ E(\vec{k}_1) + E(\vec{k}_2) &:= E(\vec{k}'_1) + E(\vec{k}'_2) \end{aligned}$$

In exact analogy to the single-carrier case, we now define a *pair transition rate*:

$$S(\vec{k}_1, \vec{k}_2; \vec{k}'_1, \vec{k}'_2)$$

In the center-of-mass reference, this resembles ionized impurity scattering, and we get

$$S(\vec{k}_1, \vec{k}_2; \vec{k}'_1, \vec{k}'_2) = \frac{2\pi e^4}{\hbar(\epsilon_0 \epsilon_r)^2 \text{Vol}} \left(\frac{1}{|k_1 - \vec{k}'_1|^2 + \frac{1}{L_D^2}} \right) \quad (4.74a)$$

$$\delta(k_1 + k_2 - k'_1 - k'_2) \cdot \delta(E(k_1) + E(k_2) - E(k'_1) - E(k'_2))$$

$$\frac{1}{\tau(k_1)} = \sum_{k_2} \sum_{k'_1} \{S(\vec{k}_1, \vec{k}_2; \vec{k}'_1, \vec{k}'_2) f(k_2) \cdot (1 - f(k'_1))(1 - f(k'_2))\} \quad (4.74b)$$

Note that from the conservation of momentum, k'_2 , is fixed once the other three wavevectors are chosen, so that we do not have to sum over k'_2 . Unlike ionized impurity scattering, carrier-carrier scattering is not elastic. Also, one needs to put in carrier occupancy factors to ensure that the initial states are full and the final states are empty (so that the Pauli principle is not violated). These carriers can also set up plasma oscillations (plasmons) in the free carrier gas and cause additional energy-loss mechanisms.

4.5.5 Phonon Scattering

As mentioned above, the atoms in a solid are not static, but execute more or less simple harmonic oscillations about the equilibrium positions.³ The amplitude of these oscillations increases with temperature. One can illustrate the basic properties of the resulting elastic waves in the crystal (whose quanta are termed phonons) by assuming nearest-neighbor interactions in a linear monoatomic chain of atoms. For small displacements, one assumes Hooke's law to be valid so that we get simple harmonic oscillations. Therefore, the restoring force becomes proportional to the relative displacement of the adjacent atoms. If the atomic locations are indexed by n , and the displacements by u , the force (= mass \times acceleration) becomes

$$F_n = m\ddot{u}_n = -\alpha(u_n - u_{n+1}) - \alpha(u_n - u_{n-1}) \quad (4.75a)$$

where the "spring constant" due to the interatomic forces is α .

We have traveling-elastic-wave solutions of the form

$$u_n = A e^{i(\omega t - qna)} \quad (4.75b)$$

There are four subsets of these phonons: acoustic (A) and optic (O) branches, which respectively refer to nearest-neighbor atoms moving in phase or out of phase; and for each of these branches the displacements of the atoms can be parallel to the direction of the phonon wavevector (longitudinal mode, L), or perpendicular to the wavevector (transverse mode, T). Thus one gets LA, TA, LO, and TO phonons. To determine the scattering rate of electrons by the time-dependent perturbation of

the phonons, one invokes the acoustic or optic phonon deformation potential theory of Bardeen and Shockley. The essence of the theory is that the phonon modes cause a local deformation of the lattice constant. Since the bandgap is a function of the interatomic spacing, there will thus be a ‘‘local’’ modulation of the conduction- and valence-band edges in a random fashion as the phonons propagate through the crystal, leading to a perturbing (scattering) Hamiltonian. In compound semiconductors, there can be additional scattering terms due the coulombic fields that are set up. This leads to polar optic-phonon scattering (for the optic modes) or piezoelectric scattering (for the acoustic modes). We will not treat these here because they are not relevant for Si-based ULSI devices.

One calculates the matrix element of this Hamiltonian, now between a final coupled electron–phonon state and an initial electron–phonon state, instead of just between electron states, as in the previous calculations. This is because electron–phonon interactions will not only change the electron wavevector and energy, but will also change the phonon occupancy of the phonon bath in the crystal. If the scattering involves phonon absorption, the final state will have one fewer phonon, while phonon emission will lead to an increase of the phonon occupancy number. The main difference between the deformation potentials for the acoustic and optic branches is that in the former, since adjacent atoms move in phase, the local deformation of the lattice is dependent on the divergence of the atomic displacement vectors. For the optic branch, where the nearest-neighbor atoms move against each other, out of phase, the local deformation depends on the displacement itself.

For the acoustic mode, using the Taylor series, we write the band-edge shifts with lattice deformation as

$$\Delta E_c = E_c(a) - E_c(a + \delta a) = \left(\frac{dE_c}{da} \right) (\delta a) \quad (4.76a)$$

or in three dimensions

$$\Delta E_c \propto \frac{\Delta \text{Vol}}{\text{Vol}} = Z_A \nabla \cdot \vec{u} \quad (4.76b)$$

where \vec{u} = lattice displacement. The proportionality constant is Z_A , the acoustic deformation potential constant (which depends on the semiconductor).¹⁰

Before we continue any further, we need to introduce the language of the quantum simple harmonic oscillator (SHO). We can quantize the simple harmonic oscillations as phonons, where instead of allowing any displacement amplitude (or classical elastic-wave energy) varying in a continuous fashion, we now allow only discrete, integer values of the phonon occupancy number, n_q , for phonons with wavevector q , as dictated by Bose–Einstein statistics and the phonon dispersion relationship. As temperature increases, instead of saying that the atomic oscillation amplitude increases, we now say that we have a greater *number* of phonons in the crystal. For the quantum SHO, one can describe the behavior in terms of the annihilation and creation operators, which add or remove a phonon from the system,

respectively, as follows:

$$a_q |n_q\rangle = \sqrt{n_q} |n_q - 1\rangle \quad (\text{phonon absorption}) \quad (4.77a)$$

$$a_q^+ |n_q\rangle = \sqrt{n_q + 1} |n_q + 1\rangle \quad (\text{phonon emission}) \quad (4.77b)$$

Using these operators, which are termed *second-quantized* operators in quantum field-theoretic notation, one can write the perturbation Hamiltonian.⁷ First writing the displacement in terms of the second-quantized operators

$$\vec{u}(r) = \sqrt{\frac{\hbar}{2NM\omega_q}} \hat{e} \cdot e^{i\vec{q}\cdot\vec{r}} [a_q + a_{-q}^+] \quad (4.78)$$

we get for the acoustic phonons

$$V(r) = Z_A \sqrt{\frac{1}{\omega_q N}} [a_q + a_{-q}^+] e^{i\vec{q}\cdot\vec{r}} (i\vec{q}) \cdot \hat{e} \quad (4.79)$$

From equations (4.76b) and (4.79) we notice that $V(r) \propto \nabla \cdot \vec{u} \propto \vec{q} \cdot \hat{e}$, where \hat{e} is the unit vector along the displacement of the atoms. This is important because it implies that for carriers in a spherical band (e.g., holes in Si) there cannot be any transverse acoustic deformation potential (TADP) scattering. As mentioned earlier, we need to calculate the matrix element using both the electron and the phonon wavefunctions. For example, the electronic contribution to the matrix element for phonon absorption is

$$M_{kk'} = \left[\sqrt{\frac{1}{N\omega_q}} \sqrt{n_q} Z_A |\vec{q}| (\hat{q} \cdot \hat{e}) \cdot \underbrace{\frac{1}{\text{Vol}} \int_{\text{Vol}} e^{-i\vec{k}'\cdot\vec{r}} e^{+i\vec{q}\cdot\vec{r}} e^{-i\vec{k}\cdot\vec{r}} d\vec{r}}_{\delta(\vec{k}' - \vec{k} - \vec{q})} \right] \quad (4.80a)$$

$$S(\vec{k}, \vec{k}') = \frac{2\pi}{\hbar} \left(\frac{Z_A^2}{N\omega_q} \right) (q)^2 [n_q \cdot \delta(E(k') - E(k) - \hbar\omega_q)] \quad (4.80b)$$

There will be similar contributions for phonon emission. The key difference is that the dependence on the phonons now is

$$S(k, k') \sim (n_q + 1) \delta(E(k') - E(k) + \hbar\omega_q) \times \delta(k' - k + q) \quad (4.80c)$$

Note that the final electron energy in this case is *higher* than the initial energy and the sign of the phonon wavevector is reversed. Finally, the rate depends on $(n_q + 1)$ instead of just n_q because we can have both stimulated ($\propto n_q$) and spontaneous

phonon emission ($\propto 1$), but only stimulated phonon absorption ($\propto n_q$). Acoustic deformation potential (ADP) scattering is almost elastic because the acoustic phonons have a more or less linear dispersion relationship with a slope equal to the sound velocity in the solid (c_ℓ), where, near $q = 0$ (gamma point), they have low energy.

We can calculate the momentum and energy relaxation times as before:

$$\frac{1}{\tau} = \sum_{k'} S(k, k') = \frac{2\pi Z_A^2}{\hbar c_\ell} k T_L \frac{1}{\text{Vol}} \sum_{k'} \delta(E(k') - E(k)) \quad (4.81)$$

The sum over k states can be converted into an integral over energy, E , using the density of states (DOS) exactly as before, except that we do not consider spin degeneracy because the electron spin does not flip during phonon scattering.

$$\frac{1}{\text{Vol}} \sum_{k'} \delta(E(k') - E(k)) = \frac{(2m^*)^{3/2} E(k)^{1/2}}{4\pi^2} = \frac{g(E)}{2} \quad \text{where } g(E) = \text{DOS} \quad (4.82)$$

$$\frac{1}{\tau_m} = \sum_{k'} S(k, k') (1 - \cos\theta) = \sum_{k'} S(k, k') = \frac{1}{\tau} \quad (4.83)$$

because longitudinal acoustic deformation potential (LADP) scattering is isotropic.

$$\frac{1}{\tau_m} = \frac{\pi Z_A^2 k T_L}{\hbar c_\ell} g(E) \quad (4.84)$$

Physically this makes sense, because we indeed expect the scattering rate to increase with the DOS where there are more final states to scatter to.¹⁰

The treatment for optical deformation potential (ODP) scattering is very similar, except that we must remember that now the adjacent atoms move in opposite phase. Therefore,

$$V(r, t) = Z_0 u(r, t) \quad (\text{instead of } \nabla \cdot \vec{u}) \quad (4.85)$$

where Z_0 is now the optic deformation potential constant. Then

$$M(k, k') = \sqrt{\frac{\hbar}{2NM\omega_q}} Z_0 \hat{e} \left[\underbrace{\sqrt{n_q}}_{\text{abs}} \text{ or } \underbrace{\sqrt{n_q + 1}}_{\text{emiss}} \right] \quad (4.86a)$$

$$S(k, k') = \frac{2\pi}{\hbar} \left(\frac{\hbar}{2NM\omega_q} \right) Z_0^2 [n_q \delta[E(k) - E(k') + \hbar\omega_q] + [(n_q + 1) \delta[E(k) - E(k') - \hbar\omega_q]]] \quad (4.86b)$$

Unlike that for the acoustic mode, the phonon dispersion relationship for the optic branch is $E(q) = \hbar\omega_0$ (constant). Since the optic phonons carry a lot more energy than acoustic phonons, ODP scattering is inelastic unlike ADP scattering, but it is still isotropic.

We compute the relaxation times as follows:

$$\frac{1}{\tau(k)} = \frac{1}{\tau_m(k)} = \frac{(2m^*)^{3/2} Z_0^2}{4\pi\hbar^3 \rho\omega_0} [(n_q + 1)[E(k) - \hbar\omega_0]^{1/2} + n_q[E(k) + \hbar\omega_0]^{1/2}] \quad (4.87a)$$

$$\frac{1}{\tau_m(k)} = \frac{\pi Z_0^2}{2\rho\omega_0} [(n_q + 1)g[E(k) - \hbar\omega_0] + n_q g[E(k) + \hbar\omega_0]] \quad (4.87b)$$

$$\frac{1}{\tau_E(k)} = \sum_{k', \text{ no spin}} S(k, k') \left[1 - \frac{E(k')}{E(k)} \right] \quad (4.87c)$$

For high carrier energies, $E(k)$, we have mainly emission of optic phonons and, therefore,

$$\frac{1}{\tau_E(k)} = \sum_{k'} S(k, k') \left[\frac{\hbar\omega_0}{E(k)} \right] = \frac{\hbar\omega_0}{E(k)} \left(\frac{1}{\tau_m(k)} \right) \quad (4.87d)$$

4.5.6 Intervalley Scattering

The final category of scattering mechanisms that we will discuss is a special type of phonon scattering that moves carriers from one $E(k)$ valley to another,^{7,10} called g type for scattering between parallel valleys and f type for scattering between perpendicular valleys. These involve a large change of the electron wavevector k , and hence, to satisfy momentum conservation, we need phonons close to the edge of the Brillouin Zone, where they have an energy of approximately $\hbar\omega_i$. Intervalley scattering is similar to ODP scattering in that it is isotropic and inelastic, except that one needs to account for the number of equivalent valleys in the final DOS:

$$\frac{1}{\tau} = \frac{1}{\tau_m} = \frac{\pi Z_i^2 N_{eq}}{2\rho\omega_i} (n_i g(E - \hbar\omega_i - \Delta) + (n_i + 1)g(E + \hbar\omega_i - \Delta)) \quad (4.88)$$

where Z_i is the coupling constant (deformation potential) for intervalley scattering, N_{eq} is the number of equivalent valleys to which it can scatter, n_i is the phonon occupancy, $\hbar\omega_i$ is the phonon energy, and Δ is the energy separation between valleys (if applicable).

We conclude this section by pointing out which scattering mechanisms are of the greatest importance in Si. For electrons in the conduction band, the most important mechanisms are

1. Acoustic phonons (LADP + TADP) at room temperature
2. Ionized impurity (above 10^{17}cm^{-3})
3. Carrier-carrier scattering (above 10^{17}cm^{-3})
4. Intervalley f and g scattering.

There is no ODP scattering of electrons due to the symmetry of the conduction band ellipsoids, as can be shown by group theoretic arguments.

For holes in the valence band, the key mechanisms are (1) ODP (because the valence band is a “warped” sphere), (2) ADP (especially for the heavy holes), (3) intersubband scattering, and (4) ionized impurity and carrier–carrier scattering at high concentrations.

In addition, depending on the device context, there can be other important mechanisms. As we shall see later, carriers in the inversion layer of MOSFETs undergo surface roughness scattering. For SiGe-based HBTs, one can have alloy scattering.

4.6 MONTE CARLO SIMULATIONS

In single-particle Monte Carlo (MC) simulations with full bandstructure, one determines the carrier trajectory according to the flowchart shown in Figure 4.2. First, one determines the carrier concentration and electric field profile from a solution of the Poisson equation and launches carriers from the contact into the device according to a Maxwellian (or hemi-Maxwellian) distribution. Using a pseudo-random number, R , generated on a computer, one determines a so-called free-flight time, t_f :

$$t_f = \frac{-\ln(1-R)}{S} \quad (4.89)$$

where S is the total scattering rate due to all scattering mechanisms.¹¹ A mathematical artifice that is used here is so-called *self-scattering*. This is an artificial scattering mechanism that is invoked that does nothing; specifically, it leaves the electron wavevector, k , and the energy, $E(k)$, unchanged after scattering. The reason this self-scattering is put in is to make the sum total of all the scattering mechanisms (real ones such as phonon and ionized impurity scattering, plus the self-scattering) constant, as a function of electron energy. In spite of the density-of-states and scattering rates varying as a function of energy, $E(k)$, a constant total scattering rate makes the numerical algorithm easier to implement in MC simulations. Hence, one can determine the free-flight time of the electrons as described. During the free-flight time, one uses the equations of semiclassical electron dynamics to update the time evolution of the wavevector, k . Looking up the corresponding $E(k)$ from the bandstructure, one can determine the corresponding velocity of the electron in real space. This can be used to update the position, r , of the electron. Periodically, the

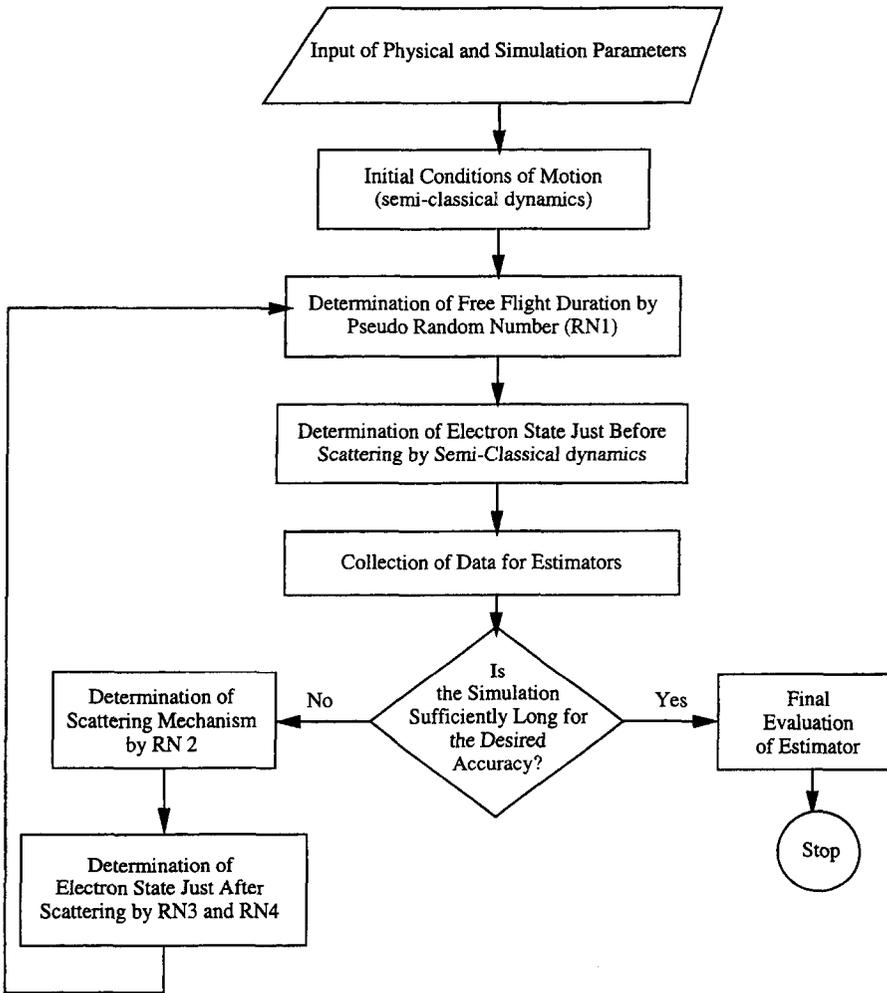


Figure 4.2 Flowchart of a Monte Carlo simulator.

electrons are forced to scatter and suddenly change their wavevector, k , and energy, $E(k)$, according to the various scattering mechanisms. This is done at different rates for the different scattering mechanisms, probabilistically using another pseudo-random number. Finally, two other random numbers are used to determine the two polar angles (in spherical coordinates) that determine the direction of the final wavevector after scattering.

One launches, typically, tens of thousands of electrons into a device, and keeps track of their trajectories and other parameters in the device. Finally, statistical averages of different physical quantities such as electron densities, currents, and

electron energies can be computed. MC simulations are extremely powerful because it is possible to incorporate more detailed bandstructure information (and consequently density-of-states information), and wavevector-dependent scattering rates than is possible analytically, or even with the solution of the BTE. One problem, however, with such brute-force numerical techniques is the numerical noise or scatter introduced. Generally speaking, the noise is inversely proportional to the square root of the number of particles simulated. There is, thus, a tradeoff between speed, which obviously goes down with increasing particle number, and accuracy, which improves with particle count.

A variation of single-particle MC is ensemble MC (or EMC), where instead of following the trajectories of electrons in real space and k space one at a time, one populates a device to start with by a large number of carriers.¹¹ Typically, the initial distribution of carriers and their velocities are guessed using a crude analytical or numerical solution such as the drift-diffusion model. Then, one looks at the temporal and spatial evolution of all the electrons in the ensemble according to the two fundamental equations of semiclassical electron dynamics. Although such EMC simulations are clearly computationally more demanding, they are being used more and more for realistic device simulations. They are more realistic than single-particle MC because one can now solve for the evolution of the ensemble self-consistently with the Poisson equation, instead of doing so in a decoupled manner, as in single-particle MC simulations. After every few time steps (depending on the scattering rates, and the spatial variation of the fields in the device), one puts in the updated spatial distribution of the electron ensemble into the Poisson equation to determine the new field to be used in the semiclassical electron dynamics to evolve the electrons further. Another advantage of EMC is that one can now handle scattering mechanisms such as carrier-carrier scattering, which cannot be done realistically in single particle MC.

4.6.1 HBT Simulations

We illustrate such EMC simulations for a heterojunction bipolar junction transistor (HBT).¹² The basic physics of such a device has been described in Chapter 2. It is possible to glean detailed information of minority-carrier transport across the base from such EMC simulations. Typical results from simulations by Hughes et al. for SiGe NPN HBTs are shown in Figure 4.3. For example, they show the effects of high collector currents on base transit times by explicitly accounting for the electron occupancies and velocities in the transverse and longitudinal conduction-band valleys, as a function of position in the base. These simulations show that increasing the collector current density from 1×10^7 A/m² to 5×10^8 A/m² causes a change in the electron concentration profile and electric field profile in the base. This leads to an increase of the base transit time due to a base widening effect and a reduction of the base collector depletion-width transit time due to an extended collector field, or the so-called inverted collector structure. In another set of simulations for a graded bandgap, Hughes et al. were able to determine

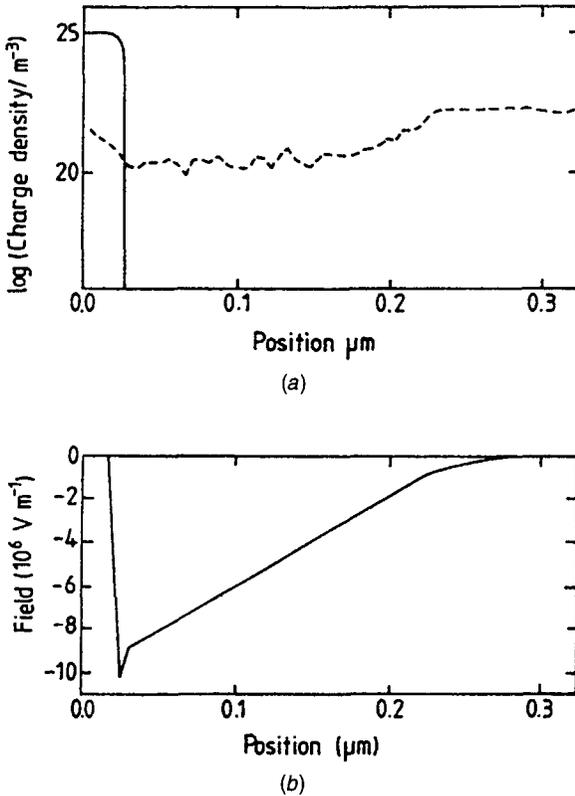


Figure 4.3 Ensemble Monte Carlo simulations of a $\text{Si}_{0.7}\text{Ge}_{0.3}$ heterojunction bipolar transistor from Hughes et al.¹² showing the details of the field, populations of the electrons (solid lines), holes (dashed lines), in the longitudinal and transverse valleys, and the velocities of those carriers for low and high current levels, all plotted as functions of position in the base. Panels *a–d* correspond to a current density of 10^3 A/cm^2 , while panels *e–g* correspond to a current density of $5 \times 10^4 \text{ A/cm}^2$. (With permission from IEEE).

the optimal Ge grading to minimize the base transit time. The point is that detailed computations of this nature can only be done numerically using MC techniques.

4.7 BOLTZMANN TRANSPORT EQUATION

Instead of looking at the individual carriers as in MC simulations, we can consider transport from a collective viewpoint. To derive ensemble-average transport properties in a device, we need

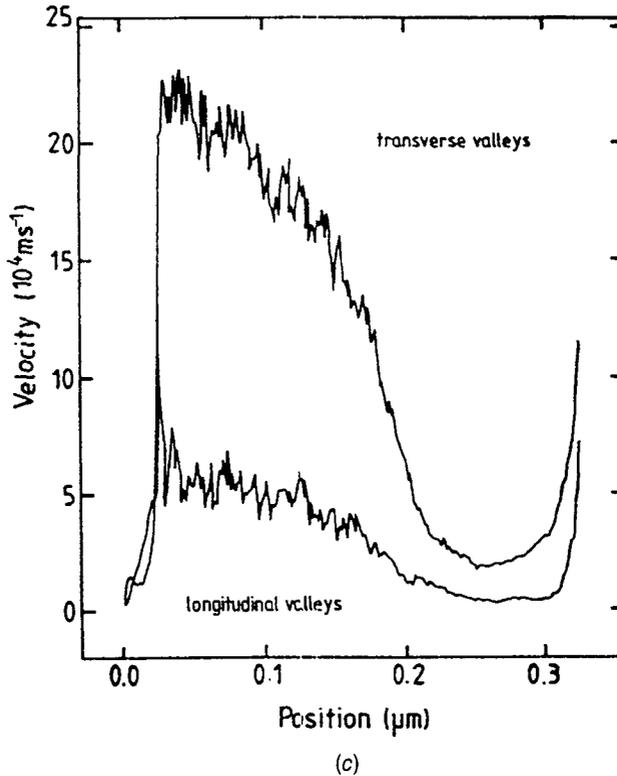


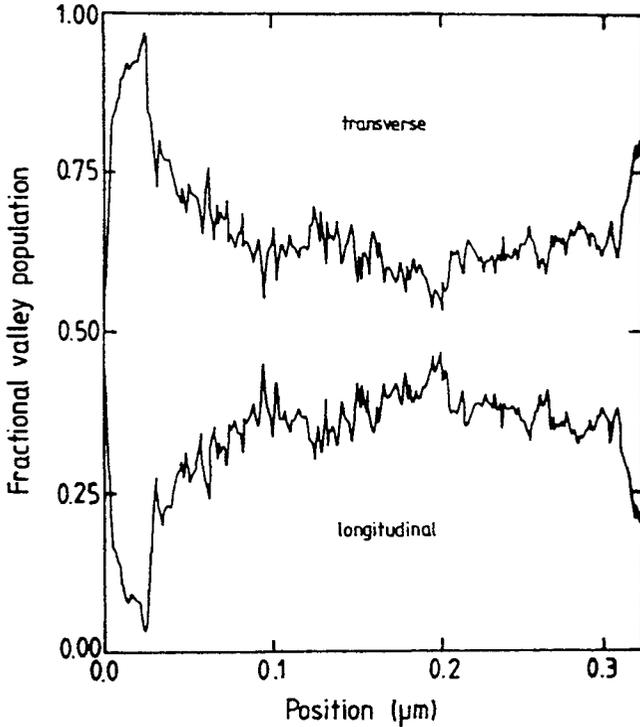
Figure 4.3 (Continued)

1. Information about single-particle behavior, which we get from semiclassical dynamics
2. Carrier statistics to average over

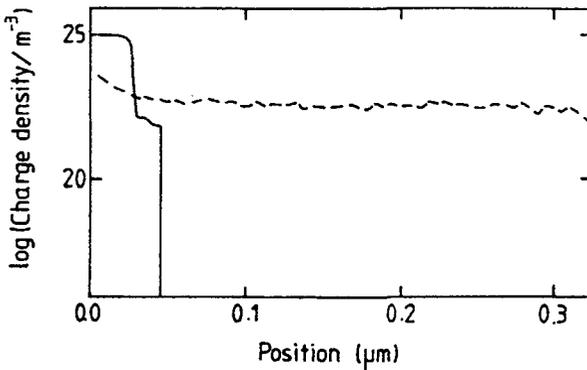
The distribution function $f(\vec{r}, \vec{k}, t) d\vec{r} d\vec{k}$ (Fermi–Dirac or its classical approximation, Maxwell–Boltzmann) is a function in six-dimensional (6D) phase space, the three real-space coordinates, the three momentum coordinates, and time. This function gives us the time-dependent probability of finding an electron or (hole) between

$$\begin{array}{lll} \vec{r} \text{ and } \vec{r} + d\vec{r} & \text{at} & (x, y, z) \\ \vec{k} \text{ and } \vec{k} + d\vec{k} & \text{at} & (k_x, k_y, k_z) \text{ or } (\vec{v} \text{ or } \vec{p}). \end{array}$$

It may be noted that this is a single-particle distribution function, which ignores (except for self-consistent potentials) the true N -particle, distribution function in the device. In that sense, this is like replacing the correct, many-body Hamiltonian by a single-electron, effective-mass equation. The spatial and temporal evolution of the carrier distribution function is obtained from the BTE.



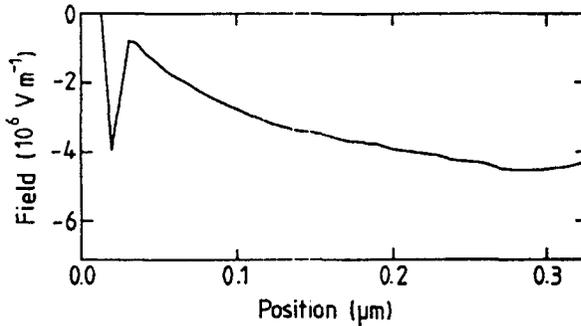
(d)



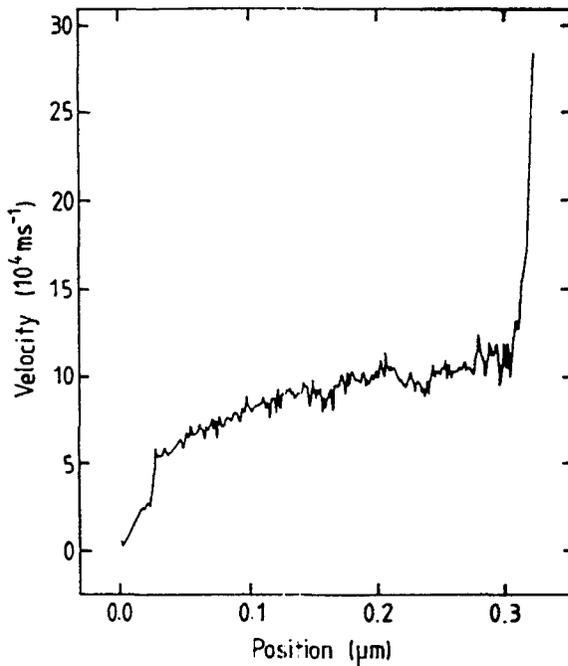
(e)

Figure 4.3 (Continued)

The derivation of the BTE basically involves application of the equation of continuity to a “probability fluid”, f , in (\vec{r}, \vec{k}) 6D phase space, keeping the two fundamental equations of semiclassical dynamics in mind to track the evolution of \vec{r} and \vec{k} .¹³ Thus, to obtain the temporal change of f in an elemental cube in real space, one first considers flow along the x direction. The change of probability f between



(f)



(g)

Figure 4.3 (Continued)

times t and $(t + dt)$ is the difference of the probability fluid flowing “in” minus the flow “out.” This is given by

$$\begin{aligned}
 & \int f(x, y, z, \vec{k}, t) d\vec{k} dy dz (\dot{x}dt) - \int f(x + dx, y, z, \vec{k}, t) d\vec{k} dy dz (\dot{x}dt) \\
 & \quad \text{(flow “in”)} \qquad \qquad \qquad \text{(flow “out”)} \\
 & = -\dot{x} \left[\frac{\partial f}{\partial x} \cdot dx \right] dy dz d\vec{k} dt \qquad \qquad \qquad (4.90a)
 \end{aligned}$$

Extending this to three dimensions, we get the total temporal change of f due to flow in real space:

$$= -\dot{\vec{r}} \cdot \nabla_r f \, d\vec{r} \, d\vec{k} \, dt \quad (4.90b)$$

Similarly, the temporal change of f due to flow in \vec{k} space

$$= -\dot{\vec{k}} \cdot \nabla_k f \, d\vec{r} \, d\vec{k} \, dt \quad (4.90c)$$

In addition to such gradual changes, there can be drastic changes of f due to scattering:

$$\begin{aligned} &= \sum_{k'}^{\text{in}} S(k', k) f(rk't) (1 - f(rkt)) \, d\vec{r} \, d\vec{k} \, dt \\ &- \sum_{k'}^{\text{out}} S(k, k') f(rkt) (1 - f(r\vec{k}'t)) \, d\vec{r} \, d\vec{k} \, dt \end{aligned} \quad (4.90d)$$

Note that here we make the *ansatz* that we can change \vec{k} instantly, but not \vec{r} . An instantaneous change of r would be tantamount to an infinite electron velocity, which is unphysical. However, an instantaneous change of k is also, strictly speaking, unphysical because it implies an infinite force. In the final section, we shall return to this approximation again, where we shall see that one ought to consider a gradual evolution of k during the scattering in a “collision sphere,” which gives rise to the intracollisional field effect (ICFE). For low carrier concentrations, the $(1 - f)$ terms are often considered $= 1$. Hence, considering all three terms, the BTE becomes

$$\frac{\partial f}{\partial t} = -\dot{\vec{r}} \cdot \nabla_r f - \dot{\vec{k}} \cdot \nabla_k f + \left. \frac{\partial f}{\partial t} \right|_{\text{coll}} \quad (4.91)$$

Here, $\dot{\vec{r}}$, or the group velocity, and $\dot{\vec{k}}$ are given by the two equations of semiclassical electron dynamics (Eq. 4.30). Device physics solutions in the collective picture using statistical distribution functions entail solving the BTE, self-consistently with the Poisson equation (Fig. 4.4). However, this partial-differential equation can be quite intractable for realistic boundary and initial conditions.^{13,14}

Hence, we next examine the solutions of the BTE under various simplifying conditions.⁷ We have seen that the distribution function in equilibrium is given by a Fermi–Dirac distribution of the form

$$f_{\text{eq}} = \frac{1}{e^{(E-E_F)/kT_L} + 1} \quad \text{where for parabolic bands} \quad E = E_c + \frac{\hbar^2 k^2}{2m^*} \quad (4.92a)$$

Here, T_L is the lattice temperature. With a bias applied, we find that

$$f_0 = \frac{1}{e^{(E-E_{F_n})/kT_c} + 1} \quad \text{where the carrier temperature} \quad T_c \geq T_L, \quad E_{F_n} \neq E_F \quad (4.92b)$$

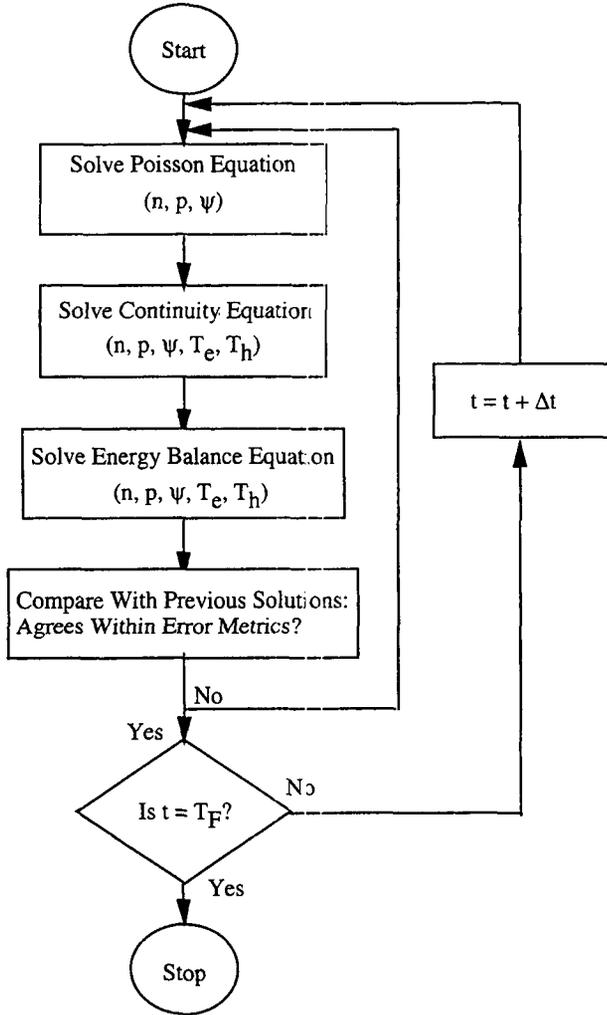


Figure 4.4 Flowchart of collective transport models based on the self-consistent solution of moments of the Boltzmann transport equation and the Poisson equation. This can be done either in a coupled (Newton scheme) or decoupled way (Gummel scheme).

However, although this is a good approximation, it cannot be strictly correct because this is still an even function of k , which cannot be responsible for net carrier transport. (The contribution to the current for positive values of k in the Fermi sphere would cancel with the contributions for negative k .) Hence, for transport, we must have a displaced Fermi sphere with an “odd” part of the distribution function $f_1(k)$, as follows:

$$\begin{aligned}
 f &= f_0(k) + f_1(k) \\
 f_0(k) &= f_0(-k); \quad f_1(k) = -f_1(-k)
 \end{aligned}
 \tag{4.93}$$

In equilibrium we have detailed balance, or

$$\frac{\partial f_{\text{eq}}}{\partial t} = 0 = [f_{\text{eq}}(k')S(k', k) - f_{\text{eq}}(k)S(k, k')] \quad (4.94)$$

In nonequilibrium with an electric field ε , the temporal change of the distribution function is written as partially due to the even part, and partially due to the odd part.

$$\left. \frac{\partial f}{\partial t} \right|_{\text{coll}} = \sum_{k'} \underbrace{[f_0(k')S(k', k) - f_0(k)S(k, k')]}_{\left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}} \propto \varepsilon^2} - f_1(k) \underbrace{\sum_{k'} S(k, k')}_{\frac{1}{\tau_f}} \quad (4.95a)$$

One can then invoke the RTA and, for low fields where the first term is negligible, write

$$\left. \frac{\partial f}{\partial t} \right|_{\text{coll}} \cong -\frac{f_1}{\tau_f} = -\frac{(f - f_{\text{eq}})}{\tau_f} \quad (4.95b)$$

Solving this, we get

$$f = f_{\text{eq}} + (f - f_{\text{eq}})e^{-(t/\tau_f)} \quad (4.96)$$

The τ_f is determined by momentum, k , randomization during scattering or collisions:

$$\frac{1}{\tau_f} = \sum_{k'} S(k, k') \left(1 - \frac{k'}{k}\right) \quad (4.97a)$$

where term $1 - (k'/k)$ is the relative change of k .

If the angle between k and $k' = \theta$, and the collisions are elastic, we obtain

$$\frac{k - k'}{k} = (1 - \cos\theta) \quad (4.97b)$$

$$\frac{1}{\tau_f} = \sum_{k'} S(k, k')(1 - \cos\theta) \quad (4.97c)$$

On the other hand, if the scattering mechanism is isotropic, then

$$\frac{1}{\tau_f} = \sum_{k'} S(k, k') = \frac{1}{\tau} \quad (4.97d)$$

Thus the RTA is valid only for low fields $[(\partial f_0/\partial t) \propto \varepsilon^2]$ and for scattering mechanisms that are elastic or isotropic (or both).

A useful, approximate solution of the BTE is for steady state, with no (or slow) spatial variations. Splitting up f into an even and an odd part, we then get

$$\frac{e\vec{\epsilon}}{\hbar} \left(\underbrace{\nabla_k f_0}_{\text{odd}} + \underbrace{\nabla_k f_1}_{\text{even}} \right) = \underbrace{-\frac{f_1}{\tau_1}}_{\text{odd}} + \underbrace{\frac{\partial f_0}{\partial t}}_{\text{even}} \Big|_{\text{coll}} \quad (4.98a)$$

$$\frac{e\vec{\epsilon}}{\hbar} \cdot \nabla_k f_0 = -\frac{f_1}{\tau_1}; \quad \frac{e\epsilon}{\hbar} \nabla_k f_1 = \frac{\partial f_0}{\partial t} \Big|_{\text{coll}} \quad (4.98b)$$

The starting f_{eq} in the Maxwell–Boltzmann limit is

$$= \frac{1}{1 + \exp\left\{ \frac{E_c + \frac{\hbar^2 k^2}{2m^*} - E_F}{kT_L} \right\}} \cong \exp\left\{ \left[\frac{E_F - E_C}{kT_L} \right] - \frac{\hbar^2 k^2}{2m^* kT_L} \right\} \quad (4.99)$$

With bias, solving the BTE, we get a displaced Maxwellian or drifted Maxwellian:

$$f(r, k) = \exp\left\{ \left[\frac{E_{F_n} - E_c}{kT_c} \right] - \frac{\hbar^2 (k - k_d)^2}{2m^* kT_c} \right\} \quad (4.100)$$

where $k_d \equiv$ drift velocity.

4.7.1 Average Quantities

Once the distribution function is solved, we can use it to calculate various averages corresponding to different physical quantities.^{7,14} The recipe is detailed below. It involves expressing the physical quantity as a function, $Q(k)$. Then

$$\langle Q(r, t) \rangle = \frac{1}{\text{Vol}} \sum_k Q(k) f(r, k, t) \quad (4.101a)$$

or

$$\langle Q(r, t) \rangle = \frac{\int_{-\infty}^{\infty} Q(k) f(r, k, t) dk}{\int_{-\infty}^{\infty} f(r, k, t) dk} \quad (4.101b)$$

This is very useful because, in general, the solution of the BTE gives us more information than we can digest (or need) under most conditions. It gives us detailed information about the carrier velocities at each point in the semiconductor. For different choices of this function $Q(k)$, we get expectation values of different physical variables. We illustrate this for three important cases:

Case 1 Carrier Concentrations: $Q = 1$

$$\langle Q \rangle = n(\vec{r}, t) = \frac{1}{\text{Vol}} \sum_k f(\vec{r}, \vec{k}, t) \quad (4.102a)$$

Earlier, while discussing equilibrium statistics, we showed that $\sum_k = \text{Vol}/4\pi^3 \int d\vec{k}$, which can be converted into an integration with respect to energy, E , by introducing the DOS. We then get

$$\begin{aligned} n(\vec{r}, t) &= \frac{1}{\text{Vol}} \left(\frac{\text{Vol}}{4\pi^3} \right) \int \exp \left(\frac{E_F - E_C - \frac{\hbar^2 k^2}{2m^*}}{kT_L} \right) d\vec{k} \\ &= \frac{1}{4} \left(\frac{2m^* kT_L}{\pi \hbar^2} \right)^{3/2} e^{(E_F - E_C)/kT_L} \\ &= N_c e^{(E_F - E_C)/kT_L} \end{aligned} \quad (4.102b)$$

It is reassuring to note that we recover the expressions for carrier concentration derived earlier (e.g., Eq. 4.54). It may be noted that by integrating with respect to k , we have thrown out the detailed information of the carrier velocities at various positions and times; but that is generally acceptable for most device physics problems.

Case 2 Carrier Kinetic Energy; $KE = Q = \hbar^2 k^2 / 2m^*$

In equilibrium,

$$KE = \frac{1}{\text{Vol}} \sum_k \frac{\hbar^2 k^2}{2m^*} f(r, k, t) = \frac{1}{8\pi^3 m^*} (\hbar^2) \int k^2 \exp \left[\frac{E_F - E_C - \frac{\hbar^2 k^2}{2m^*}}{kT_L} \right] d\vec{k} \quad (4.103a)$$

and we get, as predicted by the equipartition theorem,

$$\langle \text{KE per carrier} \rangle = \frac{KE}{n} = \frac{3}{2} kT_L$$

On the other hand, under bias, if we use the displaced Maxwellian, we get

$$\langle \text{KE per carrier} \rangle = \underbrace{\frac{\hbar^2 k_d^2}{2m^*}}_{\text{drift}} + \underbrace{\frac{3}{2} kT_c}_{\text{random}} \quad (4.103b)$$

Case 3 Current Density: $Q = \hbar k_z/m^*$

As mentioned before, the current flow can be due only to the odd part of the distribution function:¹⁴

$$\vec{J}(r, t) = \frac{-e}{4\pi^3} \int \vec{v}(f_0 + f_1) d\vec{k} \tag{4.104a}$$

$$\vec{J} = \frac{-e}{4\pi^3} \int \underbrace{\vec{v} f_1}_{\text{even}} d\vec{k} \tag{4.104b}$$

We can then introduce a conductivity tensor given by

$$\sigma_{ij} = \frac{j_i}{\epsilon_j} = \frac{-e^2}{4\pi^3} \int \tau_f(k) \frac{\partial f_0}{\partial E} v_i v_j d\vec{k} \tag{4.105a}$$

which can be diagonalized. For spherically symmetrical bands, we end up with a scalar

$$en\mu = \frac{-e^2}{4\pi^3} \int \tau_f(k) \frac{\partial f_0}{\partial E} (v_i)^2 d\vec{k}$$

where

$$\mu = \frac{-e \int \tau_f(k) \frac{\partial f_0}{\partial E} (v_i)^2 d\vec{k}}{4\pi^3 \underbrace{\frac{1}{4\pi^3} \int f_0 d\vec{k}}_{=n}} \tag{4.105b}$$

$$\mu = \frac{e\langle\tau\rangle}{m^*} = \frac{\frac{2}{3} e \int_0^\infty \tau(E) e^{-(E/kT_c)} \left(\frac{E}{m^*kT_c}\right) g(E) dE}{\int_0^\infty e^{-(E/kT_c)} g(E) dE} \tag{4.105c}$$

For some scattering mechanisms, the momentum relaxation time has an approximate power-law dependence of the form

$$\tau(k) = \tau \left(\frac{E}{kT_c}\right)^p \tag{4.106a}$$

Then, it can be shown from the preceding integral, that the ensemble momentum relaxation time for the electron gas is given by gamma functions:

$$\langle\tau\rangle = \frac{\tau_0 \Gamma\left(p + \frac{5}{2}\right)}{\Gamma\left(\frac{5}{2}\right)}. \tag{4.106b}$$

Ensemble Relaxation Times

We can generalize these concepts of ensemble averaging as shown below. For any physical quantity, $Q(k)$, the relaxation rate can be written as

$$R_Q = \frac{Q - Q^{eq}}{\langle \tau_Q \rangle} = \frac{1}{4\pi^3} \int \frac{f(k)Q(k)}{\tau_Q(k)} d\vec{k} \quad (4.107)$$

For example, the momentum relaxation rate for the entire ensemble of electrons with a well-collimated velocity distribution then is given in terms of $\langle \tau_m \rangle$ as

$$\frac{d(nm^* \vec{V}_d)}{dt} = - \left[\frac{(nm^* \vec{V}_d) - 0}{\langle \tau_m \rangle} \right] \quad (4.108a)$$

Earlier we discussed how one first averages over k' , to get

$$\frac{1}{\tau_m(k)} = \sum_{k'} S(k, k') \left(1 - \frac{k' \cos \theta}{k} \right) \quad (4.108b)$$

The ensemble relaxation time then involves a second averaging over the distribution function.

Similarly, for energy relaxation for the ensemble of “hot” carriers, we get $\langle \tau_E \rangle$, by first calculating

$$\frac{1}{\tau_E} = \sum_{k'} S(k, k') \left(1 - \frac{E(k')}{E(k)} \right) \quad (4.109a)$$

Then

$$\frac{d\langle E \rangle}{dt} = - \frac{[\langle E \rangle - E^{eq}]}{\langle \tau_E \rangle} = - \frac{\left(\frac{3}{2} kT_c - \frac{3}{2} kT_L \right)}{\langle \tau_E \rangle} \quad (4.109b)$$

4.7.2 Balance Equations and Method of Moments

As mentioned before, a self-consistent solution of the BTE and the Poisson equation is difficult. Also, the detailed k -space information is often redundant. Hence, one can take averages or scalar “moments” of the various terms of the BTE using the technique mentioned in Eq. 4.101 and convert it to a more tractable set of coupled “balance equations.”¹⁴ These form the basis of the equations used in conventional device analysis.

The recipe is to multiply each term of the BTE by $Q(k)$ and integrate with respect to $(1/4\pi^3) \int d\vec{k}$. For different powers of k , $Q(k)$, we get the different (hydrodynamic) balance equations, which have more familiar names:

0th moment, $k^0 : Q(k) = 1$ is the *continuity equation* (particle balance)

1st moment, $k^1 : Q(k) = \hbar k_z / m^*$ is the *drift–diffusion equation* (momentum balance)

2nd moment, $k^2 : Q(k) = \hbar^2 k^2 / 2m^*$ is the *energy-balance equation* (kinetic-energy balance).

The BTE using the relaxation-time approximation is

$$\frac{\partial f}{\partial t} + \underbrace{\frac{1}{\hbar} \nabla_k E \cdot \nabla_r f}_{i=v} - \underbrace{\frac{e\bar{\varepsilon}}{\hbar} \cdot \nabla_k f}_{\kappa} = \left. \frac{\partial f}{\partial t} \right|_{\text{coll}} - \frac{f_1}{\tau(k)} \quad (4.110a)$$

We multiply each term by $Q(k)$ and integrate over k . For parabolic bands, the BTE is transformed into

$$\frac{\partial}{\partial t} (\langle Q \rangle n) + \frac{\hbar}{m^*} \nabla_r \cdot (\langle Q k \rangle n) + \frac{e\bar{\varepsilon}}{\hbar} n \langle \nabla_k Q \rangle = \frac{1}{4\pi^3} \int Q \left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}} dk - \frac{1}{4\pi^3} \int \frac{f_1}{\tau} Q dk \quad (4.110b)$$

Case 1 Zeroth Moment The BTE gives us the continuity equations for electrons and holes:^{1,2}

$$\frac{\partial n}{\partial t} = \frac{1}{e} \nabla \cdot J_n + G(n) - R(n) \quad (4.111a)$$

$$\frac{\partial p}{\partial t} = -\frac{1}{e} \nabla \cdot J_p + G(p) - R(p) \quad (4.111b)$$

where G and R refer to the generation and recombination rates of carriers. The generation–recombination process can be optical or thermal (i.e., infrared photons), and can be band-to-band, or involve a trap [Shockley–Read–Hall (SRH) process]. For a SRH process, using rate equations and the principle of detailed balance, one can write the net electron recombination, for example, as

$$U_s = R(n) - G(n) = \frac{c_n c_p N_{TT} n p - e_n e_p N_{TT}}{c_n n + e_n + c_p p + e_p} \quad (4.112a)$$

$$U_s = (c_n c_p N_{TT}) \frac{np - n_i^2}{c_n n + e_n + c_p p + e_p} \quad (4.112b)$$

Here, N_{TT} represents the trap concentration and the c and e are the capture and emission coefficients, respectively. In p-type semiconductors, for instance, this simplifies to,

$$\begin{aligned} n &= n_0 + \delta n \\ p &= p_0 + \delta p \cong p_0 \\ U_s &= \bar{c} p_0 \delta n = \frac{\delta n}{\tau_n} \end{aligned}$$

Here, τ_n is the electron recombination (or generation) lifetime depending on whether we have an excess (or deficit) of carriers.

Case 2 First Moment The BTE gives us the drift–diffusion equations:^{1,2}

$$\tau \frac{\partial j}{\partial t} + j = en \frac{e\langle\tau\rangle}{m^*} \varepsilon + e \nabla_r \left(\underbrace{\frac{e\langle\tau\rangle k T_c}{m^* e}}_D n \right) \quad (4.113a)$$

$$= \sigma \varepsilon + e \nabla_r D n \quad (4.113b)$$

(drift) + (diffusion)

The current density in steady state becomes

$$J_n = e \mu_n n \varepsilon + q D_n \nabla_r n \quad (4.113c)$$

$$J_p = e \mu_p p \varepsilon - q D_p \nabla_r p \quad (4.113d)$$

Here, we have used the Einstein relationship to relate the diffusivity to the mobility. These are the Shockley equations, which, along with Maxwell equations below, form the basis of conventional device analysis. Of the Maxwell equations, the Poisson equation, shown last, is the most important for dc or low-frequency analysis:

$$\nabla \times \vec{\varepsilon} = - \frac{\partial \vec{B}}{\partial t} \quad (4.114a)$$

$$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}_{\text{cond.}} \quad (4.114b)$$

$$\nabla \cdot \vec{B} = 0 \quad (4.114c)$$

$$\nabla \cdot \vec{D} = \rho \quad (4.114d)$$

Case 3 Second Moment For high field transport, we need to consider the next-higher moment of the BTE. We then get the following for parabolic bands:

$$\begin{aligned} & \frac{\partial}{\partial t} (n \langle E(k) \rangle) + \underbrace{\frac{\hbar}{m^*} \nabla_r \{ n \langle E(k) k \rangle \}}_{\text{energy flux}} + \underbrace{\frac{e \vec{\varepsilon}}{\hbar} \cdot n \langle \langle \nabla_k E(k) \rangle \rangle}_{\text{average velocity}} \\ &= \frac{1}{4\pi^3} \int \underbrace{d\vec{k} E(k) \left[\frac{\partial f_0}{\partial t} \right]_{\text{coll}}}_{\text{energy loss to lattice}} - \frac{1}{4\pi^3} \int d\vec{k} \frac{f_1}{\tau(k)} E(k) \end{aligned} \quad (4.115a)$$

Here, the second term gives the “flux” or flow of electron energy.

$$\langle E(k)k \rangle = \frac{1}{n} \int \frac{\hbar^2 k^2}{2m^*} \vec{k} (f_0 + f_1) dk = \frac{1}{n} \frac{\hbar^2}{2m^*} \int k^2 \vec{k} f_1 dk = \frac{1}{n} \int E(k) k f_1 dk \quad (4.115b)$$

The collision term on the right-hand side of the BTE gives us

$$\frac{1}{4\pi^3} \int dk E(k) \left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}} = \underbrace{-nB(T_c)}_{\text{energy loss to lattice}} \quad (4.115c)$$

Thus the BTE yields the energy-balance equation¹⁴

$$\frac{\partial}{\partial t} (n\langle E \rangle) + \underbrace{\frac{\hbar}{m^*} \nabla_r \cdot \left(\int E(k) \vec{k} f_1 dk \right)}_{\nabla_r \cdot \vec{S}(T_c)} + \underbrace{e\vec{\epsilon} \cdot (n\langle \vec{v} \rangle)}_{-\vec{j} \cdot \vec{\epsilon}} = -nB(T_c) \quad (4.115d)$$

where $S(T_c) = (\hbar/m^*) \cdot (\int E(k) \vec{k} f_1) dk =$ energy flux. Using the RTA for the energy loss to the lattice, we can rewrite the energy-balance equation more compactly as

$$\underbrace{\frac{\partial}{\partial t} (n\langle E \rangle)}_{\text{rate of energy increase}} = \underbrace{\frac{\vec{j} \cdot \vec{\epsilon}}{n}}_{\text{power input}} - \underbrace{\frac{1}{n} (\nabla_r \cdot \vec{S}(T_c))}_{\text{energy flux}} - \underbrace{\left(\frac{E - E^{eq}}{\tau_E} \right)}_{\text{loss to lattice}} \quad (4.115e)$$

$$S(T_c) = \left\{ \left(\begin{array}{cc} \text{drift KE} & + & \text{random KE} \\ \frac{1}{2} m v_d^2 & & \frac{1}{2} m v_{th}^2 = \frac{3}{2} k T_c \end{array} \right) \right\} v_d$$

$$= \{ \text{velocity} \times \text{pressure of electron gas} \} + \text{heat flow} (= -\kappa \nabla_r T_c) \quad (4.115f)$$

If one examines the sequence of balance equations, one finds that each involves the next-higher power of k . One terminates this infinite series of balance equations by making the heuristic assumption that the Wiedemann–Franz law can be invoked for these electrons and the heat flow of the electron gas, which is the final term in the energy balance equation, can be approximated in terms of the (electronic) thermal conductivity, κ .

For large-geometry devices, where the internal electric fields are small, and, thus, the carrier heating effects are insignificant, one only needs to consider up to the first moment, and a drift–diffusion model suffices. However, as ULSI devices have been miniaturized, the internal electric fields have gone up, and the spatial variations of the fields have become large, so that the energy balance equation must also be solved.

Even if the fields are high, if the spatial variations of the fields are small, the hot electron effects can be approximated as follows.¹⁵ For slowly varying fields, the spatial gradient of the energy flux term can be ignored, so that

$$\frac{\partial}{\partial t} \langle E \rangle = \frac{1}{n} \vec{\varepsilon} \cdot \vec{j} - \underbrace{\frac{\partial}{\partial t} \langle E \rangle}_{\text{energy loss to lattice}} \Big|_{\text{coll}} \quad (4.115g)$$

$$\text{energy loss to lattice} = \frac{1}{8\pi^3} \int E \left\{ \frac{\partial f_0}{\partial t} \Big|_{\text{coll}} \right\} d\vec{k}$$

For Boltzmann statistics and small electric fields ε , we obtain

$$\frac{\partial}{\partial t} \langle E \rangle \Big|_{\text{coll}} = \left(\frac{\langle E \rangle - \frac{3}{2} kT_L}{\tau_E} \right) \quad (4.115h)$$

where $\langle E \rangle = \frac{3}{2} kT_c$. As we pump energy into an electron gas from an external field $\vec{\varepsilon}$, the carrier temperature, T_c is higher than the lattice temperature, T_L . However, the key point here is that the carrier heating now depends *only* on the *local* electric field, and not on the history of the carrier motion. Hence, we can approximately account for carrier heating effects by assuming that the carriers equilibrate with respect to the local electric fields, and by introducing field-dependent transport parameters such as carrier mobility, in the framework of a drift-diffusion model. This approach has been used for a long time to model ULSI devices, and is often still useful. It is only when there are rapid spatial and temporal variations of the electric fields that the history of the carrier motion needs to be accounted for, through the explicit solution of the energy-balance equation.

Using the field-dependent mobility concept, the energy-balance equation written in terms of the lattice and carrier temperatures is

$$\frac{\partial T_c}{\partial t} = \frac{2}{3k} \frac{1}{n} \vec{\varepsilon} \cdot \vec{j} - \left(\frac{T_c - T_L}{\tau_E} \right) \quad (4.116a)$$

In steady state and for spatial homogeneity, $\vec{j} = en\mu\vec{\varepsilon}$.

For Si devices and acoustic phonon scattering $\tau_E = (4 \times 10^{-12}) \sqrt{(T_c/T_L)}$ s and $\mu = \mu_0 \sqrt{(T_L/T_c)} \text{cm}^2 / (\text{V}\cdot\text{s})$

Then

$$e\mu\varepsilon^2 = \left(\frac{C = 10^{-9} \text{W}}{300 \text{K}} \right) (T_c - T_L) \sqrt{\frac{T_L}{T_c}}$$

$$\Rightarrow T_c = T_L + e\mu_0\varepsilon^2 \left(\frac{300 \text{K}}{C} \right) \quad (4.116b)$$

Then, for hot carriers

$$\vec{j} = en\mu_0 \frac{\varepsilon}{\sqrt{1 + e\mu_0\varepsilon^2 \left(\frac{300 \text{ K}}{T_L C}\right)}} \quad (4.116c)$$

$$j_{\text{sat}} = \frac{en\mu_0}{\sqrt{e\mu_0 \left(\frac{300 \text{ K}}{T_L C}\right)}} = env_{\text{sat}} \quad (4.116d)$$

Note that we have accounted for high-field transport, for example, in the pinchoff region of a MOSFET, but have done so using a field-dependent mobility in the context of the drift-diffusion formalism. As ULSI devices are miniaturized, and the field variations are very rapid in space and time, we cannot ignore the energy flux term in the energy balance equations (i.e., the history of carrier motion cannot be ignored).^{14,15} Then, one needs to solve the full-blown energy-balance equation, self-consistently with the drift-diffusion equation.¹⁶

4.7.3 MOSFET Simulations

We illustrate the application of the various formalisms to 0.1- μm -channel-length MOSFETs, using TMA MEDICI¹⁷ simulations. In Figure 4.5 we show a sequence of 2D plots of the carrier concentration profiles, equipotential contours, and the equienergy contours in the channel region of the NMOSFET using three types of models: low-field mobility in the drift-diffusion equation, field-dependent mobility in the drift-diffusion equation, and finally the energy-balance equation. We also show the 1D cuts along the channel, showing the longitudinal electric field profile, along with a cut of the electron energy contours. Comparing the field-dependent model, with the full energy-balance model, one sees that there is less carrier heating near the drain end in the latter case. This is because there is a divergence of the energy flux from the high-temperature region, leading to some cooling of the carriers. One also finds that in the energy-balance case, the carrier energies do not track the local electric field, because the past history of the carrier motion has to be taken into account. Such high-field effects are clearly very important for miniaturized ULSI MOS devices to model substrate current (which involves impact ionization across the 1.1-eV Si bandgap), gate current (which involves electron injection across the 3.1-eV conduction band-edge discontinuity between Si and SiO₂), and velocity overshoot effects, which can increase the drain current in deep-submicrometer MOSFETs. The output currents are higher in the energy-balance case because if there is less carrier heating, there is less high-field channel mobility degradation.

Quantization of 2DEG in MOSFET Inversion Layers

Schrieffer and Bardeen recognized as early as the 1950s that the carriers in the inversion layer of a MOSFET are confined in a narrow (approximately

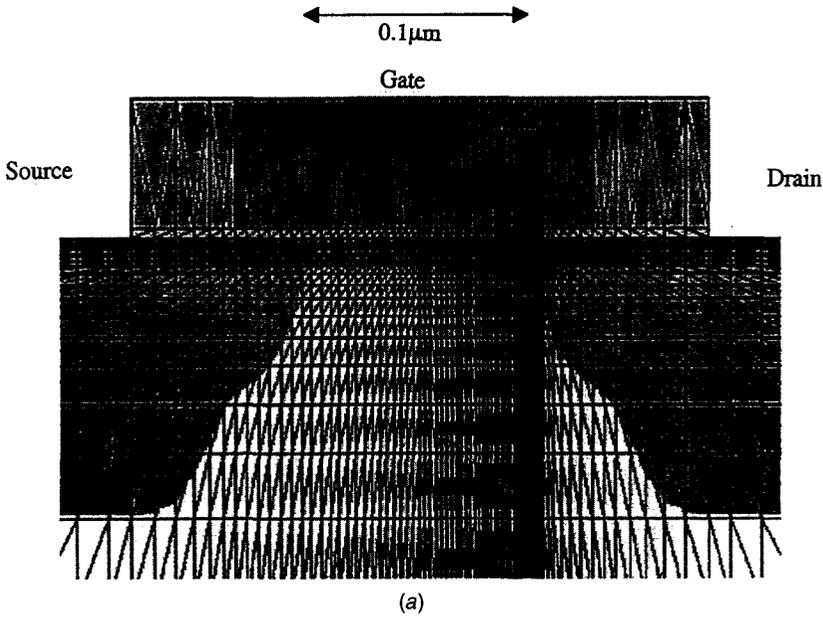
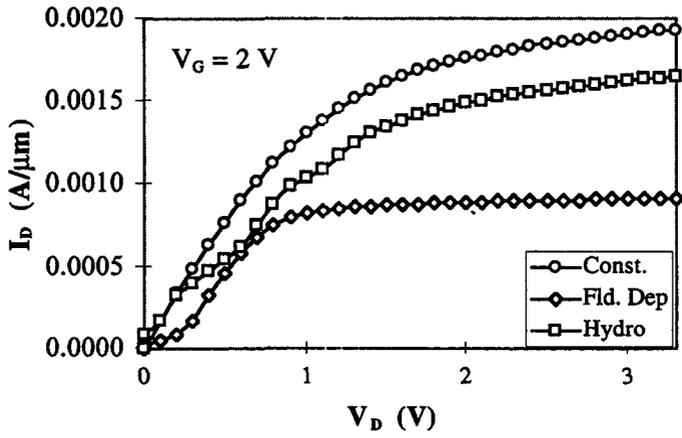
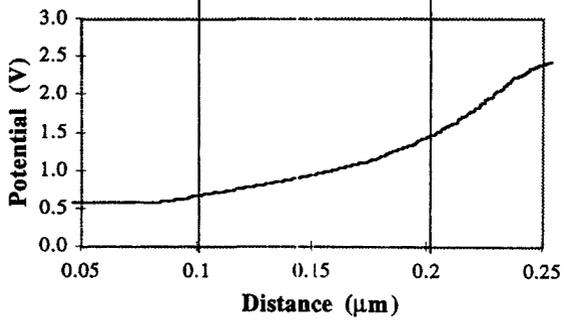
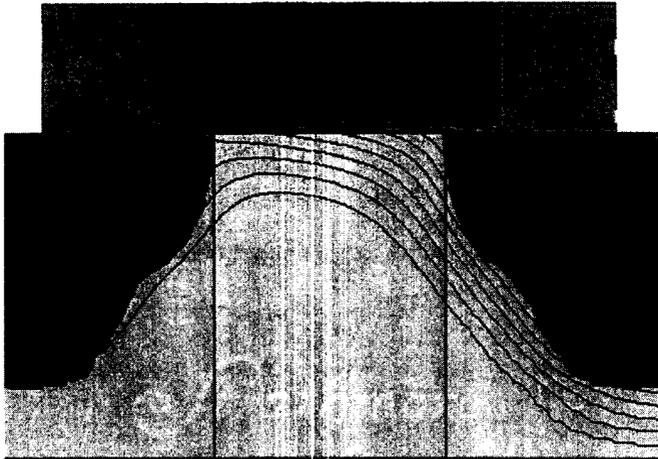


Figure 4.5 The results of TMA-MEDICI simulations of a deep-submicrometer MOSFET using drift diffusion models, and the energy-balance equation. The equipotential contours, equiconcentration contours, and equienergy contours are shown to demonstrate the differences in the results for the different hierarchies of hydrodynamic models. (a) Device and grid structure used in simulations using the device simulator MEDICI provided by TMA.¹⁷ (b) I_D versus V_D 1 μm (Left) MOSFET with $V_G = 2\text{ V}$. Three mobility models used are (1) constant mobility, (2) field-dependent mobility, (3) hydrodynamic model. (c) Potential contours from 0 to 3 V in 0.25-V steps ($V_{DS} = 2\text{ V}$, $V_{GS} = 2\text{ V}$) for constant-mobility model. Cut taken at the Si/SiO₂ interface. (d) Potential contours from 0 to 3 V in 0.25-V steps ($V_{DS} = 2\text{ V}$, $V_{GS} = 2\text{ V}$) for field-dependent mobility model. Cut taken at the Si/SiO₂ interface. (e) Potential contours from 0 to 3 V in 0.25-V steps ($V_{DS} = 2\text{ V}$, $V_{GS} = 2\text{ V}$) for hydrodynamic model. Cut taken at the Si/SiO₂ interface. (f) Equiconcentration contours from $1 \times 10^{18}\text{ cm}^{-3}$ to $1 \times 10^{21}\text{ cm}^{-3}$ for constant mobility. (g) Equiconcentration contours from $1 \times 10^{18}\text{ cm}^{-3}$ to $1 \times 10^{21}\text{ cm}^{-3}$ for field-dependent mobility. (h) Equiconcentration contours from $1 \times 10^{18}\text{ cm}^{-3}$ to $1 \times 10^{21}\text{ cm}^{-3}$ for hydrodynamic model. (i) Constant velocity contours (log scale values greater than $1 \times 10^7\text{ cm/s}$ separation of $2.5 \times 10^7\text{ cm/s}$) and cut at the Si/SiO₂ interface for constant-mobility model ($V_{DS} = 2\text{ V}$, $V_{GS} = 2\text{ V}$). (j) Constant-velocity contours (log scale values greater than $1 \times 10^7\text{ cm/s}$ separation of $2.5 \times 10^7\text{ cm/s}$) and cut at the Si/SiO₂ interface for field-dependent mobility model ($V_{DS} = 2\text{ V}$, $V_{GS} = 2\text{ V}$). (k) Constant-velocity contours (log scale values greater than $1 \times 10^7\text{ cm/s}$ separation of $2.5 \times 10^7\text{ cm/s}$) and cut at the Si/SiO₂ interface for hydrodynamic model ($V_{DS} = 2\text{ V}$, $V_{GS} = 2\text{ V}$).

triangular) potential well, about 5 nm in width. Since this external potential varies on a distance scale shorter than the De Broglie wavelength, the wave nature of electrons is manifested in the z direction and we have to solve the Schrödinger equation in this direction (Fig. 4.6). We get particle-in-a-box states, which are so-

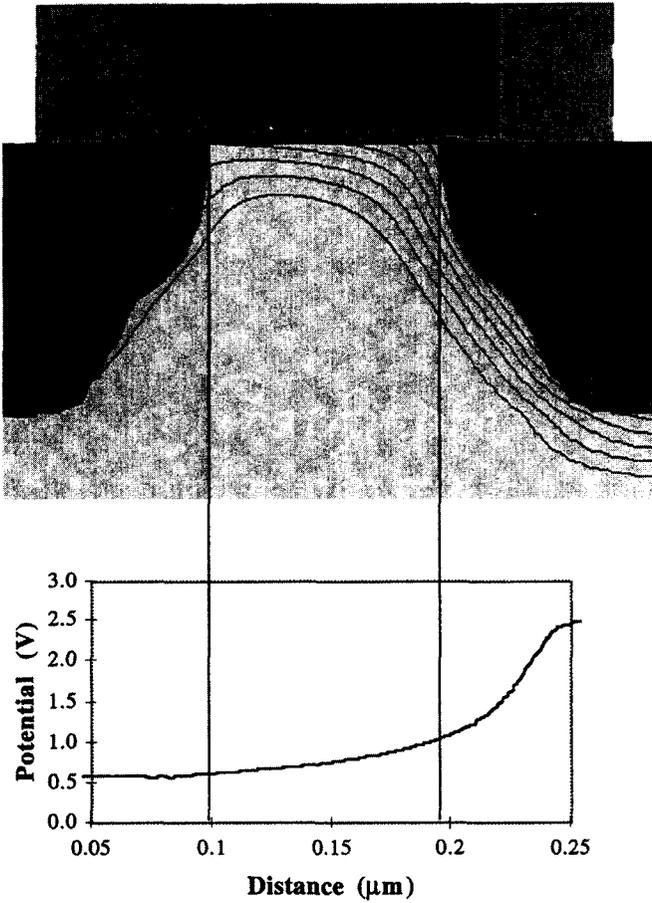


(b)



(c)

Figure 4.5 (Continued)



(d)

Figure 4.5 (Continued)

called Airy functions for a triangular well

$$\Phi_i(z) = Ai\left(\left(\frac{2me\varepsilon}{\hbar^2}\right)^{1/3} \left(\frac{z - E_i}{e\varepsilon}\right)\right) \quad (4.117a)$$

with eigenenergies:

$$E_i = \left(\frac{\hbar^2}{2m}\right)^{1/3} (1.5\pi e\varepsilon(i + 0.75))^{2/3} \quad (4.117b)$$

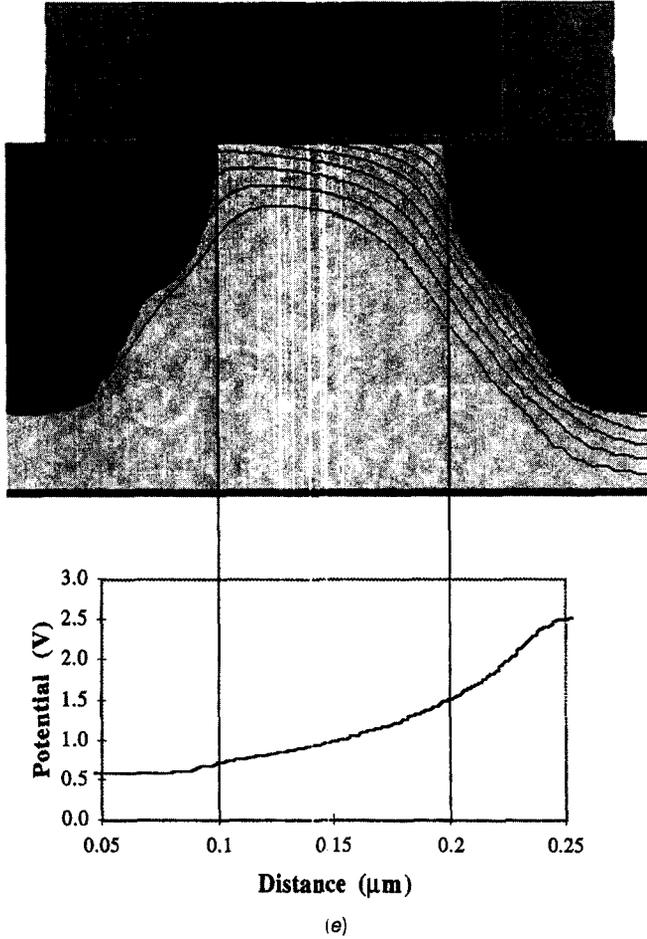
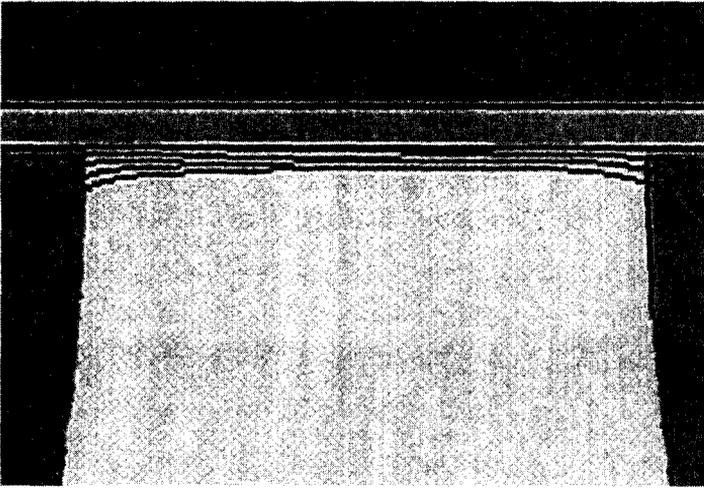


Figure 4.5 (Continued)

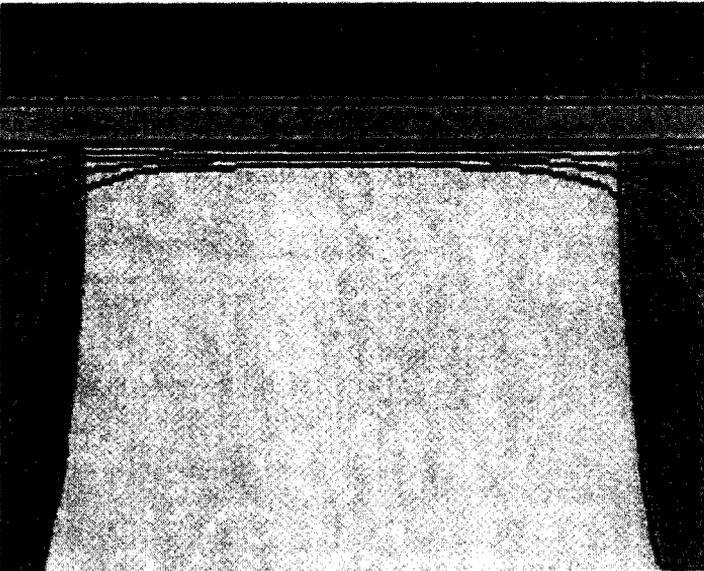
Here, ε is the average electric field in the z direction. The Airy functions are analogous to the sine or cosine function solution in a rectangular potential well.¹⁸ Unlike a rectangular well, where the eigenenergies are spaced farther apart for the higher energies, here in a triangular inversion-layer well, the energies are seen to become increasingly closely spaced at higher energy values, until they all merge with the bulk conduction-band states. Physically, the reason for this difference is that the well width continues to increase at higher energies for a triangular well.

On the other hand, the electrons are not confined in the x and y directions, leading to free-electron-like plane waves, and a 2D electron gas (2DEG). Hence, in the x - y plane, the eigenenergies are

$$E_{\parallel} := \frac{\hbar^2 k_{\parallel}^2}{2m^*} \tag{4.117c}$$



(f)

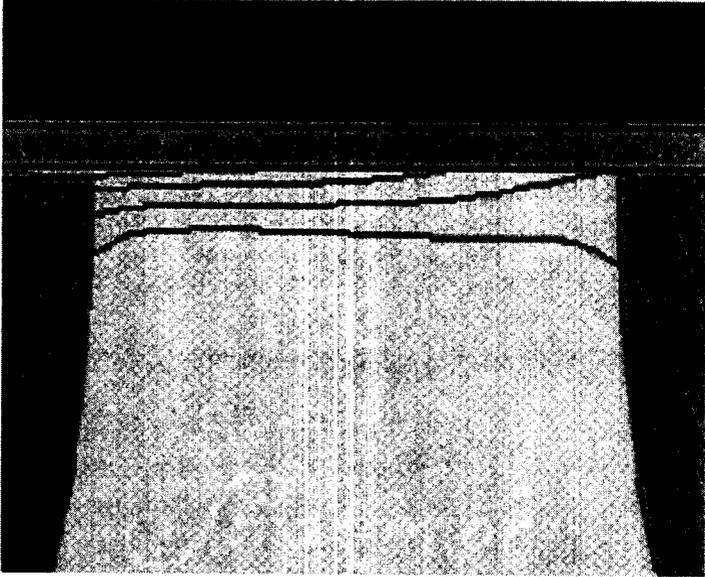


(g)

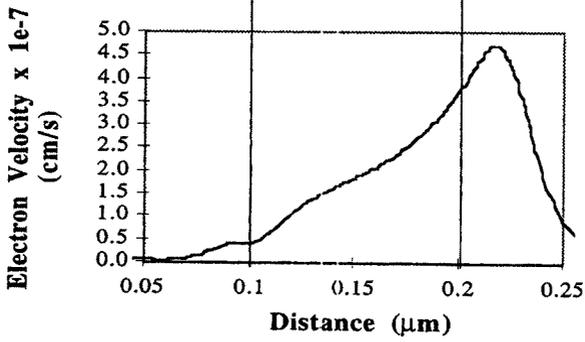
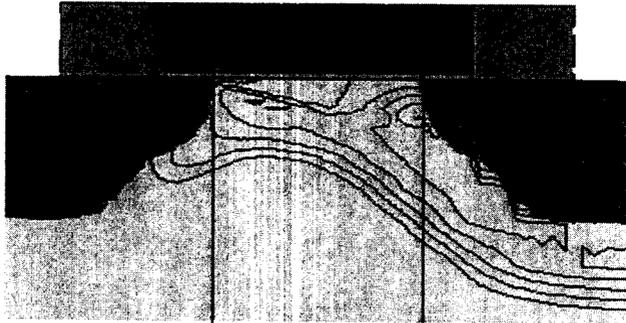
Figure 4.5 (Continued)

The total energy of an electron will then be this energy plus whatever energy, E_i , we have in a particular subband due to quantization in the z direction.

One needs to iteratively solve the Schrödinger and Poisson equations to get the energy levels and the wavefunctions for the subbands, where the number of subbands that must be considered depends on the level of accuracy demanded. Typically, two to five subbands suffice. We get the one-electron Schrödinger



(h)



(i)

Figure 4.5 (Continued)

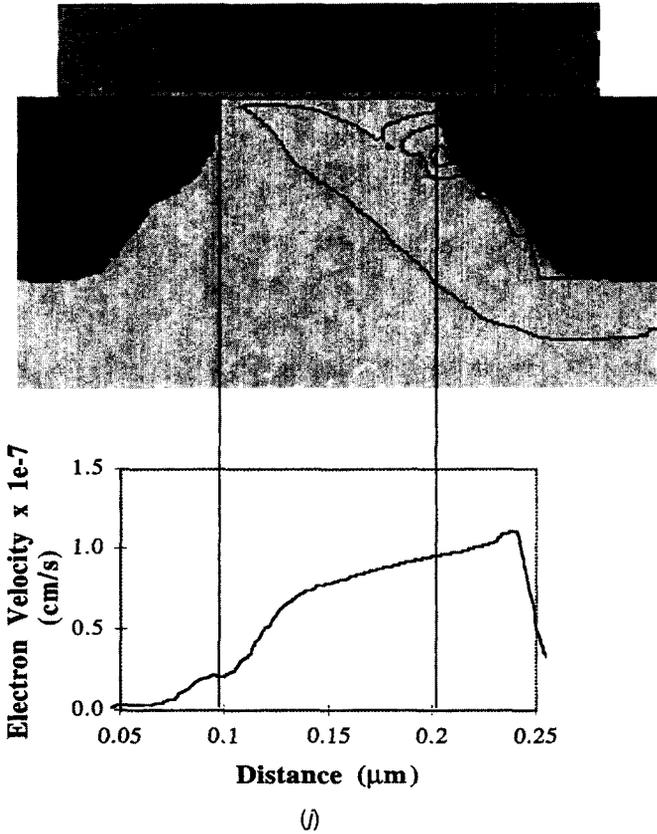


Figure 4.5 (Continued)

equation in the envelope function approximation

$$-\frac{\hbar^2}{2m^*} \frac{d^2\phi_i}{dz^2} - eV(z)\phi_i(z) = E_i \overbrace{\phi_i(z)}^{\text{envelope functions}} \tag{4.118a}$$

where

$$V(z) = V_{\text{electrostatic}}(z) + V_h(z) + V_{xc}(z) + V_{im}(z) \tag{4.118b}$$

Here, the four terms are, respectively, due to the electrostatic potential (due to the fixed dopant and mobile carrier charges), a step discontinuity between the conduction band in the Si channel and the gate oxide (=3.1 eV), an exchange correlation “repulsive” potential due to electron fermion wavefunctions being antisymmetrized, and finally an image force term due to the difference of dielectric constant between Si and SiO₂. For our discussion here, we focus only on the most

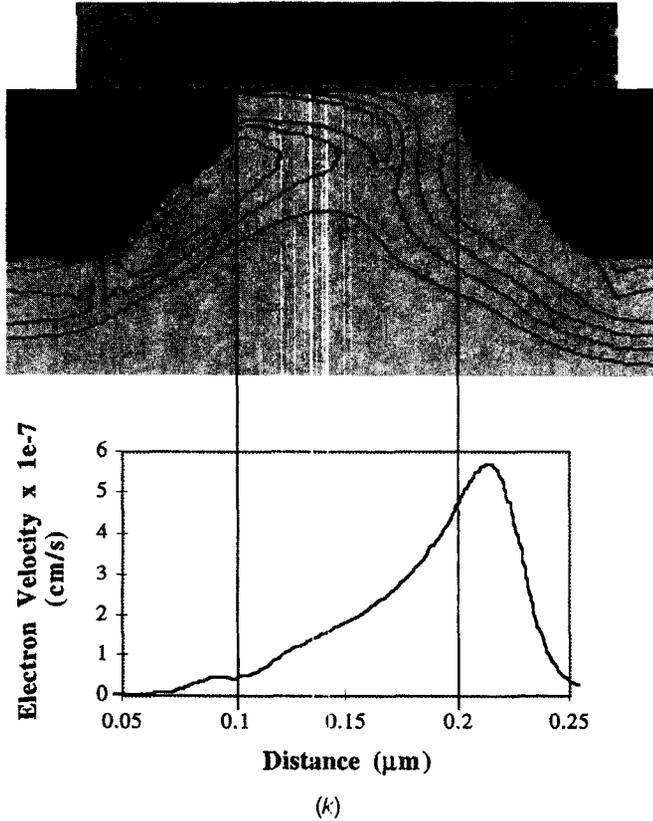


Figure 4.5 (Continued)

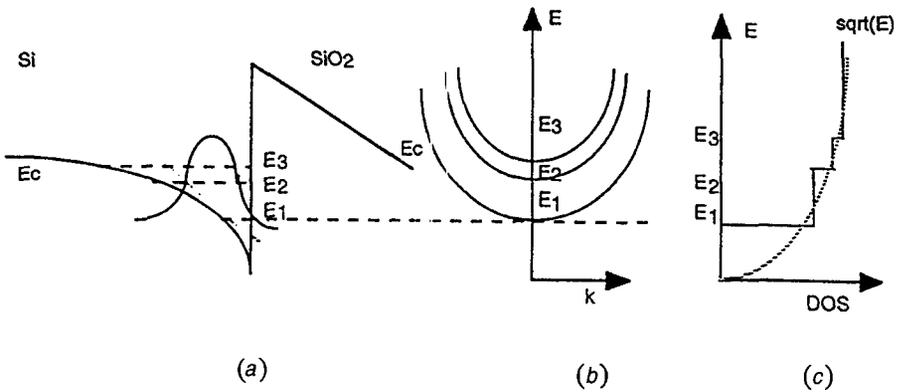


Figure 4.6 The 2D electron gas in the inversion-layer triangular quantum well in the channel of a MOSFET, showing (a) the envelope functions in (b) the different subbands, and the (c) “staircase” stepwise density of states for the electron gas.

important terms (the first two). Since the electrostatic potential depends on the mobile carrier distribution, one needs to self-consistently solve the Schrödinger equation with the Poisson equation:

$$\frac{d^2 V_{\text{electrostatic}}}{dz^2} = \frac{e}{\epsilon_r \epsilon_o} \left[\sum_{i=1}^M N_i \phi_i^2(z) + N_A(z) - N_D(z) \right] \quad (4.119)$$

Here, N_i is given by the DOS in a 2DEG, which we found earlier to be a constant. Actually, we end up with a so-called “staircase” DOS when we add the different constant DOS terms, starting at the different subband energies, E_i (Fig. 4.6). The envelope of this staircase DOS is the parabolic DOS for a 3D bulk system. In fact, the inversion layer 2D DOS asymptotically merges with the 3D parabolic DOS at high subband energies. We then get

$$N_i = \frac{m^* k_B T_c}{\hbar^2 \pi} \ln \left[1 + \exp \left[\frac{E_F - E_i}{k_B T_c} \right] \right] \quad (4.120)$$

Here, the envelope function the z direction gives us the electron distribution as a function of depth in the inversion layer. A powerful approach to this formidable computational problem is to use a variational technique (illustrated for two subbands), which involves “trial” wavefunctions in terms of unknown, variational parameters, b_0 and b_i , which are found by minimizing the total energy of the system:

$$\phi_0(z) = \left(\frac{b_0^3}{2} \right)^{1/2} z \exp\left(-b_0 \frac{z}{2}\right) \quad (4.121a)$$

$$\phi_1(z) = \left(\frac{3}{2} \frac{b_1^5}{b_0^2 - b_0 b_1 + b_1^2} \right)^{1/2} \left[z - \frac{b_0 + b_1}{6} z^2 \right] \exp\left(-\frac{b_1 z}{2}\right) \quad (4.121b)$$

The wavefunctions vanish at the Si/SiO₂ interface, and do not penetrate into the oxide (Fig. 4.6a). This statement has dramatic ramifications, and implies that unlike the classical inversion charge distribution, which has a maximum at the Si/SiO₂ interface, the quantum-mechanical distribution peaks away from the interface at $(1/b) \sim 0.5$ nm. This implies that there is a quantum-mechanical “dead layer” in the Si where there are very few inversion carriers, which function as a Si dielectric with a quantum-mechanical capacitance:

$$C_{QM} = \epsilon_o \epsilon_r b \quad (4.122)$$

This C_{QM} is in series with C_{OX} .^{18,19} This means that as gate oxides continue scaling, one will reach a point of diminishing returns in terms of the gate drive. The quantum mechanical model will also affect the classical calculation of the threshold voltage, V_T , through the modification of the C_{OX} term and because of an apparent increase of the Si bandgap because the lowest subband energy is higher than the conduction-

band edge, E_c . On the other hand, the surface roughness scattering will be less in a quantum-mechanical picture because the carriers are farther from the atomically rough Si/SiO₂ interface. The other scattering mechanisms are also affected because the parabolic DOS is now modified as a staircase DOS, and this has to be accounted for in the Fermi golden rule scattering-rate calculations.

Examples of Monte Carlo and Hydrodynamic Studies of MOSFETs

We next discuss examples of MOSFET simulations based on Monte Carlo and hydrodynamic techniques, with quantization effects and realistic bandstructure taken into account, as described in the previous sections.^{20–25} Figure 4.7a shows carrier energy as a function of position in the channel determined with ensemble

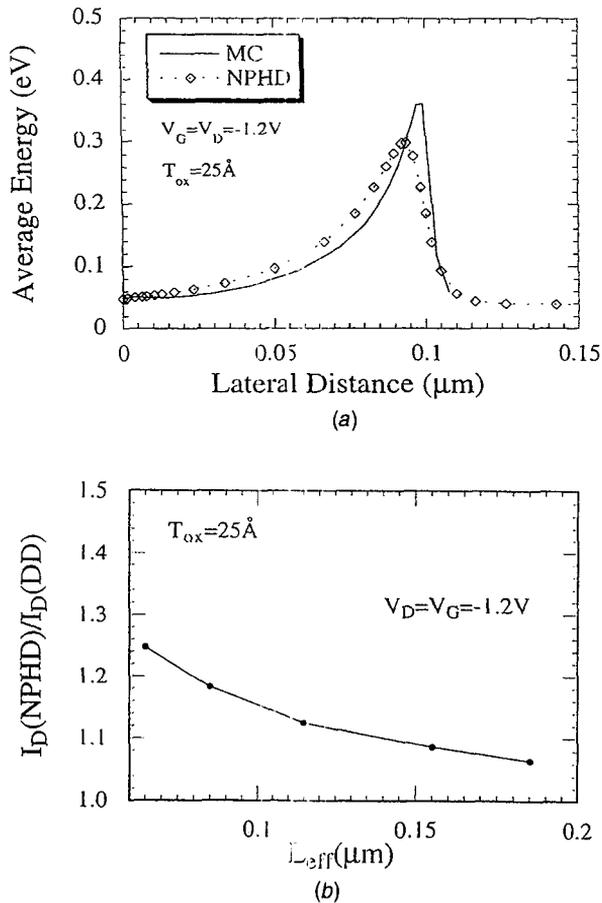


Figure 4.7 (a) Monte Carlo (MC) and nonparabolic hydrodynamic (NPHD) simulations of electron energy with nonparabolicities of bandstructure accounted for as a function of position in the channel of a 0.1 μm MOSFET; (b) velocity overshoot as a function of channel length for PMOSFETs.

Monte Carlo (MC) simulations with computationally efficient fitted bandstructure in terms of band curvature and density of states (DOS). The dynamics of the electrons are correct because of the accurate energy-dependent band curvature or effective mass, while the scattering rates are correct because of the proper DOS. While Monte Carlo gives an extremely accurate description of the carrier distribution functions, it is computationally slow. For comparison, we also show the carrier heating computed by an efficient hydrodynamic model with nonparabolic bandstructure (NPHD), which captures the complexities of the DOS of a realistic bandstructure better than a simple effective-mass model. Figure 4.7*b* shows the results of velocity overshoot on drain current in deep-submicrometer pMOSFETs. It shows that nonlocal effects can enhance the drain current below 0.1- μm channel lengths by as much as 25% over that predicted by a simple, local-field, drift-diffusion-type model.

The quantization effects in the channel can cause significant shifts in the MOSFET parameters, compared to a classical model (Fig. 4.8). Figure 4.8*a* shows the difference between the classical threshold voltage (V_T) and the quantum-mechanical prediction of V_T for the reasons discussed in this section.^{26–28} There are also significant differences of the capacitance-voltage behavior calculated classically and quantum-mechanically (Fig. 4.8*b*). As mentioned before, the difference is due to the “effective” quantum capacitance in the inversion layer in series with the oxide thickness.

The effects of quantization on transport in a 2DEG in the inversion layer are shown in Figure 4.9. Figure 4.9*a* shows relative electron concentration as a function of depth in the channel near the source end (where the electric field parallel to the channel is low) and near the pinchoff region of the channel near the drain end (where the parallel component of the field is high). One sees that there is a significant difference between the classical and quantum-mechanical distributions near the source end, clearly indicating the importance of quantum effects. The difference is much less near the pinched-off drain because there is much less carrier confinement there. The integrated sheet charge densities in the inversion layer are shown for the classical case and the quantum case as a function of gate bias in Figure 4.9*b*.

4.8 SUMMARY AND FUTURE TRENDS

The projected evolution of prototypical ULSI devices such as MOSFETs suggests that channel lengths and equivalent gate oxide thicknesses will be ~ 150 and ~ 2 nm, respectively, by the year 2001.²⁹ By 2012, the respective numbers will be 50 and ~ 1 nm. In terms of gate oxide scaling, Fowler-Nordheim or direct tunneling is an obvious problem, and the quantum-mechanical capacitance in the Si inversion layer will become increasingly important. As mentioned above, the thickness of the quantum-mechanically induced Si dielectric layer is ~ 0.5 nm, which will constitute a significant percentage of the gate oxide thickness. In other words, gate drive will not increase linearly with C_{OX} in this regime.

An even more important question is that of transport of carriers, for example, in the inversion layer. The question arises as to how far the device simulation

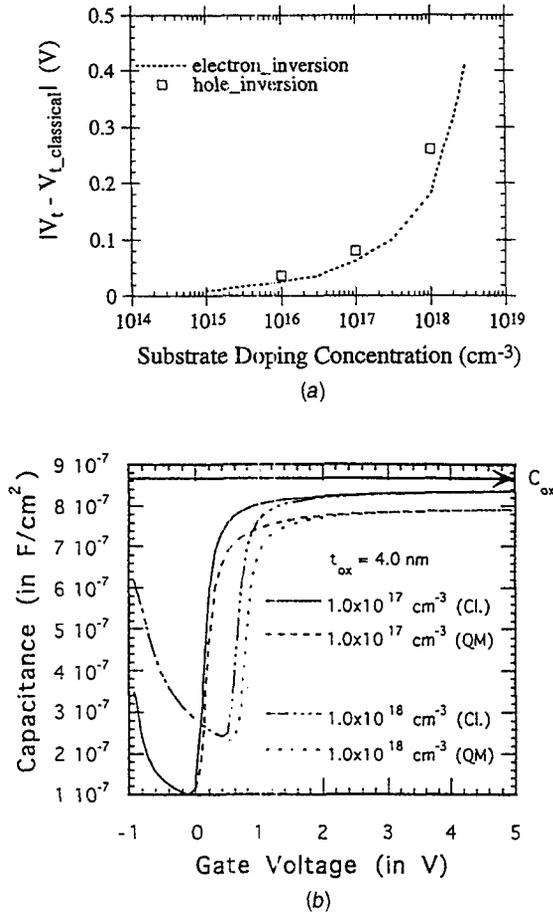


Figure 4.8 (a) Simulated threshold voltage shift versus substrate doping concentration due to the quantization effects for both electrons and holes; (b) classical and quantum calculations of the quasistatic capacitance are presented for n-MOS structures. The oxide capacitance is also shown for comparison. The additional degradation in the capacitance is due to a larger “effective” oxide thickness in the quantum case as compared to that from a classical calculation.

methodology based on semiclassical dynamics and the BTE will be valid, and at what technology-scaling limit one will enter the realm of quantum transport.³⁰ To answer this question, it might be instructive to briefly review the different length scales and time scales of interest in semiconductor device simulation.

One of the most important factors is the so-called thermal De Broglie wavelength (~ 12.5 nm in Si), defined in terms of the thermal velocity of carriers, v_{th} ($\approx 10^7$ cm/s at room temperature in Si), and the equipartition theorem as

$$\frac{m^* v^2}{2} = \frac{3}{2} kT_0 \quad \text{and} \quad \lambda_B = \frac{h}{p} = \frac{h}{m^* v} \quad (4.123)$$

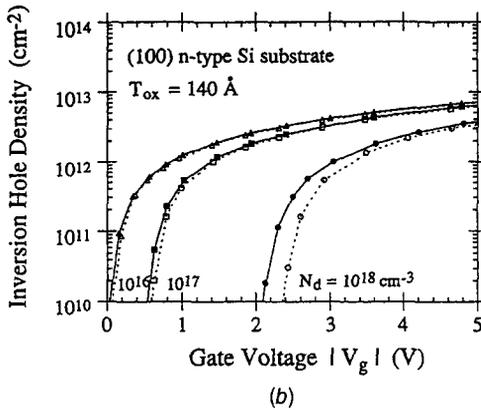
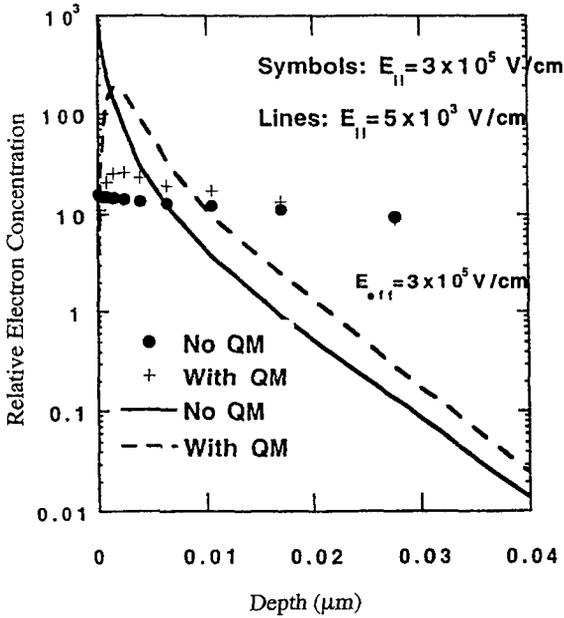


Figure 4.9 (a) Relative electron concentration as a function of depth in the channel with and without quantum-mechanical (QM) effects taken into account near the source end (lines) and near the drain end (points); (b) classical (dotted lines) and quantum-mechanical (solid lines) hole sheet change density in the channel of a MOSFET as a function of gate voltage.

While this is the relevant expression for dilute electron gases at high temperature, for degenerate conditions, one should use the Fermi velocity, v_F ($> 10^7$ cm/s, depending on carrier concentration) in place of the thermal velocity. As mentioned above, whenever the potentials vary on distance scales comparable to the De Broglie wavelength, a quantum mechanical treatment is necessary. Another length scale of

relevance is the mean free path, L_m . This, in turn, depends on the momentum relaxation time, τ_m , which is related to but greater than the collision time, τ_c . The reason is that, as we discussed in the section on scattering, not all scattering processes randomize the wavevector, k , or electron momentum to the same extent. It is thus necessary to weight the scattering rate by the *relative* change of k on scattering. One then gets $L_m = (v_{th} \cdot \tau_m)$ or $(v_F \cdot \tau_m)$. Finally, one needs to worry about the phase relaxation time, L_p , which is analogously dependent on the phase relaxation time, τ_p , which is also generally greater than τ_c for the same reason as above. τ_p is generally shorter than τ_m . In other words, the phase memory of electrons on scattering is generally destroyed more quickly (and on shorter distance scales) than the destruction of the collimated momenta of an ensemble of electrons. In this context, it is important to mention that static scatterers such as ionized impurities relax momentum, but do not destroy phase information. Only dynamic scatterers such as phonons or carrier-carrier scattering can do so.³¹

With this brief preamble, let us examine the limits of validity of semiclassical electron dynamics. The semiclassical treatment is strictly applicable only for electrons confined to a particular band. This requires that the electric fields be small enough such that when electrons move by a lattice constant, a , the carriers cannot gain kinetic energy equal to the bandgap:

$$e \varepsilon a < E_g \quad \text{or} \quad e \varepsilon a \ll \frac{E_g^2}{E_F} \quad (4.124a)$$

Otherwise, we get interband transitions or electric breakdown (Zener tunneling). One also cannot have the electric fields vary too rapidly, on the distance scale of the De Broglie wavelength, which is the spatial extent of the electron wavepacket. Otherwise, it becomes necessary to modify the equations of semiclassical dynamics. Similarly, the applied fields must vary slowly in time, or we get photon-assisted interband transitions

$$\hbar\omega = E_g; [f_{\max} \sim 2 \times 10^{14} \text{ Hz at } E_g = 1.1 \text{ eV}] \quad (4.124b)$$

The magnetic fields and the cyclotron frequency must also be low enough to avoid interband transitions, $\hbar\omega_c \ll (E_g^2/E_F)$, or we get what is called *magnetic breakdown* (in analogy to electric breakdown)³.

Another class of more serious limitations is imposed on the BTE formalism by the Heisenberg uncertainty relations.⁷ As mentioned before, the BTE uses a distribution function, $f(r, k, t)$ or $f(r, p, t)$ in the 6D phase space of positions and momenta. However, it tacitly implies that one can simultaneously specify the position and momentum of the electron with arbitrary precision, in violation of the uncertainty relationship ($\Delta r \Delta p \geq \hbar$). In reality, one cannot localize the carrier to better than the De Broglie λ_B . In other words, one must introduce a quantum-mechanical uncertainty in the distribution function, and one does so in quantum transport using the language of density matrices or the so-called Wigner distribution

function (which is the analog of the BTE). Another place where the uncertainty principle appears is in energy–time uncertainty, $\Delta E \Delta t \geq \hbar$. If $\Delta t \sim \tau$ (time between collisions), then, for carrier energies to be well defined in the initial and/final states in the Fermi golden rule expressions for carrier scattering, we must have $\tau \gg [\hbar/E(p)]$. Otherwise, for very high scattering rates, we have what is called collisional (Lorentzian lineshape) broadening of the energy levels, and the $\delta(E_f - E_i)$ denoting energy conservation is changed from discrete energy levels.

The final set of limitations of the BTE formalism is introduced by the fact that there is no phase information for electrons in the distribution function, $f(r, k, t)$. This is one of the premises behind the random phase approximation (RPA). One also makes the tacit assumption here that the scattering rate is low enough that the mean free path and the phase relaxation lengths are greater than the De Broglie wavelength. That is why one can treat the carrier scattering as an incoherent sum of single scattering events. If not, one has to account for the possibility that multiple scattering events can occur within a De Broglie wavelength. Also, the assumption that the scattering event instantaneously changes the carrier momentum or wavevector, k , is not strictly valid. One must take into account the evolution of k during the collisional process itself (the collision “sphere”), which, as mentioned briefly above, gives rise to the intracollisional field effect.

If the scattering rates are low enough that the mean free path and the phase relaxation lengths begin to approach the device dimensions, one cannot use the semiclassical BTE formalism, but must use quantum transport methodology. While it is not clear exactly at what technology scaling limit quantum transport treatment will be mandatory, and which particular quantum transport formalism will be the most useful [Wigner functions, Feynman path integrals, Landauer–Buttiker approach or nonequilibrium Green functions (NEGF)],³¹ the consensus is that some of these equivalent methodologies will be needed as ULSI devices are miniaturized in the near future.^{30,31} The NEGF method is becoming one of the more attractive approaches, and can be used to briefly illustrate our final device simulation example: dopant fluctuation phenomena in the channel region of deep-submicrometer MOSFETs.

The volume in the channel region under the gate of such deep-submicrometer MOSFETs is so small that the number of dopants within that volume will be very small (~ 50). It is clear that there can be significant statistical variations of the number and precise location of these dopants. In the BTE formalism, one does not keep track of the phases of the electrons as they scatter off the different ionized impurities. One simply assumes incoherent scattering, and sums up the individual scattering events. Now, however, as the phase relaxation lengths approach the channel length, clearly one must account for the possibility of destructive and constructive interference of the electron waves scattering off the individual dopant charges. Because these dopants are static scatterers, they do not randomize the phase, as mentioned above. The way to handle this is to introduce Green’s functions, which are correlation functions of the form $G(r, r'; t, t')$ or $G(k, k'; t, t')$. These functions tell us the likelihood of an electron injected in the device at a location r (or momentum, k) at time t , appearing later at a location r' (or momentum, k') at a time

t' , retaining the phase information. Six different Green functions correspond to various physical quantities. For example, the so-called G "less than" function, $G^<$, corresponds to the electron concentration, while the G "greater than" function, $G^>$, corresponds to the holes. Using these formulations, one can handle dopant fluctuation phenomena where the drain current fluctuates (statistical noise) corresponding to the precise locations of the dopants in the channel.^{30,31} These are sometimes known as *universal conductance fluctuations in mesoscopic systems* (which are intermediate between microscopic and macroscopic structures). In fact, there is a school of thought that predicts that it is statistical fluctuations in device characteristics of this type which will determine the ultimate scaling limit of ULSI devices, rather than limitations of the fabrication process (e.g., lithography) per se.

ACKNOWLEDGMENTS

The authors acknowledge Soji John for help with the MOSFET simulations, and IEEE for Figure 4.3. Figures 4.7–4.9 were provided by Profs. Al Tasch and Chris Maziar. We also thank Kay Ellen Shores for assistance with the typing and the figures.

REFERENCES

1. S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1981.
2. B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, 2000.
3. N. Ashcroft and N. Mermin, *Solid State Physics*, Holt, Reinhart and Winston, New York, 1976.
4. J. Chelikowsky and M. Cohen, "Nonlocal Pseudopotential Calculations for the Electronic Structure of Eleven Diamond and Zind-Blende Semiconductors," *Phys. Rev. B* **14**, 556 (1976).
5. J. Luttinger and W. Kohn, "Motion of Electrons and Holes in Perturbed Periodic Fields," *Phys. Rev.* **97**, 869 (1955).
6. G. Bastard and J. Brum, "Electronic States in Semiconductor Heterostructures," *IEEE J. Quant. Electron.* **22**, 1625 (1986).
7. K. Hess, *Advanced Theory of Semiconductor Devices*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
8. H. Brooks, "Scattering by Ionized Impurities in Semiconductors," *Phys. Rev.* **83**, 879 (1951).
9. E. M. Conwell and V. Weisskopf, "Theory of Impurity Scattering in Semiconductors," *Phys. Rev.* **77**, 388 (1950).
10. E. Conwell, in F. Seitz, D. Turnbull, and H. Ehrenreich, eds., *Solid State Physics*, Academic Press, New York, 1967.
11. C. Jacoboni and L. Reggiani, "The Monte Carlo Method for the Solution of Charge Transport in Semiconductors with Applications to Covalent Materials," *Rev. Mod. Phys.* **55**, 645 (1983).

12. D. Hughes, R. Abram, and R. Kensall, "An Investigation of Graded and Uniform Base $\text{Ge}_x\text{Si}_{1-x}$ HBT's Using a Monte Carlo Simulation," *IEEE Trans. Electron Devices* **42**(2), 201 (1995).
13. D. Rode, in Ed. R. Willardson and Q. Beer, eds., *Semiconductors and Semimetals*, vol. 10, Academic Press, New York, 1975.
14. G. Baccarani and M. Wordemann, "An Investigation of Steady-State Velocity Overshoot in Silicon MOSFETs," *Solid State Electron.* **28**, 407 (1985).
15. D. Ferry, K. Hess, and P. Vogl, in N. Einspruch, ed., *VLSI Electronics*, Vol. 2, Academic Press, New York, 1981.
16. S. Yoganathan and S. Banerjee, "A New Decoupled Algorithm for Non-stationary Transient Simulation of GaAs MESFETs," *IEEE Trans. Electron Devices* **39**(7), 1578 (1992).
17. TMA-MEDICI is a trademark of Technology Modeling Associates/Avant Corporation.
18. T. Ando, A. Fowler, and F. Stern, "Electronic Properties of Two-Dimensional Systems," *Rev. Mod. Phys.* **54**, 437 (1982).
19. C. Hu, S. Banerjee, K. Sadra, B. Streetman, and R. Sivan, "Quantization Effects in Inversion Layers of PMOSFETs on Si(100) Substrates," *IEEE Electron Device Lett.* **17**(6), 276 (1996).
20. J. Lopez-Villanueva, F. Gamiz, J. Roldan, Y. Ghailan, and J. Carcellar, "Study of the Effects of a Stepped Doping Profile in Short Channel MOSFETs," *IEEE Trans. Electron Devices* **44**(9), 1425 (1997)
21. K. Hasnat, C. Yeap, S. Jallepalli, W. Shih, S. Hareland, V. Agostonelli, A. Tasch, and C. Maziar, "A Pseudo-Lucky Electron Model for Simulation of Electron Gate Current in Submicron NMOSFETs," *IEEE Trans. Electron Devices* **43**(8), 1264 (1996).
22. R. Hulfachor, J. Ellis-Monaghan, K. Kim, and M. Littlejohn, "Spatial Retardation of Carrier Heating in Scaled 0.1 μm N-MOSFETs Using Monte Carlo Simulations," *IEEE Trans. Electron Devices* **43**(4), 661 (1996).
23. C. Wann, K. Noda, T. Tanaka, M. Yoshida, and C. Hu, "A Comparative Study of Advanced MOSFET Concepts," *IEEE Trans. Electron Devices*, **43**(10), 1742 (1996).
24. J. H. Sim, "An Analytical Deep Submicron MOS Device Model Considering Velocity Overshoot Behavior Using Energy Balance Model," *IEEE Trans. Electron Devices* **42**(5), 864 (1995).
25. J. Ellis-Monaghan, R. Hulfachor, K. Kim, and M. Littlejohn, "Ensemble Monte Carlo Study of Interface-State Generation in Low-Voltage Silicon MOS Devices," *IEEE Trans. Electron Devices* **43**(6), 1123 (1996).
26. S. A. Hareland, S. Krishnamurthy, S. Jallepalli, C.-F. Yeap, K. Hasnat, A. F. Tasch, Jr., and C. M. Maziar, "A Computationally Efficient Model for Inversion Layer Quantization Effects in Deep Submicron N-Channel MOSFETs," *IEEE Trans. Electron Devices*, **43**(1), 90 (1996).
27. S. Jallepalli, J. Bude, W.-K. Shih, M. R. Pinto, C. M. Maziar, and A. F. Tasch, Jr., "Electron and Hole Quantization and Their Impact on Deep Submicron Silicon P- and N-MOSFET Characteristics," *IEEE Trans. Electron Devices* **44**(2), 297 (1997).
28. A. Pacelli, "Self-consistent Solution of the Schrodinger Equation in Semiconductor Devices by Implicit Iteration," *IEEE Trans. Devices* **44**(7), 1169 (1997).
29. International Technology Roadmap for Semiconductors, 1999 edition.

30. D. K. Ferry and C. Jacoboni, *Quantum Transport in Semiconductors*, Plenum Press, New York, 1992.
31. S. Datta, *Electronic Transport in Mesoscopic Systems*, Univ. Press, Cambridge, 1995.

PROBLEMS

- 4.1 For a body-centered cubic (BCC) real lattice, determine the reciprocal lattice.
- 4.2 Using semiclassical electron dynamics, calculate the time taken by an electron to go from the gamma point to the edge of the Brillouin zone for a FCC lattice with lattice constant $a = 5 \text{ \AA}$, for a field of 1 kV/cm in the $[100]$ direction.
- 4.3 Calculate the total number of quantum states, or k values, in a Si slab $10 \times 2 \times 5 \text{ cm}$.
- 4.4 Calculate the average volume number density of electrons at low temperature in a 2D electron gas (effective mass $= 0.2 m_0$) in a Si quantum well (100 \AA wide) where the Fermi level lies midway between the second and third subbands.
- 4.5 For a 1D lattice with two different atom masses, M and m , in an interleaved configuration, an atomic spacing of a and a spring constant of α , determine the phonon–dispersion relationship, assuming Hooke’s law to be valid.
- 4.6 For a delta-function scattering potential in a semiconductor, derive the expression for the total scattering rate using the Fermi golden rule.
- 4.7 Derive an expression for the average kinetic energy per electron for a nondegenerate electron gas in a semiconductor, with a drifted Maxwellian distribution function (displaced by k_d).
- 4.8 Prove that we get the drift diffusion model by taking the first moment of the Boltzmann transport equation.
- 4.9 Solve the Boltzmann transport equation in 1D, in steady state, for a homogeneously doped bar of Si for low electric fields.
- 4.10 Prove Eq. 4.120 for the electron concentration in an inversion layer.

PART II

**DEVICE BUILDING BLOCKS AND
ADVANCED DEVICE STRUCTURES**

SOI and Three-Dimensional Structures

JEAN-PIERRE COLINGE
University of California
Davis, CA

5.1 INTRODUCTION

The idea of realizing semiconductor devices in a thin silicon film that is mechanically supported by an insulating substrate has been around for several decades (at the time of writing). The first description of the insulated-gate field-effect transistor (IGFET), which evolved into the modern silicon metal oxide semiconductor field-effect transistor (MOSFET), is found in the historical patent of Lilienfeld dating from 1926.¹ This patent depicts a three-terminal device in which the source-to-drain current is controlled by a field effect from a gate that is dielectrically insulated from the rest of the device. The piece of semiconductor that constitutes the active part of the device is a thin semiconductor film deposited on an insulator. In a way, one can say that the first MOSFET was a semiconductor-on-insulator device.

Bulk silicon MOSFETs are imperfect devices. The use of junction isolation gives rise to *nonnegligible junction capacitances and to some undesirable effects, such as latchup in CMOS circuits*. Because the depletion region beneath the gate is physically connected to the substrate, a relatively poor electrostatic coupling is obtained between the gate and the channel. In terms of device performance, this poor coupling results into a relatively high body factor (also called the *body-effect coefficient*).² This reduces the current drive of the devices, especially when their source is at a higher potential than the substrate, and it reduces the gain and the transconductance of the transistors. For the same reason it is impossible to reduce the threshold voltage of bulk MOSFETs below values of approximately 600 mV without increasing their OFF current, which jeopardizes the use of such devices for low-voltage applications requiring a supply voltage of 1 V or below.

Silicon-on-insulator (SOI) devices a low one to overcome these problems found in bulk silicon. In addition, with the exception of silicon-on-sapphire substrates, SOI wafers can be processed using exactly the same standard tools and techniques as bulk silicon. This chapter is devoted to the description of the benefits and the characteristics associated with SOI technology, as well as that of novel, 3D structures developed using SOI substrates.

5.2 SOI SUBSTRATES

An SOI substrate consists on a thin, single-crystal, defect-free sheet of silicon sitting on top of an insulator. Over a dozen generic techniques have been developed to produce such a material. In addition to a good crystalline quality, SOI films must have a good thickness uniformity and passivated silicon-insulator interfaces. Producing such materials is quite a challenge. The most successful techniques for producing SOI material are described next.

5.2.1 Silicon-on-Sapphire Material

Silicon-on-sapphire (SOS) material was first proposed in 1963 and has been commercially available since 1971. The fabrication of SOS wafers starts with the growth of sapphire (α - Al_2O_3) crystals which are produced using different possible techniques, including Czochralski growth. After sawing and polishing, the sapphire wafers receive a final hydrogen etching at 1150°C in an epitaxial reactor and a silicon film is deposited using the pyrolysis of silane at temperatures between 900 and 1000°C . The lattice parameters of silicon and sapphire are 0.357 and 0.355 nm, and their mean thermal expansion coefficient is $3.8 \times 10^{-6}^\circ\text{C}^{-1}$ and $9.2 \times 10^{-6}^\circ\text{C}^{-1}$, respectively. Due to lattice mismatch, the defect density in the silicon film is quite high, especially in very thin films. Furthermore, the thermal expansion coefficient mismatch results in a compressive stress in the silicon film as the wafers are cooled after silicon deposition. As the film thickness increases, however, the defect density appears to decrease as a simple power-law function of the distance from the Si-sapphire interface. The main defects present in the as-grown SOS films are aluminum autodoping from the Al_2O_3 substrate, stacking faults, and microtwins. Typical defect densities near the Si-sapphire interface reach values as high as 10^6 planar faults/cm and 10^9 line defects/cm². These account for the low values of resistivity, mobility, and lifetime near the interface.

The surface electron mobility observed in SOS MOSFETs [250 – 350 cm²/(V · s)] is lower than in bulk devices [600 – 700 cm²/(V · s)]. This is a result of both the high defect density found in as-grown SOS films and the compressive stress measured in the silicon film. This stress increases the effective mass of the electrons, which decreases their mobility. The effective mass of holes, on the other hand, is smaller in SOS than in bulk silicon, because of the same compressive stress. The hole surface mobility, which could in principle be higher than in bulk because of the compressive stress, is, however, affected by the presence of defects, such that the value of surface

mobility for holes in SOS p-channel MOSFETs is comparable to that in bulk silicon. The minority carrier lifetime found in as-grown SOS films is a fraction of a nanosecond.

Several techniques have been developed to reduce both the defect density and the stress in the SOS films, such as the solid-phase epitaxy and regrowth (SPEAR) and the double solid-phase epitaxy (DSPE) techniques.³ These techniques employ the following steps. First, silicon implantation is used to amorphize the silicon film, with the exception of a thin superficial layer, where the original defect density is lowest. Then a thermal annealing step is used to induce solid-phase regrowth of the amorphized silicon, the top silicon layer acting as a seed. A second silicon implant is then used to amorphize the top of the silicon layer, which is subsequently recrystallized in a solid-phase regrowth step using the bottom of the film as a seed. In the SPEAR process, an additional epitaxy step is performed after solid-phase regrowth. Using such techniques, substantial improvement of the defect density is obtained. Noise in MOS devices is reduced, and the minority carrier lifetime is increased by two to three orders of magnitude, up to 50 ns. Typical improvements brought about by the DSPE process are an increase of the electron mobility from 300 to 450 cm²/(V · s) and an increase of the hole mobility from 185 to 250 cm²/(V · s). More recently, thin (0.1–0.2 μm), high-quality SOS films have been produced, and MOSFETs with excellent performances have been fabricated in these films.⁴ Field-effect mobilities of 800 and 250 cm²/(V · s) have been reported for electrons and holes, respectively, in devices made in these thin SOS films.

5.2.2 Laser Recrystallization

Silicon deposited on an amorphous substrate such as SiO₂ (typically, an oxidized silicon wafer) is polycrystalline. MOS transistors fabricated in such a polysilicon layer exhibit surface mobility values, on the order of 10 cm²/(V · s), as well as high threshold voltages (several volts) due to the high density of surface states (several 10¹² cm⁻²) present at grain boundaries. To obtain good device properties, it is, therefore, suitable to transform the polycrystalline layer into single-crystal material to eliminate grain boundaries. This can be achieved by melting the silicon film using a focused laser beam and recrystallize it in such a way that a continuous, single-crystal silicon film is obtained.

Continuous-wave (CW) lasers such as CO₂, Ar, and YAG:Nd lasers can be used for producing SOI films. CW Ar lasers, with two main spectral lines at 488.0 and 514.5 nm (blue and green) are the most widely used lasers. Indeed, these wavelengths are well absorbed by silicon. In addition, the reflectivity of silicon increases abruptly once melting is reached. This effect is very convenient since it acts as negative feedback on the power absorption and prevents the silicon from overheating above the melting point.

Laser recrystallization of a polysilicon films deposited on an amorphous SiO₂ substrate produces elongated crystals with a width of a few micrometers and a length of 10–20 μm. The grains are still separated by grain boundaries that have detri-

mental effects on device properties. These crystallites have a random crystal orientation. This material is clearly unacceptable for device fabrication.

Ideally, one wants a uniform (100) orientation for all crystallites. From this comes the idea of opening a window (seeding area) in the insulator to allow contact between the silicon substrate and the polysilicon layer. On melting and recrystallization, lateral epitaxy can be induced and the recrystallized silicon presents a uniform (100) orientation.

To obtain a large single-crystal area, the laser beam has to be raster-scanned on the wafer with some overlap between the beam scans. Unfortunately, small random crystallites arise at the edges of the large crystals, which precludes the formation of large single-crystal areas, and grain boundaries are formed between the single-crystal stripes. The location of these grain boundaries depends on the scanning parameters and the stability of the beam. In other words, from a macroscopic point of view, the location of the boundaries is quasirandom, and the yield of large circuits made in this material will be zero. A solution to this problem is to use stripes of an anti reflecting (AR) material (SiO_2 and/or Si_3N_4) to obtain the photolithographically controlled beam shaping of an otherwise Gaussian beam.⁵ When such a technique is used, chipwide (several mm \times several mm), defect-free, (100)-oriented single-crystal areas can be produced. The laser recrystallization technique has been used mostly for the fabrication of three-dimensional (3D) integrated circuits, where the fabrication of up to four active device layers has been demonstrated.^{6,7}

5.2.3 Zone-Melting Recrystallization

One of the main limitations of laser recrystallization is the small molten zone produced by the focused beam, which results in a long processing time to recrystallize a whole wafer. Recrystallization of a polysilicon film on an insulator can also be carried out using incoherent light (visible or near IR) sources. In this case, a narrow (a few millimeters) but long molten zone can be created on the wafer. A molten zone length of the size of an entire wafer diameter can readily be obtained. As a result, full recrystallization of a wafer can be carried out in a single pass. Such a recrystallization technique is generally referred to as *zone-melting recrystallization* (ZMR) because of the analogy between this technique and the float-zone refining process used to produce silicon ingots. An excellent review of the ZMR mechanisms can be found in the book of Givargizov.⁸

In the ZMR process, the back of the wafer is heated by a bank of halogen lamps of a heating susceptor up to a temperature close to the melting point of silicon. Additional heating is produced locally at the surface of the wafer using either a heated graphite strip located a few millimeters above the sample and scanned across it,⁹ a linear halogen or mercury lamp whose light is focused on the sample by means of an elliptical reflector,¹⁰ or an elongated laser spot, in which case a linear molten zone can be created using a high-power (300-W) CW YAG:Nd laser (wavelength = 1.06 μm) and cylindrical lenses.¹¹

SOI films have a (100) normal orientation, even if no seeding is used. Grain boundary-like defects repeated at 100 μm to several mm intervals are found in the

recrystallized films. These defects are called *subgrain boundaries*. These are low-angle boundaries between adjacent (100) crystals, or dislocation networks. In the best cases, the dislocation networks can be reduced to a few isolated dislocations. This type of defect induces no noticeable degradation in the devices fabricated in ZMR SOI films.

5.2.4 Epitaxial Lateral Overgrowth

Silicon-on-insulator material can be produced by homoepitaxial growth of silicon on silicon, provided the crystal growth can extend laterally on an insulator (SiO_2 , typically). This can be achieved using a classic epitaxy reactor.

The epitaxial lateral overgrowth technique (ELO) consists in the epitaxial growth of silicon from seeding windows over SiO_2 islands. It can be performed in an atmospheric or in a reduced-pressure epitaxial reactor. Epitaxy proceeds from the seeding windows both vertically and laterally, and the silicon crystal is limited by (100) and (101) facets. When two growth fronts, seeded from opposite sides of the oxide, join together, a continuous silicon-on-insulator film is formed, which contains a low-angle subgrain boundary where the two growth fronts meet. The major disadvantage of the ELO technique is the nearly 1 : 1 lateral : vertical growth ratio, which means that a 10- μm -thick film must be grown to cover 20- μm -wide oxide patterns (10 μm from each side). Furthermore, 10 additional micrometers must be grown in order to get a planar surface. Thinner SOI films can, however, be obtained by polishing the wafers after the growth of a thick ELO film.¹² The ELO technique has been used to fabricate 3D dual-gate devices.¹³

A variation of the ELO technique, called *tunnel epitaxy, confined lateral selective epitaxy* (CLSEG), or *pattern-constrained epitaxy* (PACE), has been reported by several groups.^{14,15} In this technique, a “tunnel” of SiO_2 is created, which forces the epitaxial silicon to propagate laterally. With this method, a 7 : 1 lateral : vertical growth ratio has been obtained.

5.2.5 SIMOX

The acronym SIMOX stands for separation by implanted oxygen.” The principle of SIMOX material formation is very simple (Fig. 5.1), consisting in the formation of a buried layer of SiO_2 by implantation of oxygen ions beneath the surface of a silicon wafer. The structure is then annealed at 1320°C for 6 h, typically. The buried oxide layer is often referred to as “BOX.”

Processing conditions must be such that a single-crystal overlayer of silicon is maintained above the oxide. It is worth noticing that in conventional microelectronics, ion implantation is used to introduce atoms into silicon at the impurity level. In the case of SIMOX, ion implantation is used to synthesize a new material, namely SiO_2 . This means that two atoms of oxygen have to be implanted for every silicon atom in the region where silicon dioxide has to be formed. In other words, the implanted dose required to form a BOX layer has to be 200–500 times higher than the heaviest doses commonly used in microelectronics processing.

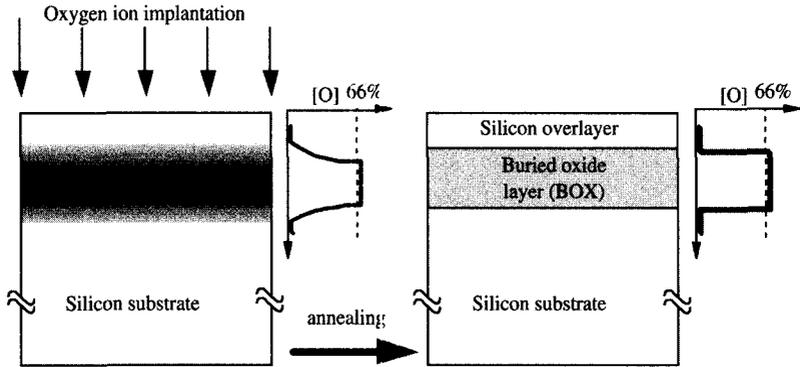


Figure 5.1 The SIMOX principle. A high dose of oxygen is implanted into silicon followed by an annealing step to produce a buried layer of silicon dioxide below a thin, single-crystal silicon overlayer. The oxygen concentration profile is also shown.

Experiments show that a dose of 1.4×10^{18} oxygen atoms cm^{-2} must be implanted to create a continuous buried oxide layer. The standard dose which is most common is $1.8 \times 10^{18} \text{ cm}^{-2}$, and produces a 400-nm-thick buried oxide layer upon annealing. At lower implant doses, a skewed Gaussian oxygen profile is obtained. When the dose reaches $1.2\text{--}1.4 \times 10^{18} \text{ cm}^{-2}$, stoichiometric SiO_2 is formed (66 at% of oxygen for 33 at% of silicon), and further implantation does not increase the peak oxygen concentration, but rather broadens the overall profile (i.e., the buried oxide layer becomes thicker). This is possible because the diffusivity of oxygen in SiO_2 ($10^{17} \text{ cm}^2/\text{s}$ at 500°C) is high enough for the oxygen to readily diffuse to the Si– SiO_2 interface where oxidation occurs. The dose at which the buried oxide starts to form ($\cong 1.4 \times 10^{18} \text{ cm}^{-2}$) is called the *critical dose*. Recent developments include low-dose implantation for production of thin buried oxides and silicon overlayers with a low defect density. The research for producing SIMOX material using low-dose implantation is also driven by economic reasons, since the production costs of SIMOX material are proportional to the dose used for the implantation. The low-dose SIMOX is obtained by implanting O^+ ions at a specific dose located in a very narrow window around $4 \times 10^{17} \text{ cm}^{-2}$. With a single implantation and a 6-h anneal at 1320°C , a continuous BOX having a thickness of 80 nm is formed.¹⁶ At a doses of $3 \times 10^{17} \text{ cm}^{-2}$, isolated oxide precipitates are formed. For a dose of $5 \times 10^{17} \text{ cm}^{-2}$, silicon precipitates form in the BOX. Only doses within a very narrow process window around $4 \times 10^{17} \text{ cm}^{-2}$ produce a continuous, precipitate-free BOX.

The silicon overlayer of SIMOX material contains dislocations due to the heavy ion-implantation conditions. Dislocation densities ranging within $10^3\text{--}10^4 \text{ cm}^{-2}$ and $10^2\text{--}10^3 \text{ cm}^{-2}$ are observed in standard and low-dose SIMOX materials, respectively. Some of the characteristics of modern SIMOX material are listed in Table 5.1, as well as those that appear to be within reach in a near future.¹⁷

TABLE 5.1 Characteristics of Modern SIMOX Material

Parameter	Current ^a	Future	Unit
Wafer diameter	100–200	200	mm
Silicon film thickness	100–200	50	nm
Silicon film thickness uniformity	<10	<5	nm
Buried oxide (BOX) thickness	400	400 or 80	nm
Buried oxide (BOX) thickness uniformity	<20	<4	nm
Surface roughness	0.3	0.2	nm
Dislocation density	<10 ⁵ ; ^b <300 ^c	<100	cm ⁻²
Metallic contamination	<10 ¹¹	<5 × 10 ¹⁰	cm ⁻²

^aYear: 1997.

^bStandard SIMOX.

^cLow-dose SIMOX.

5.2.6 Wafer Bonding and Etchback

The principle of the bonding and etchback technique is very simple; two oxidized wafers are “mated”-or-“bonded” together. One of the wafers is subsequently polished or etched down to a thickness suitable for SOI applications. The other wafer serves as a mechanical substrate, and is called a “handle wafer” (Fig. 5.2). This technique for forming SOI material is often called *BESOI* (bond and etchback SOI).

The following paragraphs describe the bonding mechanism and etchback process. When two flat, hydrophilic surfaces such as oxidized silicon wafers are placed against one another, bonding naturally occurs, even at room temperature. The contacting forces are believed to be caused by the attraction of hydroxyl groups (OH)⁻ and H₂O molecules adsorbed on the two surfaces. This attraction can be significant enough to cause spontaneous formation of hydrogen bonds across the gap between the two wafers. This attraction propagates from a first site of contact across the whole wafer in the form of a “contacting wave” with a speed of several centimeters per second (cm/s).¹⁸

After the wafers have been bonded at room temperature, it is customary to anneal the structure to strengthen the bond. The bond strength increases with annealing temperature, and three phases of bond strengthening occurring at different

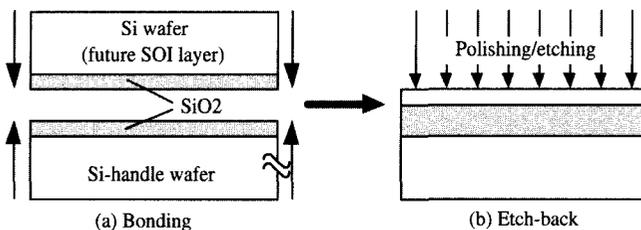


Figure 5.2 (a) Bonding of two oxidized silicon wafers; (b) polishing/etchback of one of the wafers.

temperatures can be distinguished. First, at temperatures above 200°C the adsorbed water separates from the Si–OH groups and forms tetramer water clusters. The second phase occurs at temperatures greater than 700°C when the bonds between hydroxyl groups start to be replaced by Si–O–Si bonds. Finally, at higher temperatures (1100°C and above), the viscous flow of the oxide allows for complete bonding of the wafers. The quality of bonding depends on the roughness of surfaces in contact. For example, it has been shown that much more uniform bonding is obtained if the roughness of the bonded wafers is less than 0.5 nm than if it is larger than that value.¹⁹

After bonding the wafers the top wafer has to be thinned down from a thickness of approximately 600 μm to a few micrometers or less to be useful for SOI device applications. Two basic thinning approaches can be used: grinding followed by chemical-mechanical polishing, and grinding followed by selective etchback. The etch stop is obtained by creating doping concentration gradients at the surface (i.e., right next to the oxide layer used for bonding) of the top wafer. For instance, in the double etch-stop technique, a lightly doped wafer is used, and a P⁺⁺ layer is created at its surface by ion implantation. Then, a lightly doped epitaxial layer is grown onto it. This epitaxial layer will be the SOI layer at the end of the process. After grinding, two chemical etch steps are used. The first one, either an ethylenediamine–pyrocatechol–water (EPW) etch, a potassium hydroxide solution, or a tetramethyl ammonium hydroxide (TMAH) solution etches the substrate and stops at the P⁺⁺ layer. Then, a 1 : 3 : 8 HF : HNO₃ : CH₃COOH etch is used to remove the P⁺⁺ layer. The combined selectivity of the etches is better than 10,000 : 1. The final thickness uniformity of the SOI layer depends on the uniformity of the silicon thickness grown epitaxially, as well as on the uniformity of the P⁺⁺ layer formation, but thickness standard deviations better than 12 nm can be obtained. An extensive review of the etch stop techniques can be found in Reference 20. Precision polishing techniques been developed as well to accurately control the thickness uniformity of the top silicon film.²¹ These make use of a small-area, computer-controlled polishing tool combined with continuous film thickness measurement. In the ÅcuThin process²² the SOI film is thinned using a scanning small-size plasma head. This technique, called *plasma-assisted chemical etching* (PACE) is used to planarize the silicon film and can produce 300-nm-thick SOI material with an rms thickness variation less than 2 nm ($\sigma = 1\text{--}1.6$ nm).²³

5.2.7 UNIBOND Material

The UNIBOND material is produced by the so-called Smart-Cut process.²⁴ It combines ion-implantation technology and wafer bonding techniques to transfer a thin surface layer from a wafer onto another wafer or an insulating substrate. The basic process steps of this technique are the following (Fig. 5.3).

- Ion implantation of hydrogen ions into an oxidized silicon wafer (wafer A). The implanted dose is on the order of 5×10^{16} cm⁻². At this stage micro cavities and micro bubbles are formed at a depth equal to the implantation

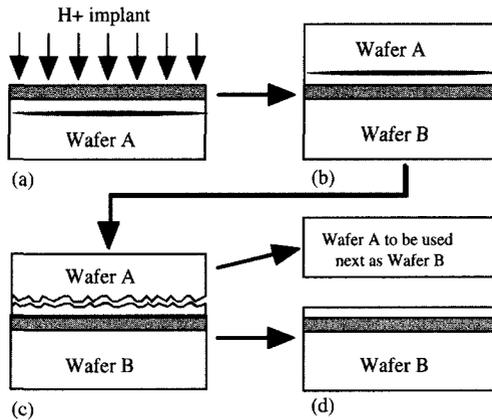


Figure 5.3 The Smart-Cut process: (a) hydrogen implant in oxidized wafer; (b) wafer bonding; (c) splitting of wafer A; (d) polishing both wafers. Wafer A is recycled as a future handle wafer.

range (R_p).²⁵ The wafer is capped preferably with thermally grown SiO_2 prior to implantation. This oxide layer will become the buried oxide of the SOI structure at the end of the process. During implantation the oxide becomes less hydrophilic due to some surface carbon contamination. It is, therefore, necessary to clean the wafers carefully before bonding to restore their hydrophilicity.²⁶

- Hydrophilic bonding of wafer A to a handle wafer (wafer B) is performed. Wafer B can either be bare or oxidized, depending on the final desired buried-oxide thickness.
- A two-phase heat treatment of the bonded wafers is then carried out. During the first phase, which takes place at a temperature of 500°C , the implanted wafer A splits into two parts: a thin layer of monocrystalline silicon that remains bonded onto wafer B, and the remainder of wafer A, which can be recycled for use as another wafer B. During the anneal, the average size of the microcavities increases. This size increase takes place along a $\langle 100 \rangle$ direction (i.e., parallel to the wafer surface) and an interaction between cavities is observed, which eventually results into the propagation of a crack across the whole wafer. This crack is quite parallel to the bonding wafer. The second heat treatment takes place at a higher temperature (1100°C) and is aimed at strengthening the bond between the handle wafer and the SOI film.²⁷
- Finally, chemomechanical polishing is performed on the SOI film to give it the desired mirrorlike surface. Indeed, this layer exhibits significant microroughness after wafer A splitting, such that a final touch-polish step is necessary. This polishing step reduces the surface roughness and consumes a few hundred angstroms of the SOI film.

The typical silicon film thickness provided by this process is 200 nm. The silicon film roughness after wafer splitting is better than 4 nm. It is better than 0.15 nm after the final polishing step. The film thickness uniformity is better than 10 nm (max-min, 200-mm wafer), and the dislocation density is lower than 10^2 cm^{-2} . The metal contamination is lower than $5 \times 10^{10} \text{ cm}^{-2}$. The minority carrier lifetime in the UNIBOND material is on the order of 100 μs , specifically, 10 times larger than in SIMOX.

From an economic point of view, the Smart-Cut process is significantly better than classical BESOI processes. Indeed, the Smart-Cut process requires only $n + 1$ starting wafers to produce n SOI wafers, while other BESOI processes require $2n$ wafers to produce n SOI wafers. In addition, this process can potentially be used to transfer a thin layer of *any* semiconductor material on top of an insulator. It is worthwhile noting that patterns etched on a silicon wafer can be transferred onto another wafer using the Smart-Cut process.²⁸

5.3 THE SOI MOSFET

SOI MOSFETs are free of some of the effects that tend to reduce the performance of their bulk counterparts. In particular they present reduced junction capacitances, a very small body factor, and, hence, a near-ideal subthreshold slope and high current drive. These different properties are detailed next.

5.3.1 Source and Drain Capacitance

In bulk MOS devices, the parasitic drain (or source)-to-substrate (or well) capacitance consists of two components: the capacitance between the junction and the substrate itself and the capacitance between the junction and the channel-stop implant under the field oxide. As devices are shrunk to smaller geometries, higher substrate doping concentrations are used and the junction capacitance increases.

In SOI devices with reach-through junctions, the junction capacitance has only one component: the capacitance of the MOS structure made of the junction (gate electrode of the MOS structure), the buried oxide (gate oxide of the MOS structure), and the underlying silicon substrate (substrate of the MOS structure). This parasitic capacitance can only be smaller than the capacitance of the buried oxide, which is typically lower than the capacitance junction of a bulk MOSFET by an order of magnitude. This difference further increases as the supply voltage is reduced (Fig. 5.4).²⁹

5.3.2 Fully Depleted, Partially Depleted, and Accumulation-Mode MOSFETs

All SOI MOSFETs are not alike. Their physics is highly dependent on the thickness of the silicon film on which they are made. Three types of devices can be

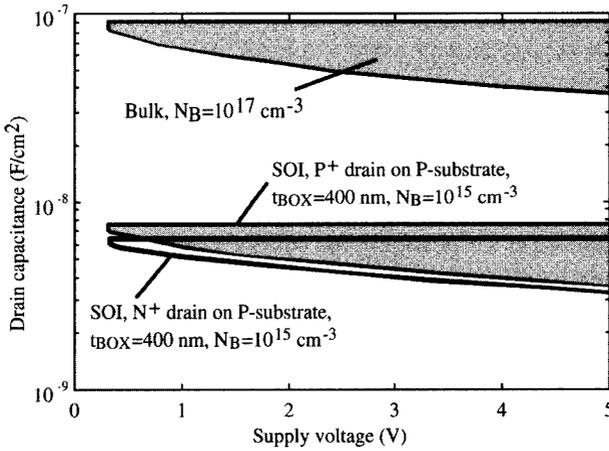


Figure 5.4 Parasitic junction capacitance per unit area as a function of supply voltage in bulk CMOS with a constant substrate doping concentration (10^{17} cm^{-3}) and in standard SIMOX with a substrate doping concentration of 10^{17} cm^{-3} .

distinguished, depending on both the silicon film thickness and the channel doping concentration: thick-film, thin-film, and “medium-thickness” devices. The medium thickness devices which can exhibit either a thin- or a thick-film behavior, depending on the back-gate bias. Figure 5.5 presents the band diagrams of a bulk, a thick-film SOI, and a thin-film SOI n-channel device at threshold.

In a *bulk* device (Fig. 5.5a), the depletion zone extends from the Si–SiO₂ interface to the maximum depletion width, x_{dmax} , which is typically given by $x_{dmax} = \sqrt{4\epsilon_{si}\Psi_F/qN_a}$, where Ψ_F is the Fermi potential, which is equal to $(kT/q) \ln(N_a/n_i)$.

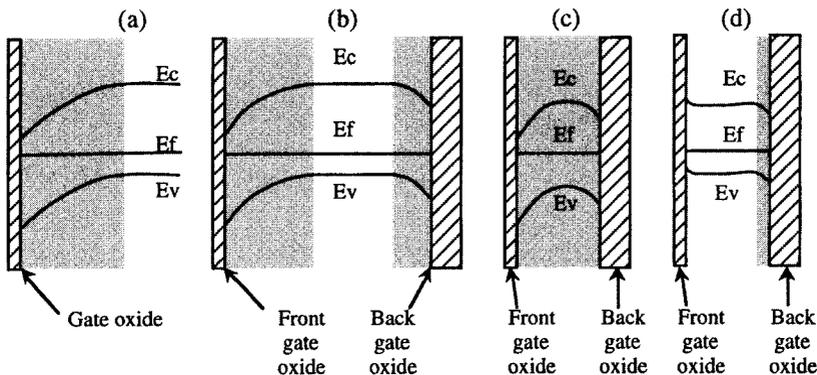


Figure 5.5 Band diagram in (a) an n-channel bulk, (b) a partially depleted, (c) a fully depleted, and (d) an accumulation-mode (p-channel) SOI device. All devices are represented at threshold (front-gate voltage = threshold voltage). The shaded areas represent the depleted zones.

In a *thick-film SOI* device (Fig. 5.5*b*), the silicon film thickness is larger than twice the value of x_{dmax} . In such a case, there is no interaction between the depletion zones arising from the front and the back interfaces, and there exists a piece of neutral silicon beneath the front depletion zone. Such a device is called partially depleted (PD). If this neutral piece of silicon, called “body,” is connected to ground by a “body contact” or “body tie,” the characteristics of the device will be exactly those of a bulk device. If, however, the body is left electrically floating, the device will basically behave as a bulk device, but with the notable exception of two parasitic effects—the first one is called the “kink effect” (a kink appears in the output characteristics of the device), and the second one is the presence of a parasitic, open-base, bipolar transistor between the source and the drain.

In a *thin-film SOI* device (Fig. 5.5*c*), the silicon film thickness is smaller than x_{dmax} . In this case, the silicon film is fully depleted at threshold, irrespective of the bias applied to the back gate (with the exception of the possible presence of thin accumulation or inversion layers at the back interface, if a large negative or positive bias is applied to the back gate, respectively). Such a device is called fully depleted (FD). Fully depleted SOI devices are virtually free of the kink effect, if their back interface is not in accumulation. Among all types of SOI devices, fully depleted devices with depleted back interface exhibit the most attractive properties, such as low electric fields, high transconductance, excellent short-channel behavior, and a near-ideal subthreshold slope.

In a thin SOI film it is also possible to realize accumulation-mode (AM) MOSFETs. If the workfunction of the gate material is properly chosen (e.g., N^+ polysilicon for a p-channel device with P^+ source and drain and P-type channel doping; Fig. 5.5*d*), the device is fully depleted in the OFF state, and no current flows between source and drain. When a gate voltage is applied to turn the device on, a quasineutral zone appears in the body of the silicon film, and a current, called “body current,” starts to flow. At higher gate voltages the depletion zone underneath the front gate disappears completely, and an accumulation channel is formed. Current thus flows from source to drain through both the silicon film body and the accumulation channel.

5.3.3 Threshold Voltage

The threshold voltage of an enhancement-mode bulk n-channel MOSFET is classically given by³⁰

$$V_{th} = V_{FB} + 2\psi_F + \frac{qN_a x_{dmax}}{C_{ox}} \quad (5.1)$$

where V_{FB} is the flat-band voltage, equal to $\Phi_{MS} - (Q_{ox}/C_{ox})$ (we will neglect the presence of fast surface states, N_{it} , in the present analysis), $\psi_F = (kT/q) \ln(N_a/n_i)$ is the Fermi potential, and x_{dmax} is the maximum depletion width, which is equal to $\sqrt{4\epsilon_{si}\psi_F/qN_a}$.

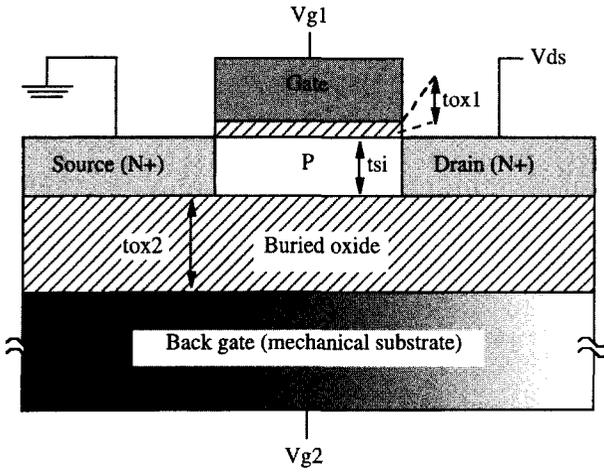


Figure 5.6 Cross section of a thin-film, n-channel fully depleted SOI MOSFET illustrating some of the notations used in this section.

In a PD SOI device, there is no interaction between the front and back depletion zones. In that case, the threshold voltage is the same as in a bulk transistor.

The threshold voltage of a FD n-channel SOI device³¹ (Fig. 5.6) can be obtained by solving the Poisson equation, using the depletion approximation: $d^2\Psi/dx^2 = qN_a/\epsilon_{si}$, which, when integrated, twice yields the potential as a function of depth in the silicon film, x :

$$\Psi(x) = \frac{qN_a}{2\epsilon_{si}}x^2 + \left(\frac{\Psi_{s2} - \Psi_{s1}}{t_{si}} - \frac{qN_a t_{si}}{2\epsilon_{si}} \right)x + \Psi_{s1} \quad (5.2)$$

where Ψ_{s1} and Ψ_{s2} are the potentials at the front and back silicon–oxide interfaces, respectively. The front- and back-gate voltages, V_{G1} and V_{G2} , are given by $V_{G1} = \Psi_{s1} + \Phi_{ox1} + \Phi_{MS1}$ and $V_{G2} = \Psi_{s2} + \Phi_{ox2} + \Phi_{MS2}$, where Φ_{MS1} , Φ_{MS2} are the front and back workfunction differences, respectively. Using these relationships, one can find the threshold voltage of the front interface, by assuming that $\Psi_{s1} = 2\Psi_F$.

If the back surface is in accumulation, Ψ_{s2} is pinned to approximately 0 V. The threshold voltage $V_{th1,acc2}$ is obtained from the expressions above, where $V_{th1,acc2} = V_{G1}$ is calculated at $\Psi_{s2} = 0$, $Q_{inv1} = 0$, and $\Psi_{s1} = 2\Psi_F$. The result is

$$V_{th1,acc2} = \Phi_{MS1} - \frac{Q_{ox1}}{C_{ox1}} + \left(1 + \frac{C_{si}}{C_{ox1}} \right) 2\Psi_F - \frac{Q_{depl}}{2C_{ox1}} \quad (5.3)$$

If the back surface is inverted, Ψ_{s2} is pinned to approximately $2\Psi_F$. The front threshold voltage $V_{th1,inv2}$ is obtained by posing $V_{th1,inv2} = V_{G1}$ with $\Psi_{s2} = 2\Psi_F$, $Q_{inv1} = 0$, and $\Psi_{s1} = 2\Psi_F$. The result is

$$V_{th1,inv2} = \Phi_{MS1} - \frac{Q_{ox1}}{C_{ox1}} + 2\Psi_F - \frac{Q_{depl}}{2C_{ox1}} \quad (5.4)$$

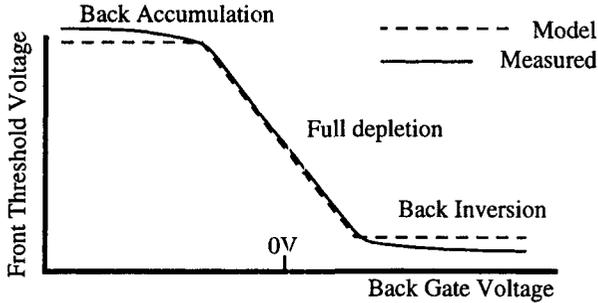


Figure 5.7 Variation of the front-gate threshold voltage with back-gate bias.

If the back surface is depleted, Ψ_{s2} depends on the back-gate voltage, V_{G2} , and its value can range between 0 and $2\Psi_F$. The value of back-gate voltage for which the back interface reaches accumulation (the front interface being at threshold), $V_{G2,acc}$, is obtained by posing $\Psi_{s1} = 2\Psi_F$ and $\Psi_{s2} = 0$. Similarly, the value of back-gate voltage for which the back interface reaches inversion, $V_{G2,inv}$, is given obtained by posing $\Psi_{s1} = 2\Psi_F$ and $\Psi_{s2} = 2\Psi_F$. When $V_{G2,acc} < V_{G2} < V_{G2,inv}$ the front threshold voltage is obtained by writing $\Psi_{s1} = 2\Psi_F$ and $Q_{inv1} = 0$. The result is

$$V_{th1,depl2} = V_{th1,acc2} - \frac{C_{si} C_{ox2}}{C_{ox1} (C_{si} + C_{ox2})} (V_{G2} - V_{G2,acc}) \quad (5.5)$$

The variation of the threshold voltage, V_{th1} , of a FD SOI MOSFET as a function of back-gate bias, V_{G2} , is presented in Figure 5.7.

In an AM MOSFET the accumulation threshold voltage is equal to the front flat-band voltage: $V_{th} = \Phi_{MS1} - (Q_{ox1}/C_{ox1}) = V_{fb1}$; however, there exists a second drain current component (body current) that flows through the body of the device and is dependent on the back-gate bias. As a result, the back-gate dependence of the threshold voltage of an AM device resembles that of a FD, enhancement-mode device.³²

5.3.4 Body Effect

In a bulk device, the body effect is defined as the dependence of the threshold voltage on the substrate bias. In an SOI transistor, it is similarly defined as the dependence of the threshold voltage on the back-gate bias.

In a bulk n-channel transistor, the threshold voltage can be expressed as

$$V_{th} = \Phi_{MS} + 2\Psi_F - \frac{Q_{ox}}{C_{ox}} + \frac{Q_b}{C_{ox}} \quad \text{with} \quad Q_b = \sqrt{2\varepsilon_{si}qN_a(2\Psi_F - V_B)} \quad (5.6)$$

which can be rewritten

$$V_{th} = \Phi_{MS} + 2\Psi_F - \frac{Q_{ox}}{C_{ox}} + \frac{\sqrt{2\varepsilon_{si}qN_a(2\Psi_F - V_B)}}{C_{ox}} \quad (5.7)$$

and, by defining $\gamma = \sqrt{2\varepsilon_{\text{si}}qN_a}/C_{\text{ox}}$, one obtains

$$V_{\text{th}} = \Phi_{\text{MS}} + 2\Psi_F - (Q_{\text{ox}}/C_{\text{ox}}) + \gamma\sqrt{2\Psi_F} + \gamma(\sqrt{2\Psi_F - V_B} - \sqrt{2\Psi_F}) \quad (5.8)$$

The last term depicts the dependence of threshold voltage on substrate bias, called the *body effect*. When a negative bias is applied to the substrate (with respect to the source), the threshold voltage increases as a square-root function of the substrate bias. If the threshold voltage with zero substrate bias is referred to as V_{th0} , one can write

$$V_{\text{th}}(V_B) = V_{\text{th0}} + \gamma(\sqrt{2\Psi_F - V_B} - \sqrt{2\Psi_F}) \quad (5.9)$$

where γ is the body-effect parameter (unit: $V^{1/2}$).

The body effect (or, more accurately, the back-gate effect) can be neglected ($\gamma = 0$) in PD SOI devices, because there is no coupling between front gate and back gate.

In a FD SOI device, the body effect is obtained by

$$\frac{dV_{\text{th1}}}{dV_{G2}} = -\frac{C_{\text{si}}C_{\text{ox2}}}{C_{\text{ox1}}(C_{\text{si}} + C_{\text{ox2}})} = \frac{-\varepsilon_{\text{si}}C_{\text{ox2}}}{C_{\text{ox1}}(t_{\text{si}}C_{\text{ox2}} + \varepsilon_{\text{si}})} = \gamma \quad (5.10)$$

The symbol γ is chosen by analogy with a bulk device. It should be noted that γ is dimensionless in the case of thin-film SOI transistors, and that the threshold voltage dependence on back-gate bias is linear.

It is practical to linearize the body effect in bulk devices. In that case one defines the *body factor*, denoted n , using the following expression: $n = 1 + C_D/C_{\text{ox}}$ with C_D , the depletion capacitance, being equal to $\varepsilon_{\text{si}}/x_{\text{dmax}}$. In the case of a FD SOI device, one obtains $n = 1 + [C_{\text{si}}C_{\text{ox2}}/C_{\text{ox1}}(C_{\text{si}} + C_{\text{ox2}})]$. Typical values for the body factor are $n = 1.3-1.5$ for a bulk device and $n = 1.05-1.1$ for a fully depleted SOI device.

5.3.5 Output Characteristics and Transconductance

The output characteristics of a FD SOI MOSFET are identical to those of a bulk device. In both types of devices the saturation current is given by the following expression:

$$I_{\text{Dsat}} \cong \frac{W\mu_n C_{\text{ox1}}}{2nL} [V_{G1} - V_{\text{th}}]^2 \quad (5.11)$$

where W , L , and μ_n are the channel width and length, and the electron surface mobility, respectively. Because the body effect is smaller in FD SOI than in bulk devices, a higher drain current can be obtained.

The output characteristics of PD SOI devices present a kink due to impact ionization phenomena taking place near the drain. This so-called “kink effect” can

be eliminated using contacts to the otherwise electrically floating body of the transistor, underneath the gate.

The transconductance of a MOSFET, g_m , is a measure of the effectiveness of the control of the drain current by the gate voltage. In a bulk n-channel MOSFET in saturation, it is given by³³

$$g_m = \frac{dI_{Dsat}}{dV_G} \quad (\text{for } V_{DS} > V_{Dsat})$$

$$= \frac{W}{L} \mu_n C_{ox} (V_G - V_{th}) + 4\Psi_F \left(\frac{C_D}{C_{ox}} \right)^2 \left(1 - \sqrt{1 + \left(\frac{C_{ox}}{C_{depl}} \right)^2 \frac{V_G - V_{FB}}{2\Psi_F}} \right) \quad (5.12)$$

with $C_D = \epsilon_{si}/x_{dmax}$, which can be approximated by

$$g_m = \frac{dI_{Dsat}}{dV_{G1}} = \frac{W\mu_n C_{ox1}}{nL} (V_{G1} - V_{t1}) \quad \text{with } n = 1 + \frac{\epsilon_{si}}{x_{dmax} C_{ox}} \quad (5.13)$$

In fully depleted SOI MOSFET, the transconductance can be obtained from in a similar way:

$$g_m = \frac{W\mu_n C_{ox1}}{nL} (V_{G1} - V_{th}) \quad (5.14)$$

As in the case of the drain current a higher transconductance can be obtained from a FD SOI MOSFET than from a bulk device, because of the smaller body effect of the SOI device.

The maximum voltage gain of a MOSFET is obtained when the value of g_m/I_D is largest, since the gain of a transistor is given by³⁴

$$\frac{\Delta V_{out}}{\Delta V_{in}} = \frac{\Delta I_d}{g_D} \frac{1}{\Delta V_{in}} = \frac{\Delta V_{in} g_m}{g_D} \frac{1}{\Delta V_{in}} = \frac{g_m}{g_D} = \frac{g_m}{I_D} V_A \quad (5.15)$$

where g_D is the output drain conductance and V_A is the Early voltage. The Early voltage, V_A , of a FD SOI transistor is basically identical to that of a bulk device.

The largest value of g_m/I_D appears in the weak inversion regime for MOS transistors.³⁵ The value of g_m/I_D can be rewritten

$$\frac{g_m}{I_D} = \frac{dI_D}{I_D dV_G} = \frac{\ln(10)}{S} = \frac{q}{nkT} \quad (5.16)$$

where S is the subthreshold slope (see next section) and n is the body factor. In strong inversion, g_m/I_D becomes³⁶

$$\frac{g_m}{I_D} = \sqrt{\frac{2\mu C_{ox} W/L}{nI_D}} \quad (5.17)$$

Because of the lower value of n in fully depleted SOI MOSFETs, values of g_m/I_D significantly higher than in bulk devices can be obtained ($g_m/I_D = 20\text{--}25V^{-1}$ in a bulk device and $g_m/I_D = 30\text{--}35V^{-1}$ in a FD SOI device, typically).

5.3.6 Subthreshold Slope

The inverse subthreshold slope (or, in short, the subthreshold slope, or subthreshold swing) is defined as the inverse of the slope of the $I_D(V_G)$ curve in the subthreshold regime, presented on a semilogarithmic plot:

$$S = \frac{dV_G}{d(\log I_D)} \quad (5.18)$$

Neglecting the presence of traps at the Si–SiO₂ interface, the subthreshold slope of a bulk device is given by³⁷

$$\begin{aligned} S &= \frac{kT}{q} \ln(10) \left(1 + \frac{C_D}{C_{ox}} \right) = n \frac{kT}{q} \ln(10) \\ &= n \times 60 \text{ mV/decade at room temperature} \end{aligned} \quad (5.19)$$

In a FD SOI device, the subthreshold slope is given by the same expression; the only difference is the smaller value of the body factor. AM SOI devices present subthreshold slope values close to those of FD devices, although their physics is quite different because the subthreshold current flows in the body of the silicon film instead of at its surface. Figure 5.8 shows the subthreshold characteristic of a FD n-channel and an AM p-channel SOI transistors. Because of the sharp subthreshold slopes, low OFF current and low threshold-voltage values can be obtained simultaneously.

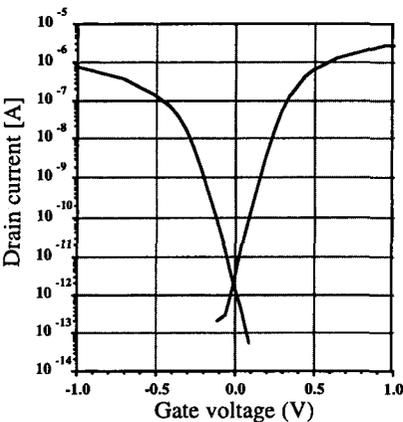


Figure 5.8 Measured current versus gate voltage for an n-channel and a p-channel SOI MOSFET with $W = L = 3 \mu\text{m}$ at $V_{DS} = \pm 50 \text{ mV}$. Because of the sharp subthreshold slope, OFF currents less than $1 \text{ pA}/\mu\text{m}$ and low threshold voltage (0.4 V) can be obtained simultaneously.

5.3.7 Microwave MOSFETs

It has been demonstrated that high-frequency n- and p-channel MOSFETs with transition frequencies above 10 GHz can be fabricated in SOI CMOS. In addition, the use of high-resistivity SOI substrates ($5000 \Omega \cdot \text{cm}$ or higher) allows the fabrication of passive elements, such as strip or slot lines with relatively low losses, and planar inductors with a relatively good-quality factor. Rectangular planar inductors with quality factors of 5, 8, and 11 have been obtained when realized on 20-, 4000-, and 10,000- $\Omega \cdot \text{cm}$ substrates, respectively. Some of the highest-performance SIMOX microwave MOSFETs were fabricated using a dedicated MOS process, called MICROX, which uses nonstandard CMOS features, such as a metal (gold) gate. In other approaches, a standard CMOS SOI process was used. The latter devices are, therefore, compatible with lower-frequency (base band) analog and digital circuits fabricated using the same process. For correct microwave operation of a MOSFET the gate sheet resistivity must be low. If this is not the case, the gate behaves like a delay line for the input signal and the part of the transistor which is farthest from the gate contact does not respond to high frequencies.³⁸ In addition, a low gate resistance is crucial for obtaining a low noise factor. Therefore, silicided gates (or metal-covered gates as in the case of MICROX) must be used. Table 5.2 presents the performance of some SOI microwave MOSFETs.^{39–46}

Figure 5.9 shows the high-frequency performance of an n-channel, fully depleted device. The silicided gate has a sheet resistance of $2.9 \Omega/\text{square}$. To obtain optimized microwave performances, the $240\text{-}\mu\text{m}$ gate width of the transistors is obtained using a comblike design of the polysilicon gate, which is composed of 10 fingers having a length of $24 \mu\text{m}$ each. Figure 5.9 presents the current gain (H_{21}), the maximum available gain (MAG), and the unilateral gain (ULG) as a function of frequency in

TABLE 5.2 Performances and Characteristics of Microwave n-Channel SOI MOSFETs^a

SOI Material	L (μm)	V_D (V)	I_D (mA)	f_T (GHz)	f_{max} (GHz)	Noise Figure/ Associated Gain (dB) at 2 GHz
BESOI	1	—	—	—	14	5/6.4
SIMOX	1	—	—	—	11	5/4.4
SIMOX ^b	0.32	3	33	14	21	3/13.4
SIMOX	0.25	3	41	23.6	32	1.5/17.5
SIMOX	0.75	0.9	3	10	11	1.5/9
SIMOX	0.75	0.9	10	12.9	30	-/13.9 (10.4 ^c)
SIMOX	0.3	2	—	—	24.3	0.9/14
SIMOX ^b	0.2	2	125	28.4	46	1/15.3 ^d
SIMOX	0.07	1.5	5	150	—	-/-

^aT-gate technology is used.

^bWith metal shunt on the gate.

^cAt 5 GHz.

^dAt 3 GHz.

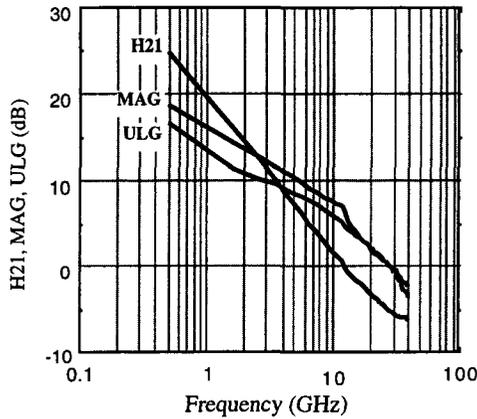


Figure 5.9 Current gain (H_{21}), the maximum available gain (MAG), and the unilateral gain (ULG) as a function of frequency in an n-channel SOI transistor, having a length of $0.75\ \mu\text{m}$ and a width of $240\ \mu\text{m}$ at $V_{DS} = V_{GS} = 1\ \text{V}$.

an n-channel SOI transistor having a length of $0.75\ \mu\text{m}$ ($L_{\text{eff}} = 0.65\ \mu\text{m}$). These parameters were measured through s-parameter extraction under a supply voltage ($V_{DS} = V_{GS}$) of $1\ \text{V}$. The MAG is $13\ \text{dB}$ at $2\ \text{GHz}$. The unit-gain frequency, f_T (found when $H_{21} = 0\ \text{dB}$) is equal to $12.9\ \text{GHz}$, and the maximum oscillation frequency, f_{max} (found when $\text{ULG} = 0\ \text{dB}$), is equal to $30\ \text{GHz}$.

Finally, it is worth mentioning that the use of an SOI instead of bulk allows for significant reduction of crosstalk in the chips and that crosstalk immunity is enhanced if high-resistivity substrates are used.⁴⁷ This feature is particularly attractive for mixed-mode circuits containing both base-band and RF parts.

5.4 3D AND NOVEL SOI DEVICES

The dielectric isolation of SOI devices, in contrast to the junction isolation of bulk devices, and the presence of second gate underneath the transistors allows one to realize novel structures with interesting properties. Hybrid bipolar MOS devices, dual-gate MOSFETs, special bipolar transistors, and thin-film, quantum-effect SOI devices have been demonstrated and are described in this section.

5.4.1 Bipolar–MOS “Hybrid” Device

Every enhancement-mode SOI MOSFET contains a parasitic bipolar transistor. The hybrid bipolar–MOS device controls the bipolar effect and makes use of the combined current drive capabilities of both the bipolar and the MOS parts of a “normally MOS” device.⁴⁸ This is achieved by connecting the gate, which controls the current flow in the MOS part of the device, to the floating substrate, which acts as the base of the lateral bipolar transistor (Fig. 5.10). The source and the drain of the

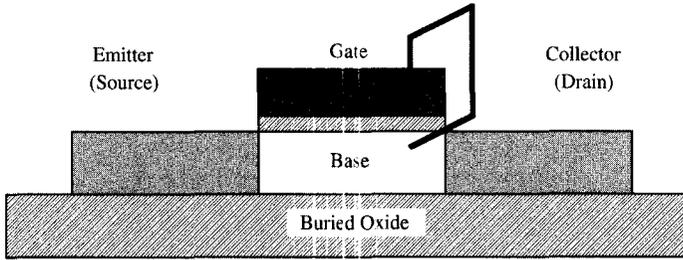


Figure 5.10 Representation of a hybrid bipolar-MOS hybrid device.

MOS transistor are also the emitter and the collector of the bipolar device, respectively.

We now consider the case of an n-channel (NPN) device, although p-channel (PNP) devices can be realized as well. When the device is OFF ($V_G = V_B = 0$), the potential of the MOS substrate (= the base) is low, which maximizes the value of the threshold voltage (and, thereby, minimizes the OFF current). On the other hand, when a gate bias is applied, the potential of the MOS substrate is increased, which decreases the threshold voltage. This lowering of the threshold voltage increases the current drive for a given gate voltage, compared to an MOS transistor without a gate-to-body connection. Similarly, the application of a gate voltage when the device is ON increases the collection efficiency of the bipolar device, and reduces the effective neutral base width of the device, thereby improving the bipolar gain. This effect is illustrated in Figure 5.11. In absence of any gate bias (we assume a flat-band condition at the Si-SiO₂ interfaces), the minority-carrier (electrons) current injected from the emitter has to diffuse across the width of the neutral base before being collected by the depletion zone near the collector (Figure 5.11a). The gain of the bipolar transistor, β , can be approximated by $\beta \cong 2(L_n/L_B)^2 - 1$ where L_n is the minority-carrier (electrons in this case) diffusion length in the base and L_B is the base width. When a positive gate voltage is applied, an inversion channel is created. This channel allows the MOS current component, I_{ch} , to flow from source to drain. Two components of bipolar current can be distinguished—the electrons injected by the emitter can either be attracted by the surface potential created by the gate bias and

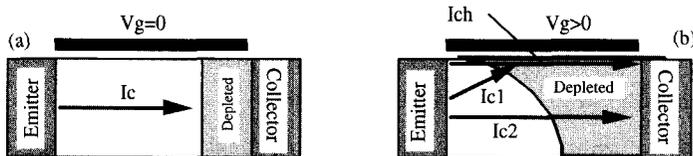


Figure 5.11 Emitter-to-collector current, I_c , (a) flowing from emitter to collector in the lateral bipolar transistor with no applied gate bias, and (b) the different current components flowing from source (emitter) to drain (collector) when a gate bias is applied to both gate and base: the MOS channel current, I_{ch} , a bipolar current collected by the channel, I_{c1} , and a bipolar current component flowing from emitter to collector, I_{c2} .

join the channel electrons (I_{C1} component) or diffuse directly from emitter to collector through the base (I_{C2} component). In both cases, the path traveled by the electrons is shorter than if no gate bias were applied (Fig. 5.11*b*). Indeed, the widening of the depletion zone caused by the gate bias reduces the effective neutral base length. The even shorter path traveled by the electrons injected from the emitter and joining the channel current is also obvious.

As a result, from these considerations, the gain of the bipolar device is increased when a positive gate voltage is applied, which is always the case in the bipolar MOS device, since the gate is connected to the base. Recently (at the time of writing), this principle of gain enhancement has been applied to bulk lateral bipolar transistors.⁴⁹ In summary, the presence of a gate improves the gain of the bipolar transistor, and the presence of a base contact improves both the ON and OFF characteristics of the device. Such a mutually beneficial phenomenon could almost be called "symbiosis." An analytical model of the device can be found in the literature.⁵⁰

The electrical characteristics of the bipolar-MOS device are presented in Figures 5.12 (NPN device) and 5.13 (quasi-PNP device). Such devices can provide common emitter current gains of 10,000 for $L = 0.3 \mu\text{m}$.⁵¹⁻⁵³

Low-power, low-voltage CMOS ring oscillators have been fabricated using complementary hybrid MOS-bipolar devices.^{54,55} These were shown to operate with supply voltages ranging from 0.5 to 1 V. Very low power dissipation is observed (Figs. 5.14 and 5.15).

More recently, a 0.5-V SIMOX CMOS circuit (8 K-gate ALU) has been realized using hybrid MOS-bipolar transistors [called *multithreshold CMOS* (MTCMOS) in Refs. 56 and 57]. The gate delay and the clock frequency are 200 ps and 40 MHz, respectively, for a supply voltage of 0.5 V. The standby (sleep mode) power dissipation of the ALU is 5 nW and the power consumption during 40-MHz operation is 350 μW .

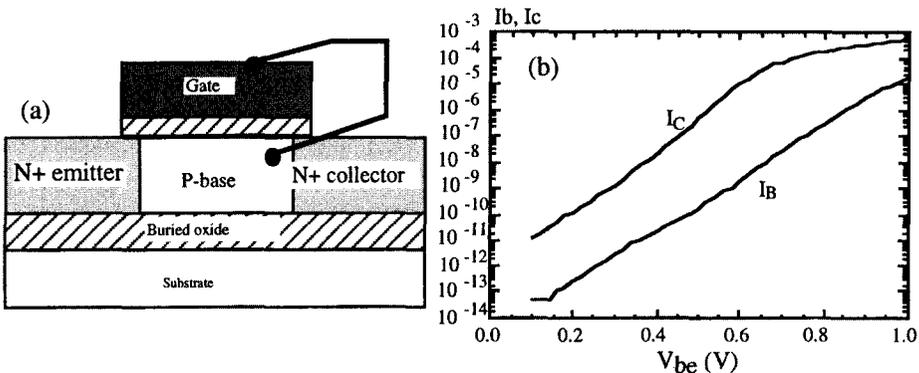


Figure 5.12 (a) Cross section of a hybrid bipolar-MOS fully depleted SOI NPN device and (b) the measured Gummel plot ($W/L_{\text{eff}} = 20 \mu\text{m}/0.6 \mu\text{m}$, $t_{\text{si}} = 80 \text{ nm}$, $t_{\text{ox}} = 30 \text{ nm}$, $\beta_F = 5000$ at $I_C = 10 \mu\text{A}$).

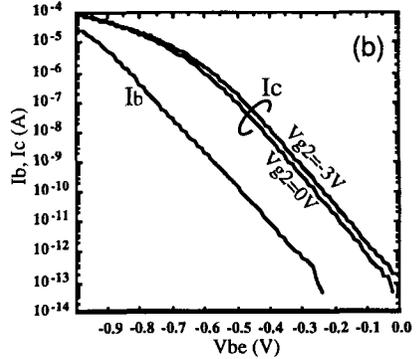
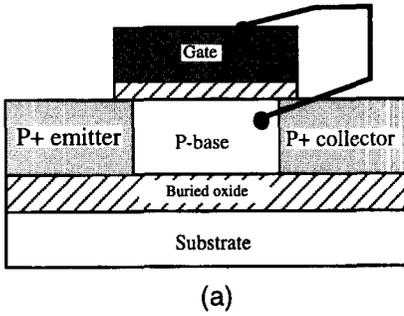


Figure 5.13 Cross section of hybrid bipolar – MOS lateral P⁺PP⁺ device (left) and measured Gummel plot (right) for back-gate voltage values of 0 and –3 V. Gate is tied to base. $V_{CE} = -1$ V and $W/L_{eff} = 20 \mu\text{m}/1.1 \mu\text{m}$.

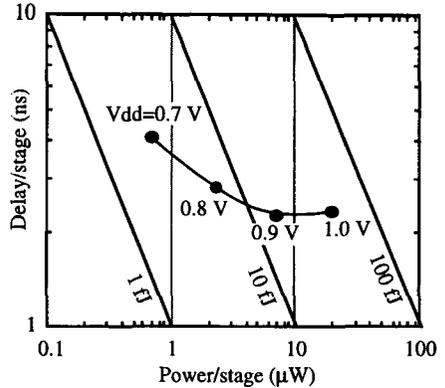


Figure 5.14 Delay per stage versus power per stage in a hybrid complementary MOS–bipolar complementary ring oscillator ($L_{eff,N} = 0.45 \mu\text{m}$, $L_{eff,P} = 0.58 \mu\text{m}$).

5.4.2 Dual-Gate MOSFET

SOI MOSFETs always have two gates. The top gate is used to control the carrier concentration in the top channel, while the back gate is usually grounded. Usually, back-channel conduction is avoided, and the back interface is kept in depletion or accumulation. The advantages of fully depleted SOI devices are well known. Most of them derive from the excellent coupling between the front gate and the front surface potential. This excellent coupling allows for a reduced body factor. The lower the body factor, the more ideal the characteristics of the device, the lowest possible value being $n = 1$. This value is not completely reached in fully depleted SOI transistors because of the capacitor divider formed by the gate oxide, the silicon film, and the buried oxide capacitances. The only ways to obtain a perfect coupling between the surface potential in the channel region and the gate are either to use an infinitely thick buried dielectric or to have the back gate connected to the front gate, with a back-gate oxide thickness equal to the front one. The latter solution defines

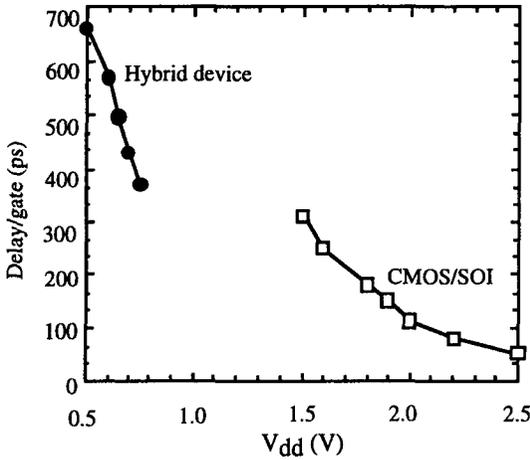


Figure 5.15 Delay per stage versus V_{dd} in a hybrid complementary MOS–bipolar ring oscillator ($L_{eff,N} = L_{eff,P} = 0.3 \mu\text{m}$). The data for an equivalent SOI CMOS ring oscillator are shown for comparison.

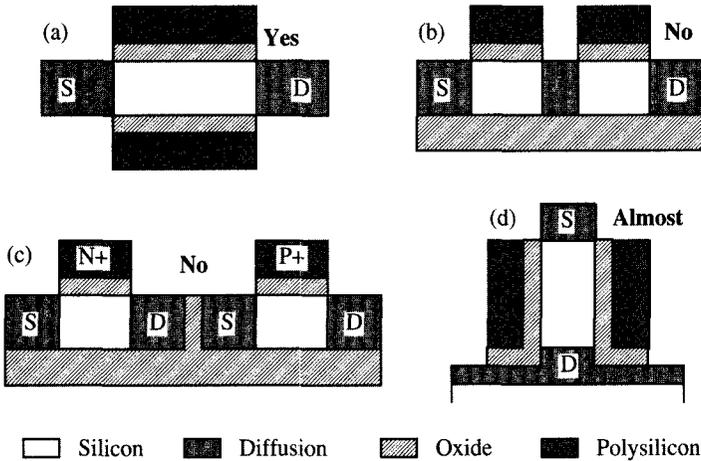


Figure 5.16 Which is a dual-gate SOI MOSFET? (a) A genuine SOI dual-gate MOSFET; (b) the twin-gate MOSFET⁵⁸ is not a dual-gate MOSFET; (c) the SOI CMOS inverter with dual polysilicon gate (N^+ and P^+ doping) is not a dual-gate MOSFET; (d) the pillar-shaped vertical transistor with surrounding gate has the same physics as the dual-gate transistor, but it is not an SOI device.

the concept of the dual-gate SOI MOSFET. Figure 5.16 describes this device: the front and back gates are symmetrical (the same gate oxide thickness is used) and tied together electrically. One should not be confused by other devices that bear names similar the dual-gate MOSFET, such as the twin-gate MOSFET and dual-polysilicon-gate MOSFETs. The pillar-shaped vertical transistor with a surrounding

gate has the same physics as the dual-gate SOI transistor, but it is not an SOI device.

One of the first publications on the dual-gate transistor concept dates back to 1984.⁵⁹ It shows that one can obtain significant reduction of short-channel effects in a device, called *XMOS*, in which an excellent control of the potential in the silicon film is achieved by using a top-and-bottom gate. The name of the device comes from its resemblance with the Greek letter Ξ . Using this configuration, better control of the channel depletion region is obtained than in a “regular” SOI MOSFET, and, in particular, the influence of the source and drain depletion regions are kept minimal. This reduces the short-channel effects by screening the source and drain field lines away from the channel. More complete modeling, including Monte Carlo simulations, is presented in Ref. 60, in which the ultimate scaling of silicon MOSFETs is explored. According to that paper, the ultimate silicon device is a dual-gate SOI MOSFET with a gate length of 30 nm, an oxide thickness of 3 nm, and a silicon film thickness of 5–20 nm. Such a (simulated) device shows no short-channel effects for gate lengths larger than 70 nm, and provides transconductance values of up to 2300 mS/mm.

One problem with such thin devices is the splitting of the conduction band into subbands. The energy minimum of the first subband controls the threshold voltage and is dependent on the silicon film thickness, according to the equation: $\sigma_{VT} = (h^2\pi^2/qm^*t_{si}^3)\sigma_{tsi}$, where σ_{VT} is the threshold uncertainty and σ_{tsi} is the silicon film thickness uncertainty. The latter result is similar to the quantum-mechanical increase of threshold voltage in ultrathin SOI devices.⁶¹

Fully depleted SOI MOSFETs are known to have a near-ideal subthreshold slope. For a slope factor of 1.05, a subthreshold slope of 63 mV/decade is indeed expected, but the presence of interface states brings it up to values around 66–68 mV/decade. In dual-gate SOI MOSFETs, values very close to the theoretical limit of 60 mV/decade are expected. Furthermore, this low value can be obtained for very short channel lengths provided the gate oxide thickness and the silicon film thickness are scaled accordingly.

Theoretical investigation shows a subthreshold slope with values lower than 63 mV/decade can be obtained for a gate length down to 0.1 μm if the device is designed using a scaling parameter $v = L/2\lambda$, where L is the gate length, and λ , called the “natural length,” is equal to

$$\lambda = \sqrt{\frac{\epsilon_{si}t_{si}t_{ox}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox}t_{si}}{4\epsilon_{si}t_{ox}} \right)} \quad (5.20)$$

In this case, the subthreshold swing is given by $S = (kT/q) \ln(10)[1/1 - 2\exp(-v)]$. As long as the value of v is larger than 3, the device is free from punchthrough, threshold voltage rolloff, and subthreshold slope degradation.⁶²

The most innovative device property of the dual-gate SOI MOSFET is the possibility of not only forming inversion layers at the top and the bottom of the channel region but also inverting the entire film thickness. This effect, which appears if the silicon film is thin enough, is called *volume inversion*. It increases the current

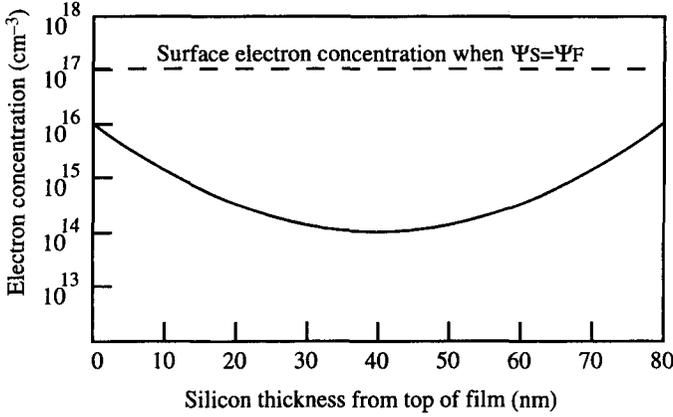


Figure 5.17 Minority-carrier (electron) concentration across a dual-gate, inversion-mode, n-channel transistor, at threshold ($t_{\text{si}} = 75 \text{ nm}$, $t_{\text{ox}} = 55 \text{ nm}$, $N_a = 10^{17} \text{ cm}^{-3}$, $V_{\text{th}} = 0.71 \text{ V}$).

drive in the device, compared to an SOI MOSFET that would have channels at both the top and the bottom of the device.⁶³

One of the most common criteria for defining the threshold voltage of a MOSFET uses the relationship $\Psi_S = 2\Psi_F$, where Ψ_S is the surface potential and Ψ_F is the Fermi potential. In inversion, therefore, threshold are reached when the surface potential reaches twice the Fermi potential, which depends on the doping level. This definition is inadequate for thin-film, dual-gate devices, where current appears following a weak inversion mechanism. Indeed, Figure 5.17 presents the electron concentration in the silicon film as current starts to flow in the device, which corresponds to the practical definition of the strong inversion threshold. Therefore, the necessity arises to adopt a mathematical definition of the threshold voltage that differs from the classical $\Psi_S = 2\Psi_F$ relationship.

Francis et al. have developed an extensive model of the dual-gate, inversion-mode SOI transistor where the threshold voltage is defined by the transconductance change (TC) method.⁶⁴ According to this method, the threshold voltage can be defined as the gate voltage where the derivative of the transconductance reaches a maximum, or, in mathematical terms, when $d^3I_D/dV_G^3 = 0$.⁶⁵

Using this condition, the surface potential at threshold can be obtained:

$$\Psi_S^* = 2\Psi_F + \frac{kT}{q} \ln \left[\delta \frac{1}{1 - \exp(-\alpha)} \right] \quad \text{where} \quad \alpha = \frac{q Q_D}{kT 8C_{\text{si}}} \quad \text{and} \quad \delta = \frac{C_{\text{ox}}}{4C_{\text{si}}}$$

and all other symbols have their usual meanings. The last term of the surface potential at threshold is negative, such that Ψ_S^* is smaller than $2\Psi_F$ by a value of 10–90 mV, which justifies the previous assumption of having a weak inversion current at threshold. The threshold voltage can be obtained analytically:

$$V_{\text{th}} = \Psi_S^* + V_{\text{FB}} + \frac{kT}{q} \frac{\alpha}{\delta} \sqrt{1 + \frac{\delta}{\alpha}} \quad (5.21)$$

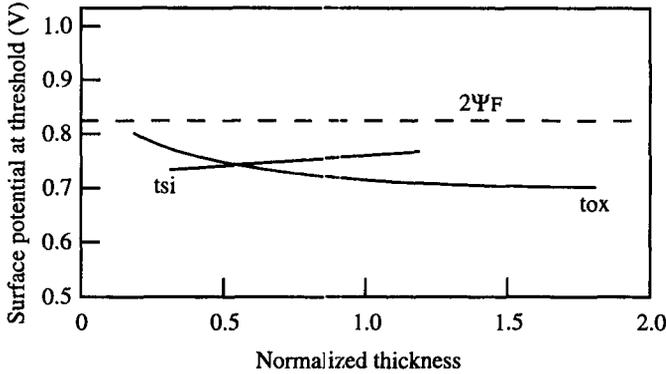


Figure 5.18 Surface potential evolution as a function of normalized silicon film and gate oxide thickness ($t_{si}/75$ nm and $t_{ox}/55$ nm; $N_a = 10^{17}$ cm⁻³).

The difference between the surface potential at inversion and $2\Psi_F$ depends on the silicon film thickness, the gate oxide thickness, and the doping concentration. The silicon film and gate oxide film thickness dependence is illustrated in Figure 5.18.

Thin-film, dual-gate transistors have a low threshold voltage (around 0 volt) when the p-type channel doping is low and the top and bottom gates are made of N⁺ polysilicon. To obtain a higher threshold voltage, it is possible to use P⁺ polysilicon for both gates. In that case a threshold voltage around 1 volt is obtained, which is too high for most applications. An intermediate solution consists into using either midgap or dual-type gate material (one gate is N⁺-doped and the other one is P⁺-doped). Figure 5.19 presents the threshold voltage in an n-channel dual-gate MOSFET with low channel doping concentration ($N_a = 10^{15}$ cm⁻³), as a function of the silicon film thickness. Another solution is, of course, to use two N⁺-poly gates and to increase the threshold voltage by increasing the channel doping level. This

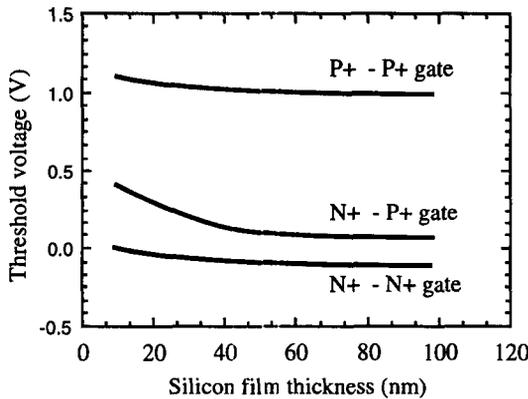


Figure 5.19 Dependence of threshold voltage in P⁺P⁺, N⁺P⁺, and N⁺N⁺ dual-gate SOI MOSFETs on silicon film thickness ($N_a = 10^{15}$ cm⁻³).

solution, however, has the disadvantage of decreasing the mobility through increased impurity scattering.

The fabrication steps necessary to produce a dual-gate MOSFET are not easily derived from a classic SOI process. Silicon wafer bonding has been used to fabricate dual-gate transistors. In this case, the bottom gate is realized on the handle wafer. After oxidation and planarization of the oxide, the SOI wafer is bonded to the handle wafer, thinned down, and SOI MOSFETs are fabricated on top of the bottom gates. This technique has been used to realize ultrafast devices with P^+ and N^+ dual gates.⁶⁶ Dual-gate SOI transistors can also be fabricated using epitaxial lateral overgrowth (ELO) or tunnel epitaxy of silicon over an oxidized polysilicon gate. A second gate is then fabricated on top of the device.⁶⁷ Another embodiment of the dual-gate SOI transistor makes use of regular SIMOX wafers and adopts a process sequence similar to that used for regular SOI MOSFET fabrication, with only one additional mask step and a wet-etch step in buffered HF. The device is called the *gate-all-around* (GAA) MOSFET.⁶⁸ The GAA device fabrication process makes use of SIMOX substrates where the original thickness of the silicon film is 120 nm. A thin pad oxide is grown, and silicon nitride is deposited. Using a mask step, the nitride and the silicon are etched to define the active areas. A local oxidation step is used to round off the edges of the silicon islands after which the nitride and the pad oxide are stripped. A mask step is then used to cover the entire wafer with resist except areas that correspond to an oversize of the intersection between the active area and the poly gate layers. The wafers are then immersed in buffered HF (BHF). At this step the oxide on the sidewalls of the silicon islands as well as the buried oxide are etched and a cavity is created underneath the center part of the silicon islands. The wafers are removed from BHF once the cavity etch is completed. At this point, the device resembles a silicon bridge supported by its extremities (which will later on become source and drain), which is hanging over an empty cavity. Gate oxidation is then carried out. In this step, a 30-nm-thick gate oxide is grown over all the exposed silicon (top, bottom, and edges of the active silicon, as well as on the silicon substrate in the bottom of the cavity). Boron is implanted to adjust the threshold voltage, and polysilicon gate material is then deposited and doped N type. Because of the extremely good step coverage of LPCVD polysilicon, the gate oxide over the cavity is completely coated with polysilicon, and a gate is formed on the top, the sides and the bottom of the channel area [hence the name *gate-all-around* (GAA) device]. The polysilicon gate is then patterned using conventional lithography and anisotropic plasma etch. Source and drain are formed using phosphorous implantation followed by an annealing step. CVD oxide is deposited, and contact holes are opened. An aluminum metallization step completes the process. The final silicon thickness is 80 nm. A schematic cross section of the device is presented in Figure 5.20. Accumulation-mode p-channel GAA devices can be obtained just as easily by using a lower-dose boron channel implant and P^+ doping of the sources and drain.

The GAA transistor is in volume inversion when the gate voltage is close the threshold voltage. This can clearly be observed in Figure 5.21, where the transconductance of a conventional device and that of a GAA device are compared.

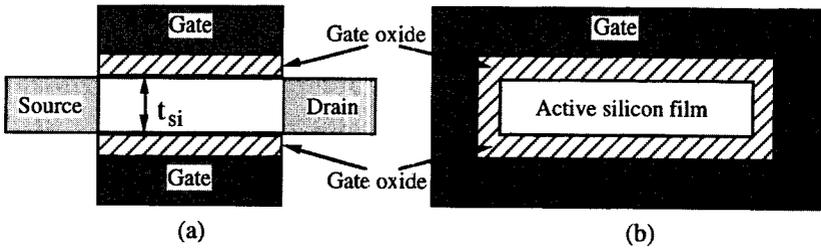


Figure 5.20 GAA device cross section (a) parallel and (b) perpendicular to the current flow direction.

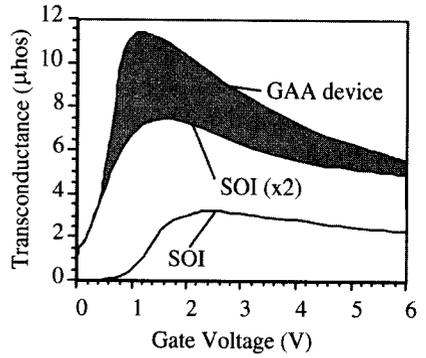


Figure 5.21 Transconductance (dI_d/dV_G) at $V_{ds} = 100$ mV in a conventional SOI MOSFET and a GAA device [$(W/L)_{\text{mask}} = 3 \mu\text{m}/3 \mu\text{m}$].

An additional curve labeled “SOI × 2” presents the transconductance of the SOI MOSFET multiplied by 2 to account for both the presence of two channels in the GAA device. The gray area represents the extra drive of the GAA device, which is attributed to volume inversion.

The contribution of volume inversion to drain current is more pronounced right above threshold where the inversion layer is distributed across the entire silicon film and where the effects of bulk mobility, in contrast to surface mobility, can be observed. At higher gate voltages, there is still inversion in the center of the silicon film, but the carriers are now mostly localized in inversion layers near the interfaces. As a result, more scattering occurs and the transconductance tends to be equal to twice that of that a conventional device. Because of the excellent coupling between the surface potentials and the gate voltage, a subthreshold slope of 60 mV/decade is obtained at room temperature.

5.4.3 Bipolar Transistors

Different types of bipolar transistors can be realized in SOI, depending on the application and the silicon film thickness under consideration. If a pure bipolar circuit is contemplated (e.g., ECL), a “hick” silicon layer can be employed and classic vertical transistors can be fabricated. The advantages of using an SOI substrate reside in the reduction of collector–substrate capacitances, the full

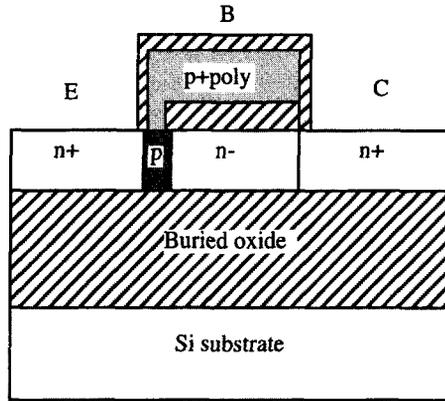


Figure 5.22 Lateral SOI bipolar transistor with a top base contact.

dielectric isolation, and the reduction of sensitivity to alpha particles (reduction of the soft-error rate).⁶⁹

If the bipolar transistors have to be integrated with thin-film CMOS devices (BiCMOS and CBiCMOS), pretty efficient lateral bipolar transistors can be realized as well. Early SOI lateral bipolar transistors suffered from high-base resistance problems because the base contact was made on one side of the device (MOSFET body contact).⁷⁰ More recent devices use base contacts made at the top of the device, which improves the base resistance but usually complicates the fabrication process (Figure 5.22).⁷¹

Lateral polysilicon emitters can also be realized, also at the expense of process complexity. It is, however, worthwhile to mention that the increase of process complexity necessary to upgrade a CMOS SOI process to (C)BiCMOS is much more modest than for a bulk process, simply because of the inherent dielectric isolation of SOI devices. Table 5.3 presents the performances of some SOI bipolar transistors.^{71–78} It is worthwhile noting that the maximum gain of lateral devices is usually obtained for relatively low collector current values. Indeed, high current densities and high-injection phenomena are reached much more quickly in thin-film lateral devices than in vertical devices because of the small silicon thickness.

TABLE 5.3 Performance of SOI Bipolar Transistors

Type	L(μm)	β_F	BV_{CE0} (V)	f_T (GHz)	Company
Lateral	0.2	120	2.5	4.5	UCB
Hybrid	0.3	10,000	2.5	—	UCB
Lateral	—	90	3	15.4	Philips
Lateral	—	80	>3	10	Motorola
Lateral	—	30	2.8	20	IBM
Vertical	—	50	8	12.4	SIM
Vertical	—	110	12	4.5	Analog Devices

Because of the need to make a buried collector layer, it may seem impossible to fabricate vertical bipolar transistors in thin SOI films. An original solution to this problem has, however, been proposed, and vertical NPN transistors have been realized in 400-nm-thick SOI films (and the use of even thinner films is quite possible).⁷⁹ In that device there is no buried collector diffusion. The bottom of the P⁻ intrinsic base is directly in contact with the buried oxide. If no back-gate bias is applied, very few of the electrons injected by the emitter into the base can reach the lateral collector, and the current gain of the device is extremely small. When a positive bias is applied to the back gate, however, an inversion layer is induced at the silicon–buried oxide interface, at the bottom of the base, and acts as a buried collector (Figure 5.23). Furthermore, the band bending in the vicinity of the bottom inversion layer attracts the electrons injected by the emitter into the base. As a result, collection efficiency is drastically increased, and the current gain can reach useful values. The use of a field-effect-induced buried collector makes it possible to obtain significant collection efficiency without the need to form a diffused buried layer and an epitaxy step. It also solves the high-injection (high current densities) problems inherent to thin-film lateral devices. Furthermore, because the neutral base width does not depend on the collector voltage, but on the back-gate voltage, the Early effect is almost totally suppressed.⁸⁰ Since the amplification factor of a bipolar transistor is given by the Early voltage \times current–gain product, this feature can be extremely attractive for high-performance analog applications. Vertical SOI bipolar transistor with field-induced collector with an Early voltage over 50,000 V and a current gain of 200 have been demonstrated. Such devices offer thus an Early voltage \times current gain product larger than 1,000,000.⁸¹

5.4.4 Quantum-Effect Devices

When a dual-gate MOSFET operates in the volume inversion regime, the electrons form a two-dimensional electron gas (2DEG), with a thickness equal to the silicon film thickness. One can, therefore, expect to observe splitting of the conduction band into subbands, since the electrons are confined in one spatial direction, along which the electron wavefunctions form standing waves. The wavefunctions and the

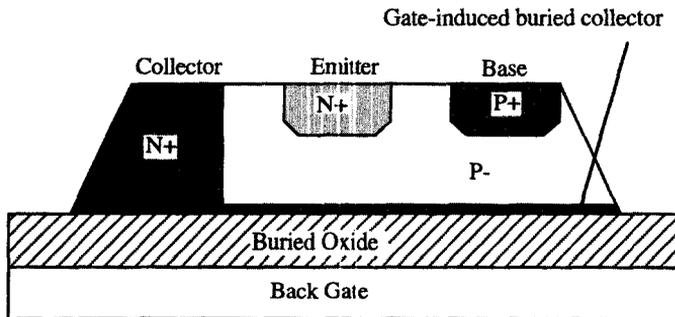


Figure 5.23 Vertical bipolar transistor with a gate-induced buried collector.

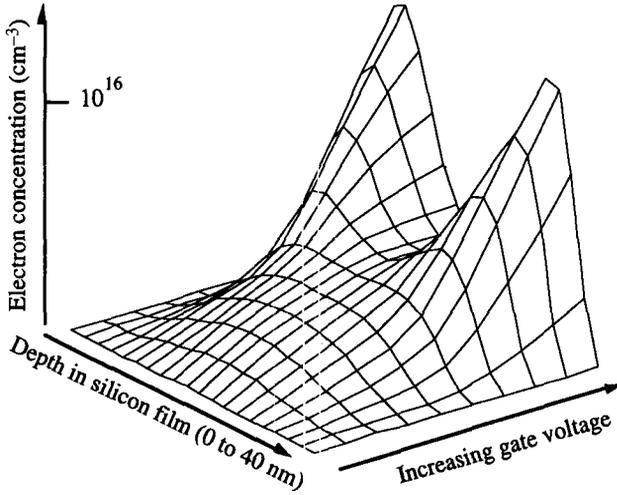


Figure 5.25 Electron concentration versus depth in the silicon film and versus $V_G - V_{th}$ (40-nm-thick device).

Figure 5.25 presents the evolution of the electron concentration in the 40-nm-thick device as a function of gate voltage. One can clearly see the volume inversion right above threshold and the formation of two inversion channels at higher gate voltages.

Figure 5.26 presents the evolution of the subband energy levels as a function gate voltage. The plot is realized in such a way that the Fermi level remains constant.

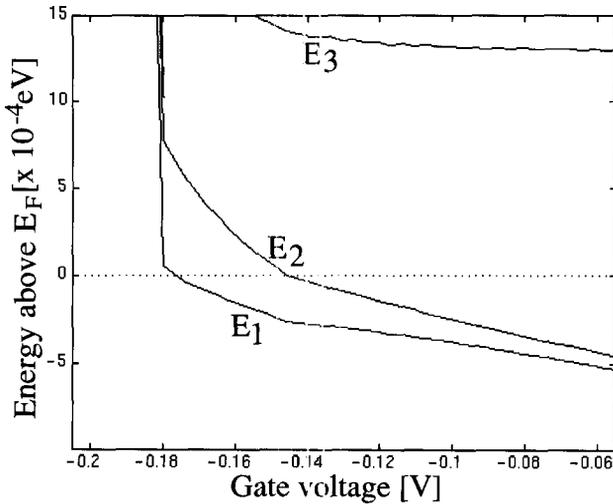


Figure 5.26 Evolution of energy levels (relative to E_F) as a function of gate voltage (40-nm device).

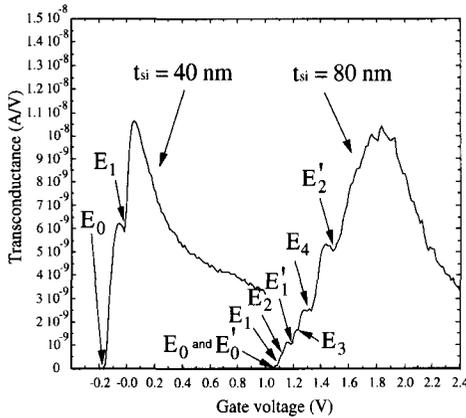


Figure 5.27 Transconductance versus gate voltage for two silicon film thicknesses. The energy levels corresponding to the transconductance humps are indicated [E_0 corresponds to the first subband ($m^* = m_1e$); E_1 , to the second subband ($m^* = m_1$); E_0' , to the first subband with $m^* = m_t$, etc.]; m_1 and m_t are the classic effective mass of an electron in silicon ($m_1 = 0.98 m_0$ and $m_t = 0.19 m_0$, where m_0 is the mass of the electron in a vacuum).

Every time an energy level falls below the Fermi level, the corresponding subband becomes populated.

As a result of intersubband scattering, mobility decreases every time a new subband becomes populated, which generates humps in the transconductance curve of the device, as presented in the experimental curves of Figure 5.27.⁸³

Short quantum-wire transistors have been fabricated in SIMOX.^{84,85} The conductance of these devices increases in a staircase manner as a function of gate voltage. After correcting for the source and drain resistance the transconductance was found to increase in multiples of $4q^2/h$, which agrees with the Landauer formula for universal transconductance fluctuations.⁸⁶ Longer quantum-wire transistors have been realized as well, and transconductance fluctuations due to subband splitting have been observed as well (Figure 5.28).⁸⁷

The fabrication of a single-electron transistor (SET) on SIMOX has been reported as well. This device is basically a short quantum wire connected to source and drain through constrictions. The higher energy levels in the constrictions isolate the wire from the outside world. When a gate voltage is applied, current can flow through the quantum wire by a Coulomb blockade mechanism, which is made possible by the very small capacitances (2 aF) involved in the structure. Conductance oscillations as a function of gate voltage have been observed at temperatures as high as 300 K.^{88,89}

The fabrication and successful operation of SOI single-electron memory cells with floating-dot gate has been reported as well (Figure 5.29). In such a device, which basically behaves as a miniature EEPROM cell, a single electron can be injected into an electrically floating dot. Because the dot has a very small capacitance the injection of such a minute charge as that of a single electron gives the dot to a potential which is sufficiently high (several volts) to modify the threshold

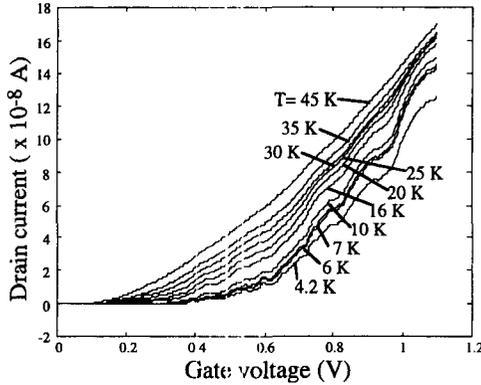


Figure 5.28 Current as a function of gate voltage in a series of seven parallel SOI quantum-wire MOSFETs at different temperatures ($V_{DS} = 10$ mV). The device thickness and width are 86 and 100 nm, respectively.

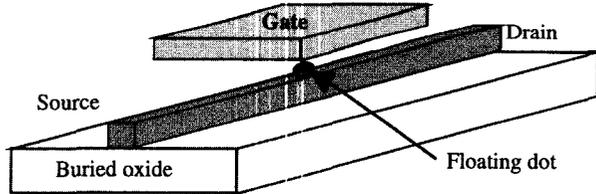


Figure 5.29 SOI single-electron memory cell MOSFET.

voltage of a quantum-wire MOSFET. This MOSFET can, in turn, be used as a memory element since it will present different threshold voltage values depending on whether there is an electron stored in the floating dot.⁹⁰

5.5 SOI CIRCUITS

The total dielectric isolation of SOI devices makes it possible to realize circuits operating at high temperature and to integrate power devices with CMOS and bipolar transistors on a single chip to realize smart-power circuits. In addition, the reduced body factor of fully depleted SOI MOSFETs allows one to fabricate high-performance circuits and, in particular, DRAMs operating with a low supply voltage.

5.5.1 High-Temperature Circuits

The excellent behavior of thin-film SOI MOSFETs at high temperature makes SOI technology highly suitable for high-temperature IC applications. Indeed, the major causes of failure in bulk CMOS logic at high temperature, specifically, excessive power consumption and degradation of logic levels and noise margin, are observed to be much reduced in SOI circuits. Latchup is, of course, totally suppressed when

SOI technology is used. SOI CMOS inverters exhibit full functionality and very little change in static characteristics at temperatures up to 320°C .⁹¹ The switching voltage remains stable, due to the remarkably weak and symmetrical variation of the n- and p-channel threshold voltages. The output voltage swing is reduced by only a few millivolts at 320°C , due to the slightly increased leakage current of the OFF devices and the reduced carrier mobility of the ON devices. This degradation is, however, totally negligible when compared to what is observed in bulk devices. In logic gates with series transistors, such as NAND gates, the increase of standby supply current with temperature remains even more limited than what can be expected on basis of the leakage current of all constituent devices (Figure 5.30). This is because in SOI circuits, the drain leakage current of each individual transistor flows toward its source, and thus into the following transistor, unlike in bulk circuits, where all drain leakage currents are collected by the substrate.

In analog circuits, such as operational amplifiers, the increase of leakage current with temperature may result in the loss of the operating point in bulk circuits. In a bulk circuit, some (or most) of this bias current will flow to the substrate if the junction leakage increases. As a result, the devices are no longer properly biased and the circuit no longer operates correctly unless the bias current is increased to compensate for the leakage losses. If SOI technology is used, the leakage current can flow only through the branches of the circuit and the circuit operates without loss of the bias operating point.

Excellent high-temperature performance can be obtained using circuit design techniques. A cascode gain-boosting stage can be used to decrease the output conductance (g_D) of the output stage. The zero-temperature coefficient (ZTC)

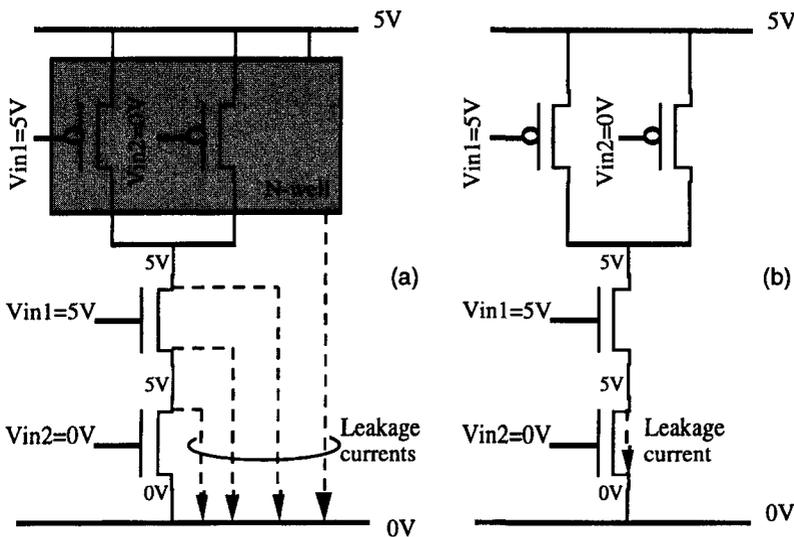


Figure 5.30 Leakage current paths in (a) a bulk circuit and (b) an SOI CMOS NAND gate. Both gates have a high and a low input bias. All bulk junction leakage currents flow in parallel toward the substrate. The leakage of the SOI gate is limited to that of a single transistor.

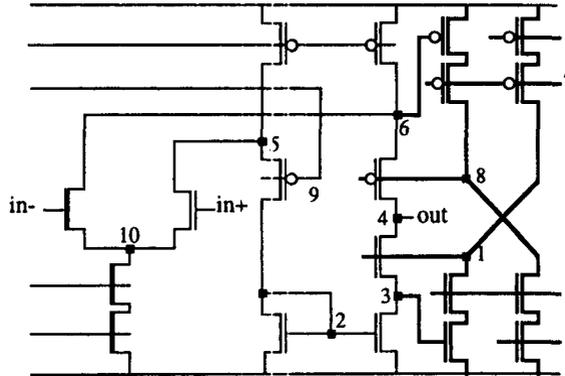


Figure 5.31 Schematic diagram of a CMOS OTA designed with a gain-boosting stage. The numbers correspond to several nodes of the circuit.

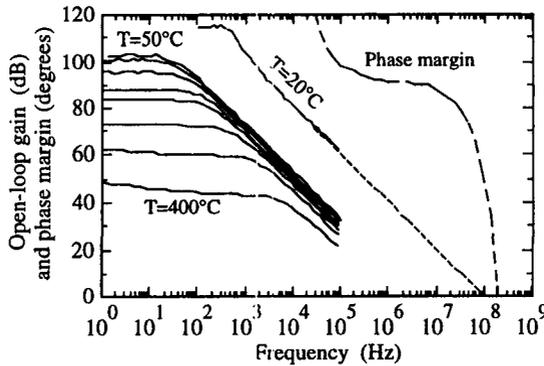


Figure 5.32 Measured Bode diagram of the gain-boosting OTA presented in the previous figure measured with a 10-pF load at 20°C and a 250-pF load at all other temperatures. The phase margin was measured at 20°C.

concept can also be applied, in which there exists a bias point where the current is independent of temperature since the reduction of mobility is compensated by the reduction of threshold voltage. Figure 5.31 shows such an amplifier. It is not a low-power device (25 mW), but it produces a high gain (115 dB) and a wide bandwidth (100 MHz). The dc gain is still 50 dB at a temperature of 400°C (Fig. 5.32).

5.5.2 Low-Voltage, Low-Power (LVLP) Circuits

As more and more systems become portable, there is a general trend toward the reduction of the supply voltage and the power consumption of integrated circuits. In these circuits SOI CMOS offers unique opportunities. The assets of SOI MOSFETs in this area are their reduced drain capacitance, low body effect, sharp subthreshold slope, and high drive current. Both digital and analog circuits using SOI CMOS technology offer excellent LVLP performance.

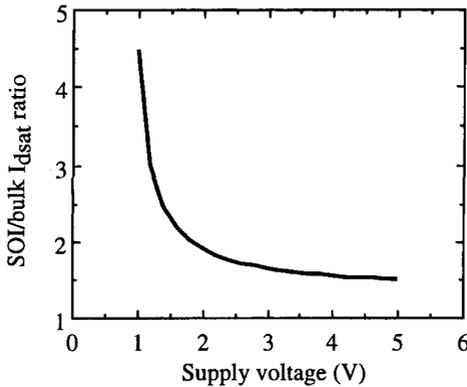


Figure 5.33 SOI/bulk I_{dsat} ratio versus supply voltage ($n_{SOI} = 1.05$, $n_{bulk} = 1.4$).

It is well known that fully depleted (FD) SOI devices have a near-optimal subthreshold slope. This allows one to reduce the threshold voltage and to increase the current drive without increasing the OFF current compared to bulk devices. Low leakage current and low threshold voltage can thus be obtained simultaneously.

If one considers maximum speed performance, the allowed reduction of threshold voltage in SOI devices has a positive, but modest, effect on the speed of the circuits when $V_{DD} = 5$ V. This effect, however, becomes more pronounced as the supply voltage is reduced. Figure 5.33 presents the ratio of saturation currents between an SOI MOSFET with a body factor $n = 1.05$ and a threshold voltage $V_{th} = 0.45$ V and a bulk transistor with $n = 1.4$ and $V_{th} = 0.7$ V. The I_{Dsat} enhancement due to the V_{th} reduction is quite remarkable at low supply voltage.

In parallel to the reduction of V_{DD} , a reduction of power consumption is desirable. The dissipated dynamic power of a circuit is equal to fCV^2 , where f is the frequency, V is the supply voltage, and C is the sum of all the capacitances in the circuit. Since all capacitances but the gate oxide capacitance are smaller in SOI than in bulk, less power is dissipated in SOI circuits.⁹²

It has been reported, for instance, that a frequency divider implemented on SOI is twice as fast and consumes half of the power of the equivalent bulk circuit and that ASICs and 16 K (16-kilobyte) SRAMs deliver the same speed performance at 2 V at 25% of the power of the equivalent bulk circuits operated at 3.3 V.⁹³ Table 5.4 presents the performances of some recent low-voltage, low-power (LVLP), high-speed SOI CMOS circuits.^{94–107} The choice between fully depleted (FD) and partially depleted (PD) devices is not fully resolved yet. Some favor the better performances of the FD devices, and others prefer the better uniformity of threshold voltage control expected from PD devices.

The dc gain (A_0) and the transition unit-gain frequency (f_T) of a MOSFET are given by $A_0 = -(g_m/I_D)V_A$ and $f_T = (g_m/2\pi C_L)$, where V_A is the Early voltage of the device and C_L is its load capacitance. The g_m/I_D ratio is a measure of the efficiency to translate current (hence power) into transconductance. The larger the g_m/I_D ratio, the larger the transconductance for a given current value. The high

TABLE 5.4 Low-Voltage, Low-Power, and High-Speed SOI CMOS Circuits

Circuit	$L(\mu\text{m})$	$V_{DD}(\text{V})$	Speed	Power	Company
Frequency divider	0.15	1	1.2 GHz	50 μW	NTT
Frequency divider	0.15	2	2.5 GHz	130 μW	NTT
Frequency divider	0.1	1	1.2 GHz	60 μW	NTT
Frequency divider	0.1	2	2.6 GHz	350 μW	NTT
Prescaler	0.4	1	1 GHz	0.9 mW	NTT
Prescaler	0.4	2	2 GHz	7.2 mW	NTT
PLL	0.4	1.2	1 GHz	1.4 mW	NTT
PLL	0.4	2	2 GHz	8.4 mW	NTT
PLL	0.24	1.5	2.2 GHz	4.5 mW	NTT
512K SRAM	0.2	1	3.5-ns access time	—	IBM
64K DRAM	0.6	1.5	—	—	Mitsubishi
Microcontroller CPU	0.5	0.9	5.7 MHz	—	Motorola
8K-gate 16-bit ALU	—	0.5	40 MHz	5 nW to 350 μW	NTT
300K gate array	0.25	1.2	38 MHz	30 mW	NTT
16 \times 16-bit multiplier ^a	—	0.5	18 ns	4 mW	Toshiba
PLL	0.35	1.5	1.1 GHz	—	Sharp
32-bit ALU*	—	0.5	260 MHz	2.5 mW	Toshiba
8 \times 8 ATM switch	0.25	2	40 Gb/s	8.4 W	NTT
560K master array	0.35	1	50 MHz	50 nW/MHz per gate	Mitsubishi
Power PC processor	—	—	1 GHz	—	IBM
Alpha processor	0.18	1–2	1.5 GHz	—	Samsung

^aUses hybrid bipolar–MOS devices.

g_m/I_D value of fully depleted SOI devices should, therefore, allow one to realize near-optimal micropower analog designs (g_m/I_D values of 35 V^{-1} are obtained, while g_m/I_D reaches typical values of 25 V^{-1} in bulk MOSFETs).¹⁰⁸

CMOS operational amplifiers (op amps) can be designed by knowing the g_m/I_D and the Early voltage, V_A , as a function of the scaled drain current, $I_D/(W/L)$.¹⁰⁹ This design technique was used to synthesize a simple one-stage op amp with a specified active device size (W/L) and output capacitance (C_L). Figure 5.34 clearly shows that the gain of the SOI amplifier is 25–35% larger than in the bulk counterpart, while the static bias current I_{dd} is reduced in the same proportion. Taking into account the reduction of parasitic source–drain and intrinsic gate capacitances found in SOI in addition to the lower n value, it can be further demonstrated that a two-stage Miller op amp can be synthesized to achieve an open-loop gain 15 dB larger and an I_{dd} 3 times smaller in SOI than in bulk for similar area and identical bandwidth.¹⁰⁹ Conversely, total area and I_{dd} can be divided by a factor of up to 2–3 when all other specifications are kept constant. These theoretical predictions have been confirmed by the realization of a 100 dB open-loop gain $3 \mu\text{A}\text{-}I_{dd}$ Miller opamp using a fully depleted SOI CMOS process (Figure 5.35). These results establish the potential of FD SOI CMOS to boost the speed (f_T), gain and power ($I_{dd} \times V_{dd}$) performances of op amps well over bulk implementations.

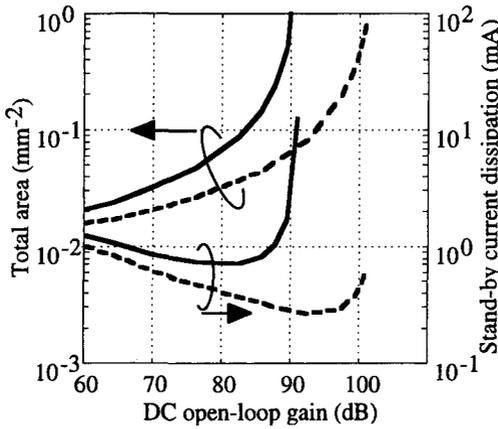


Figure 5.34 Synthesis of bulk (—) and SOI (---) Miller amplifier for $C_L = 10$ pF and $f_T = 10$ MHz.

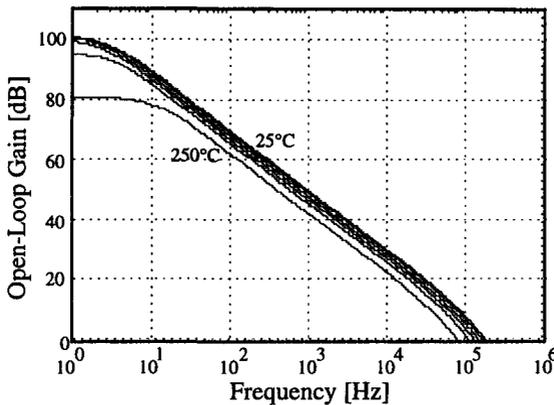


Figure 5.35 Measured Bode diagram of the SOI Miller operational amplifier ($V_{dd} = 3$ V, $I_{dd} = 3$ μ A). Measurement was carried out at six different temperatures (25, 50, 100, 150, 200, and 250°C ($C_{load} = 60$ pF)).

5.5.3 Smart Power Circuits

The full dielectric isolation provided by the SOI substrate allows one to fabricate high-voltage devices without the burden of having to use a complicated junction isolation process. Compared to bulk, the advantages of the SOI full dielectric isolation are a higher integration density, no latchup, less leakage current, high-temperature operation, improved noise immunity, easier circuit design, and the possibility to integrate vertical DMOS transistors. Furthermore, power devices can be integrated on the same substrate as CMOS logic to produce “smart power” circuits. Relatively thick films of SIMOX or BESOI with thicknesses ranging from 5

TABLE 5.5 Characteristics of Power Devices Used in Smart Power SOI Circuits

Power Device	t_{si} (μm)	V_{max} (V)	I_{max} (A)	R_{on} ($\Omega\cdot\text{mm}^2$)	Company
VDMOS	6	50	—	0.15	FhG IMS
VDMOS	15	150	2	2	IMS
LDMOS	5	400	—	15	Philips
LIGBT	5	400	—	5	Philips
VDMOS	“Bulk”	500	10	18	FhG IMS
LDMOS	20	640	—	—	Daimler
n-MOSFET	Thin	44	0.5	$19\ \Omega\cdot\text{mm}$	Allied Signal
VDMOS	“Bulk”	1200	—	—	FhG IMS

to $20\ \mu\text{m}$ are usually employed (Table 5.5), although some devices have been fabricated in thin SOI films as well.^{111–117} It is also possible, using a mask during oxygen ion implantation, to produce SIMOX material with bulk regions (i.e., unimplanted areas). Using such a technique, it is possible to combine SOI CMOS devices and bulk power devices on a same chip.

The flexibility of dielectric isolation allows for the fabrication of mixed chips featuring, such as CMOS + NPN bipolar + VDMOS devices (Fig. 5.36) using a process that requires fewer masking steps than the equivalent bulk process.¹¹⁸

The main drawback of dielectric isolation lies in the fact that it is accompanied by an increased thermal resistance between the device and the heat sink, because the thermal conductivity of SiO_2 is lower than that of silicon. As a result, the temperature rise in SOI power devices is higher than that in bulk power devices. The differences in self-heating between SOI and bulk power devices is largest for short power transients because the initial temperature rise in SOI devices is more rapid. As the pulse length increases, the difference in temperature rise between SOI and bulk devices converges to a constant value, which is proportional to the thickness of the buried oxide. For oxide thicknesses under $2\ \mu\text{m}$, the steady-state temperature is

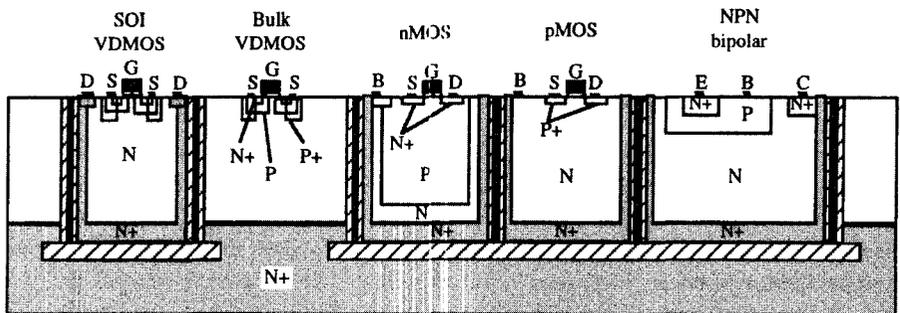


Figure 5.36 Process featuring SOI quasivertical DMOS, bulk vertical DMOS, CMOS, and NPN devices.

TABLE 5.6 Milestones in the Evolution of SOI Memory Chips

Year	Company	RAM
1982	NTT	1 K SRAM
1988	Harris	High-temperature CMOS 4 K SRAM
1989	TI/Harris	64 K SRAM
1989	CEA/LETI	Thin-film 16 K SRAM
1990	TI	Commercial 64 K SRAM
1991	TI	256 K SRAM
1991	IBM	256 K fully depleted SRAM
1993	TI	1-MB SRAM, 20 ns access time at 5 V
1993	IBM	512 K SRAM, 3.5 ns access time at 1 V
1993	Mitsubishi	64 K DRAM, 1.5 V operation
1995	Samsung	16 MB DRAM
1996	Mitsubishi	16-MB, DRAM, 0.9-V operation
1997	Hyundai	1-GB DRAM, 2-V operation

essentially determined by the thermal properties of the substrate and the device package, and not by the buried oxide.¹¹⁹

5.5.4 SRAMs and DRAMs

SOI memory chips evolved quite rapidly during the late 1990s. Until 1992 only SRAMs chips had been demonstrated, but later, when the quality of SIMOX material reached a high-enough quality to ensure low leakage current levels in pass transistors, DRAM chips were fabricated as well. Table 5.6 presents the evolution of SOI SRAMs and DRAMs over the years.^{120–132}

The advantages of SOI for memory chips are multiple. In the case of SRAMs, the use of SOI reduces the rate of SEU (single-event upset) and rules out SEL (single-event latchup). Hence the use of SOI for memory chips used in space applications. In addition, the reduced parasitic capacitances and the reduced body effect allow for faster and lower-voltage operation. In the case of DRAMs, the advantages of SOI are a reduction of the bit line capacitance (by 25% compared to bulk), a reduction of the access transistor leakage current, and the reduction of the soft-error sensitivity. As a result, a storage capacitance of 12 fF is sufficient for a 256-MB, 1.5-V DRAM realized in SOI, while 34 fF are required for the equivalent bulk device. The advantage of using SOI, in terms of reducing the memory cell area, is obvious. It is also possible to realize DRAMs that can operate with a <1-V power supply, which has not yet been demonstrated in bulk. To achieve this level of performance, a reduced C_b/C_s ratio must be achieved (C_b and C_s are the bit line and the storage capacitances, respectively). The reduction of the bit-line capacitance comes “for free” with the use of SOI, while the increase of C_s is achieved through the cell design. Considering the cell retention time, it has been shown that thin-film SOI MOSFETs have a much lower leakage current than bulk or thicker SOI devices. OFF-

state leakage currents smaller than $1 \text{ fA}/\mu\text{m}$ are observed in fully depleted, thin-film devices.¹³³

DRAM architecture can also take advantage of the low body effect found in SOI devices. Indeed, bulk pass transistors suffer from body effect during the storage capacitor charging state such that the word line voltage has to be boosted to $V_{WL} = V_{th} + V_{ds} + \Delta V$, where ΔV compensates for the V_{th} increase in the bulk pass transistor design. Thus the body effect becomes a major issue in bulk pass transistor design.¹³⁴ The use of fully depleted SOI significantly improves the performance of pass gates because of the low body effect found in SOI devices.

In contrast to what happens in a bulk device, the body potential of a partially depleted transistor follows the source node potential during the charging, so that the body-to-source potential difference is kept at a constant value. For fully depleted SOI transistors there is no body-potential change. Therefore, SOI transistors have a higher charging efficiency than bulk transistors. If the charging current is designed to be $1 \mu\text{A}/\mu\text{m}$, then the charging efficiency, defined as the ratio of source (the storage node) voltage at $1 \mu\text{A}/\mu\text{m}$ to the drain (the bit line) voltage, is 80% for bulk transistors, and 88% and 98% for partially depleted and fully depleted transistors, respectively (Fig. 5.37). The higher charging efficiency of SOI transistors results in higher programming speed for the DRAM cell. At time $t = 1 \text{ ns}$ after being activated by the word line, a fully depleted transistor can transfer 10 times more charge than a bulk transistor. For a bulk pass transistor to have the same charging efficiency as an SOI transistor, the word line voltage has to be increased during the charging. As shown in Figure 5.37, to reach a 90% charging efficiency, the gate voltage of a bulk transistor has to be 0.6 V higher than that of a fully depleted SOI MOSFET. Such an increased word line voltage can degrade gate oxide integrity and is also not favorable for low-power operation.¹³⁴

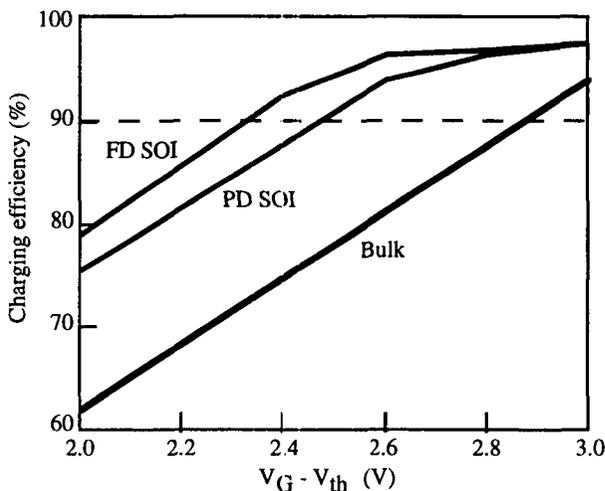


Figure 5.37 Charging efficiency as a function of gate voltage overdrive.

5.6 SUMMARY AND FUTURE TRENDS

After over 10 years of extensive research and development SOI substrates now have excellent quality and their availability increases with the number of vendors. There is no doubt that modern SIMOX and UNIBOND wafers can meet the requirements imposed by CMOS and DRAM processing.

Owing to reduced parasitic capacitances, low body factor, sharp subthreshold slope, increased drive current, and absence of latchup, SOI MOSFETs are ideal candidates for high-performance CMOS, especially in the sub-1 volt regime. The high leverage of analog, digital, RF, and DRAM devices and circuits has been demonstrated by many research and development teams.

The exceptional flexibility in device design offered by the full dielectric isolation of SOI gives one the opportunity to conceive novel devices with interesting performances. SOI thus becomes the material of choice for silicon-based nanostructures and quantum-effect devices.

There is little doubt that, beside some niche applications such as high-temperature electronics, SOI is becoming a major contender in the race towards low-voltage, low-power integrated circuits. Indeed, SOI devices behave exceptionally well at supply voltages of 1 V and below and can be used to realize digital, analog, RF and DRAM parts integrated on a single chip. Furthermore, the design and processing flexibility offered by the full dielectric isolation permits one to fabricate silicon nanostructures, such as single-electron flash memory cells that will, without any doubt, play an important role in future electronic components.

REFERENCES

1. J. E. Lilienfield, U. S. Patents 1,745,175 (filed 1926, issued 1930), 1,877,140 (filed 1928, issued 1932), and 1,900,018 (filed 1928, issued 1933).
2. J. P. Colinge, *Silicon-on-Insulator Technology*, 2nd ed., Kluwer, Norwell, 1997, p. 179.
3. J. Amano and K. Carey, "A Three-Step Process for Low Defect Density Silicon-on-Sapphire," *Appl. Phys. Lett.* **39**, 163 (1981).
4. D. J. Dumin, S. Dabral, M. H. Freytag, P. J. Robertson, G. P. Carver, and D. B. Novoty, "High-Mobility Transistors Fabricated on Very Thin SOS Films," *IEEE Trans. Electron Devices* **36**, 596 (1989).
5. J. P. Colinge, E. Demoulin, D. Bensahel, and G. Auvert, "Use of Selective Annealing for Growing Very Large Grain Silicon on Insulator," *Appl. Phys. Lett.* **41**, 346 (1982).
6. T. Nishimura, Y. Inoue, K. Sugahara, S. Kusunoki, T. Kumamoto, S. Nakagawa, M. Nakaya, Y. Horiba, and Y. Akasaka, "Three Dimensional IC for High-Performance Image Signal Processor," *Tech. Digest IEDM*, Dec. 1987, p. 111.
7. T. Kunio, K. Omay, Y. Hayashi, and M. Morimoto, "Three Dimensional ICs Having Four Stacked Active Device Layers," *Tech. Digest IEDM*, Dec. 1989, p. 837.
8. E. I. Givargizov, *Oriented Crystallization on Amorphous Substrates*, Plenum Press, New York, 1991.

9. J. C. C. Fan, M. W. Geis, and B. Y. Tsaur, "Lateral Epitaxy by Seeded Solidification for Growth of Single-Crystal Si Films on Insulators," *Appl. Phys. Lett.* **38**, 365 (1981).
10. T. J. Stultz, "Arc Lamp Zone Melting and Recrystallization of Si Films on Oxidized Silicon," *Appl. Phys. Lett.* **41**, 824 (1982).
11. E. I. Givargizov, V. A. Loukin, and A. B. Limanov, "Defect Engineering in SOI Films Prepared by Zone-Melting Recrystallization," in J. P. Colinge, V. S. Lysenko, and A. N. Nazarov, eds., *Physical and Technical Problems of SOI Structures and Devices*, Kluwer Academic Publishers, NATO ASI Series—High Technology, Vol. 4, 1995, p. 27
12. R. P. Zingg, H. G. Graf, W. Appel, P. Vöhringer, and B. Höfflinger, "Thinning Techniques for 1 μ m ELO-SOI," *Proc. IEEE SOS/SOI Technology Workshop*, Oct. 1988, p. 52.
13. J. P. Denton and G. W. Neudeck, "Fully Depleted Dual-Gated Thin-Film SOI p-MOSFET with an Isolated Polysilicon Backgate," *Proc. IEEE International SOI Conf.* Oct. 1995, p. 135.
14. P. J. Schubert and G. W. Neudeck, "Confined Lateral Selective Epitaxial Growth of Silicon for Device Fabrication," *IEEE Electron Device Lett.* **11**, 181 (1990).
15. S. Venkatesan, C. Subramanian, G. W. Neudeck, and J. P. Denton, "Thin-Film Silicon-on-Insulator (SOI) Device Applications of Selective Epitaxial Overgrowth," *Proc. IEEE Int. SOI Conf.* Oct. 1993, p. 76.
16. B. Aspar, C. Pudda, A. M. Papon, A. J. Auberton-Hervé, and J. M. Lamure, "Ultra Thin Buried Oxide Layers Formed by Low Dose SIMOX Process," in S. Cristoloveanu, ed., *Silicon-on-Insulator Technology and Devices*, The Electrochemical Society, Proceedings Vol. 94-11, 1994, p. 62, Abstract 541.
17. A. J. Auberton-Hervé, J. M. Lamure, T. Barge, M. Bruel, B. Aspar, and J. L. Pelloie, "SOI Materials for ULSI Applications," *Semiconduct. Int.* **97** (Oct. 1995).
18. R. Stengl, T. Tan, and U. Gösele, "A Model for the Silicon Wafer Bonding Process," *Jpn. J. Appl. Phys.* **28**, 1735 (1989).
19. T. Abe, M. Nakano, and T. Itoh, "Silicon Wafer-Bonding Process Technology for SOI Structures," in D. Schmidt, ed., *Proc. 4th Int. Symp. Silicon-on-Insulator Technology and Devices*, The Electrochemical Society, Vol. 90-6, 1990, p. 61.
20. S. D. Collins, "Etch Stop Techniques for Micromachining," *J. Electrochem. Soc.* **144**, 2242 (1997).
21. A. Yamada, O. Okabayashi, T. Nakamura, E. Kanda, and M. Kawashima, "A Computer Controlled Polishing System for Silicon-on-Insulator (SOI)," *Ext. Abstr. 5th Int. Workshop on Future Electron Devices*, Miyagi-Zao, Japan, May 1988, p. 201.
22. See, for example: *Solid-State Technology*, p. 155, June 1994, or back cover page of *European Semiconductor*, May 1994.
23. P. B. Mumola and G. J. Gardopee, "Advances in the Production of Thin-Film Bonded SOI and Ultra Flat Bulk Wafers Using Plasma Assisted Chemical Etching," *Extended Abstracts of the International Conference on Solid-State Devices and Materials*, Yokohama, Japan, Aug. 1994, p. 256.
24. M. Bruel, "Silicon on Insulator Material Technology," *Electron. Lett.* **31**, 1201 (1995).
25. A. J. Auberton-Hervé, "SOI: Materials to Systems," *Tech. Digest IEDM*, 1996, p. 3.

26. C. Maleville, B. Aspar, T. Poumeyrol, H. Moriceau, M. Bruel, A. J. Auberton-Hervé, T. Barge, and F. Metral, "Physical Phenomena Involved in the Smart-Cut Process," in P. L. F. Hemment, S. Cristoloveanu, K. Izumi, T. Houston, and S. Wilson, eds., *Silicon-on-Insulator Technology and Devices VII*, Proc. Electrochemical Society, Vol. 96-3, 1996, p. 34.
27. B. Aspar, M. Bruel, H. Moriceau, C. Maleville, T. Poumeyrol, A. M. Papon, A. Claverie, G. Benassayag, A. J. Auberton-Hervé, and T. Barge, "Basic Mechanisms Involved in the Smart-Cut Process," *Microelectro. Eng.* **36**, 233 (1997).
28. B. Aspar, M. Bruel, M. Zussy, and A. M. Cartier, "Transfer of Structured and Patterned Thin Silicon Films Using the Smart-Cut Process," *Electron. Lett.* **32**, 1985 (1996).
29. A. J. Auberton-Hervé, J. M. Lamure, T. Barge, M. Bruel, B. Aspar, and J. L. Pelloie, "SOI Materials for ULSI Applications," *Semiconduct. Int.* **97**, (Oct. 1995).
30. See, for example, S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.
31. H. K. Lim and J. G. Fossum, "Threshold Voltage of Thin-Film Silicon-on-Insulator SOI MOSFETs," *IEEE Trans. Electron Devices* **30**, 1244 (1983).
32. J. P. Colinge, "Conduction Mechanisms in Thin-Film, Accumulation-Mode p-Channel SOI MOSFETs," *IEEE Trans. Electron Devices* **37**, 718 (1990).
33. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967, p. 326.
34. F. Silveira, D. Flandre, and P. G. A. Jespers, "A g_m/I_D Based Methodology for the Design of CMOS Analog Circuits and Its Application to the Synthesis of a Silicon-on-Insulator Micropower OTA," *IEEE J. Solid-State Circ.* **31**(9), 1314 (1996).
35. E. A. Vittoz, "Low-Power Designs: Ways to Approach the Limits," *Tech. Digest Papers of ISSCC*, Jan. 1994, p. 14.
36. D. Flandre, L. F. Ferreira, P. G. A. Jespers, and J. P. Colinge, "Modelling and Application of Fully Depleted SOI MOSFETs for Low Voltage, Low Power Analogue CMOS Circuits," *Solid-State Electron.* **39**(4), 455 (1996).
37. S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1981, p. 446.
38. R. Gillon, J. P. Raskin, D. Vanhoenacker, and J. P. Colinge, "Modelling and Optimising the SOI MOSFET in View of MMIC Applications," *Proc. Eur. Microwave Conf.* Bologna, Italy, 1995, p. 543.
39. A. L. Caviglia, R. C. Potter, and L. J. West, "Microwave Performances of SOI n-MOSFETs and Coplanar Waveguides," *IEEE Electron Device Lett.* **12**, 26 (1991).
40. A. K. Agarwal, M. C. Driver, M. H. Hanes, H. M. Hobgood, P. G. McMullin, H. C. Nathanson, T. W. O'Keefe, T. J. Smith, J. R. Szedon, and R. N. Thomas, "MICROXTM—An Advanced Silicon Technology for Microwave Circuits up to X-Band," *Tech. Digest IEDM*, Dec. 1991, p. 687.
41. M. H. Hanes, A. K. Agarwal, T. W. O'Keefe, H. M. Hobgood, J. R. Szedon, T. J. Smith, R. R. Siergie, P. G. McMullin, H. C. Nathanson, M. C. Driver, and R. N. Thomas, "MICROXTM—An All-silicon Technology for Monolithic Microwave Integrated Circuits," *IEEE Electron Device Lett.* **14**, 219 (1993).
42. J. P. Colinge, J. Chen, D. Flandre, J. P. Raskin, R. Gillon, and D. Vanhoenacker, "A Low-Voltage, Low-Power Microwave SOI MOSFET," *Proc. IEEE Int. SOI Conf.* Oct. 1996, p. 128.

43. J. Chen, J. P. Colinge, D. Flandre, R. Gillon, J. P. Raskin, and D. Vanhoenacker, "Comparison of TSi_2 , CoSi_2 , and NiSi for Thin-Film Silicon-on-Insulator Applications," *J. Electrochem. Soc.* **144**, 2437 (1997).
44. A. Hürriich, P. Hübler, D. Eggert, H. Kück, W. Barthel, W. Budde, and M. Raab, "SOI-CMOS Technology with Monolithically Integrated Active and Passive RF Devices on High Resistivity SIMOX Substrates," *Proc. IEEE Int. SOI Conf.* Oct. 1996, p. 130.
45. D. Eggert, P. Huebler, A. Huerrich, H. Kueck, W. Budde, and M. Vorwerk, "A SOI-RF-CMOS Technology on High-Resistivity SIMOX Substrates for Microwave Applications up to 5 GHz," *IEEE Trans. Electron Devices* **44**, 1981 (1997).
46. C. Wann, F. Assaderaghi, L. Shi, K. Chan, S. Cohen, H. Hovel, K. Jenkins, Y. Lee, D. Sadana, R. Viswanathan, S. Wind, and Y. Taur, "High-Performance $0.007\ \mu\text{m}$ CMOS with 9.5-ps Gate Delay and 150 GHz f_T ," *IEEE Electron Device Letters*, **13**, 625 (1997).
47. I. Rahim, I. Lim, J. Foerstner, and B. Y. Hwang, "Comparison of SOI Versus Bulk Silicon Substrate Crosstalk Properties for Mixed-Mode ICs," *Proc. IEEE Int. SOI Conf.* Oct. 1992, p. 170.
48. J. P. Colinge, "An SOI Voltage-Controlled Bipolar-MOS Device," *IEEE Trans. Electron Devices* **34**(4), 845 (1987).
49. S. Verdonck-Vandenbroeck, S. S. Wong, and P. K. Ko, "High Gain Lateral Bipolar Transistor," *Tech. Digest IEDM*, 1988, p. 406.
50. R. Huang, Y. Y. Wang, and R. Han, "Analytical Modeling for the Collector Current in SOI Gate-Controlled Hybrid Transistor," *Solid-State Electron.* **39**(12), 1816 (1996).
51. J. P. Colinge, "SOI Technology," in G. A. S. Machado, ed., *Low-power HF Microelectronics: A Unified Approach*, IEE Circuits and Systems series 8, the Institution of Electrical Engineers, 1996, p. 139.
52. J. P. Colinge, "Thin-Film SOI Technology: The Solution to Many Submicron CMOS Problems," *Tech. Digest IEDM*, Dec. 1994, p. 817.
53. J. P. Colinge, D. Flandre, D. De Ceuster, "P⁺-P-P⁺ Pseudo-Bipolar Lateral SOI Transistor," *Electron. Lett.* **30**, 1543 (1994).
54. J. P. Colinge, "Voltage Controlled Bipolar-MOS Ring Oscillator," *Electron. Lett.* **23**, 1023 (1987).
55. F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. K. Ko, and C. Hu, "A Dynamic Threshold Voltage MOSFET (DTMOS) for Ultra-Low Voltage Operation," *Tech. Digest IEDM*, Dec. 1994, p. 809.
56. T. Douseki, S. Shigematsu, Y. Tanabe, M. Harada, H. Inokawa, and T. Tsuchiya, "A 0.5V SIMOX-MTCMOS Circuit with 200ps Logic Gate," *Digest Tech. Papers ISSCC*, 1996, p. 84.
57. M. Harada, T. Douseki, and T. Tsuchiya, "Suppression of Threshold Voltage Variation in MTCMOS/SIMOX Circuit Operating below 0.5 V," *Digest Tech. Papers, Symp. VLSI Technology*, 1996, p. 96.
58. M. H. Gao, J. P. Colinge, L. Lauwers, S. Wu, and C. Claeys, "Twin-MOSFET Structure for Suppression of the Kink and Parasitic Bipolar Effect in SOI MOSFET's at Room and Liquid Helium Temperatures," *Solid-State Electron.* **35**, 505 (1992).
59. T. Sekigawa and Y. Hayashi, "Calculated Threshold Voltage Characteristics of an XMOS Transistor Having an Additional Bottom Gate," *Solid-State Electron.* **27**, 827 (1984).

60. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Short Can Si Go?" *Tech. Digest IEDM*, 1992, p. 553.
61. Y. Omura, S. Horiguchi, M. Tabe, and K. Kishi, "Quantum-Mechanical Effects on the Threshold Voltage of Ultrathin SOI nMOSFETs," *IEEE Electron Device Lett.* **14**, 569 (1993).
62. K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling Theory for Double-Gate SOI MOSFETs," *IEEE Trans. Electron Devices* **40**, 2326 (1993).
63. F. Balestra, S. Cristoloveanu, M. Benachir, and T. Elewa, "Double-Gate Silicon-on-Insulator Transistor with Volume Inversion: A New Device with Greatly Enhanced Performance," *IEEE Electron Device Lett.* **8**, 410 (1987).
64. P. Francis, A. Terao, D. Flandre, and F. Van de Wiele, "Modeling of Ultrathin Double-Gate nMOS/SOI Transistors," *IEEE Trans. Electron Devices* **41**, 715 (1994).
65. A. Terao, D. Flandre, E. Lora-Tamayo, and F. Van de Wiele, "Measurement of Threshold Voltages of Thin-film Accumulation-Mode PMOS/SOI Transistors," *IEEE Electron Device Lett.* **12**, 682 (1991).
66. T. Tanaka, K. Suzuki, H. Horie, and T. Sugii, "Ultrafast Operation of V_{th} -Adjusted p^+-n^+ Double-Gate SOI MOSFETs," *IEEE Electron Device Lett.* **15**, 386 (1994).
67. J. P. Denton and G. W. Neudeck, "Fully Depleted Dual-Gated Thin-Film SOI P-MOSFETs Fabricated in SOI Islands with an Isolated Buried Polysilicon Back Gate," *IEEE Electron Device Lett.* **17**, 509 (1996).
68. J. P. Colinge, M. H. Gao, A. Romano, H. Maes, and C. Claeys, "Silicon-on-Insulator Gate-All-Around Device," *Tech. Digest IEDM*, 1990, p. 595.
69. K. Watanabe, T. Hashimoto, M. Yoshida, M. Usami, Y. Sakai, and T. Ikeda, "A Bonded-SOI Bipolar Process Technology," in U. Gösele, T. Abe, J. Haisma, and M. A. A. Schmidt, eds., *Semiconductor Wafer Bonding: Science, Technology, and Applications*, Proc. Electrochemical Society, Vol. 92-7, 1992, p. 443.
70. J.P. Colinge, "Half-Micron-Base Lateral Bipolar Transistors Made in Thin Silicon-on-Insulator Films," *Electron. Lett.* **22**, 886 (1986).
71. W. M. Huang, K. Klein, M. Grimaldi, M. Racanelli, S. Ramaswami, J. Tsao, J. Foerstner, and B. Y. Hwang, "TFSOI BiCMOS Technology for Low-Power Applications," *Tech. Digest IEDM*, 1993, p. 449.
72. S. A. Parke, C. Hu, and P. K. Ko, "Bipolar-FET Hybrid-Mode Operation of Quarter-Micrometer SOI MOSFETs," *IEEE Electron Device Lett.* **14**, 234 (1993).
73. S. A. Parke, C. Hu, and P. K. Ko, "Bipolar-FET Hybrid-Mode Operation of Quarter-Micrometer SOI MOSFETs," *IEEE Electron Device Lett.* **14**, 234 (1993).
74. R. Dekker, W. T. A. v. d. Einden, and H. G. R. Maas, "An Ultra Low Power Lateral Bipolar Polysilicon Emitter Technology on SOI," *Tech. Digest IEDM*, Dec. 1993, p. 75.
75. W. M. Huang, K. Klein, M. Grimaldi, M. Racanelli, S. Ramaswami, J. Tsao, J. Foerstner, and B. Y. Hwang, "TFSOI BiCMOS Technology for Low Power Applications," *Tech. Digest IEDM*, Dec. 1993, p. 449.
76. G. G. Shahidi, D. D. Tang, B. Davari, Y. Taur, P. McFarland, K. Jenkins, D. Danner, M. Rodriguez, A. Megdanis, E. Petrillo, M. Polcari, and T. H. Ning, "A Novel High-Performance Lateral Bipolar on SOI," *Tech. Digest IEDM*, Dec. 1991, p. 663.
77. U. Magnusson, H. Norström, W. Kaplan, S. Zhang, M. Jargelius, and D. Sigurd, "High Frequency Bipolar Transistor on SIMOX," in J. Borel, P. Gentil, J. P. Noblanc, A.

- Noailhat, and M. Verdone, eds., *Proc. 23rd ESSDERC*, Editions Frontières, Paris, 1993, p. 683.
78. S. Feindt, J. J. Hajjar, M. Smrtic, and J. I. Aphon, "A Complementary Bipolar Process on Bonded Wafers," in M. A. Schmidt, C. E. Hunt, T. Abe, and H. Baumgart, eds., *Semiconductor Wafer Bonding: Science, Technology, and Applications*, Electrochemical Society Proc. Vol. 93-29, 1993, p. 189.
 79. J. C. Sturm and J. F. Gibbons, "Vertical Bipolar Transistors in Laser-Recrystallized Polysilicon," *IEEE Electron Device Lett.* **6**, 400 (1985).
 80. T. Arnborg and A. Litwin, "Analysis of New High-Voltage Bipolar Silicon-on-Insulator Transistor with Fully Depleted Collector," *IEEE Trans. Electron Devices* **42**, 172 (1995).
 81. K. Yallup, S. Edwards, and O. Creighton, "A Novel Bipolar Device on SOI Wafers for Analog BiCMOS Applications," *Proc. 24th ESSDERC*, in C. Hill and P. Ashburn, eds., Editions Frontières, Paris, 1994, p. 565.
 82. J. P. Colinge, X. Baie, and V. Bayot, "Evidence of Two-Dimensional Carrier Confinement in Thin n-Channel Gate-All-Around (GAA) Devices," *IEEE Electron Device Lett.* **15**, 193 (1994).
 83. X. Baie and J. P. Colinge, "Two-Dimensional Confinement Effects in Gate-All-Around (GAA) MOSFETs," *Solid-State Electron.* **42**, 499 (1998).
 84. Y. Nakajima, Y. Takahashi, S. Horiguchi, K. Iwadate, H. Namatsu, K. Kurihara, and M. Tabe, "Quantized Conductance of a Silicon Wire Fabricated Using SIMOX Technology," *Ext. Abstr. SSDM*, Yokohama, Japan, 1994, p. 538.
 85. Y. Nakajima, Y. Takahashi, S. Horiguchi, K. Iwadate, H. Namatsu, K. Kurihara, and M. Tabe, "Quantized Conductance of a Silicon Wire Fabricated by Separation-by-Implanted-Oxygen Technology," *Jpn. J. Appl. Phys.* **34**, 1309 (1995).
 86. M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, "Generalized Many-Channel Conductance Formula with Application to Small Rings," *Phys. Rev. B* **31**, 6207 (1985).
 87. X. Baie, J. P. Colinge, V. Bayot, and E. Grivei, "Quantum-Wire Effects in Thin and Narrow SOI MOSFETs," *Proc. IEEE Int. SOI Conf.* Oct. 1995, p. 66.
 88. Y. Takahashi, H. Namatsu, K. Kurihara, K. Iwadate, M. Nagase, and K. Murase, "Size Dependence of the Characteristics of Si Single-Electron Transistors on SIMOX Substrates," *IEEE Trans. Electron Devices* **43**, 1213 (1996).
 89. K. Murase, Y. Takahashi, Y. Nakajima, H. Namatsu, M. Nagase, K. Kurihara, K. Iwadate, S. Horiguchi, M. Tabe, and K. Izumi, "Transport Properties of Silicon Nanostructures Fabricated on SIMOX Substrates," *Microelectron. Eng.* **28**, 399 (1995).
 90. A. Nakajima, T. Futatsugi, K. Kosemura, T. Fukano, and N. Yokoyama, "Room Temperature Operation of Si Single-Electron Memory with Self-Aligned Floating Dot Gate," *Tech. Digest IEDM*, Dec. 1996, p. 952.
 91. P. Francis, A. Terao, B. Gentinne, D. Flandre, and J. P. Colinge, "SOI Technology for High-Temperature Applications," *Tech. Digest IEDM*, Dec. 1992, p. 353.
 92. A. J. Auberton-Hervé, "SIMOX-SOI Technologies for High-Speed and Radiation-Hard Technologies: Status and Trends in VLSI and ULSI Applications," in D. Schmidt, ed., *Proc. 4th Int. Symp. Silicon-on-Insulator Technology and Devices*, The Electrochemical Society, Vol. 90-6, 1990, p. 455.

93. E. D. Nowak, L. Ding, Y. T. Loh, and C. Hu, "Speed, Power, and Yield Comparison of Thin Bonded SOI versus Bulk CMOS Technologies," *Proc. IEEE Int. SOI Conf. Oct. 1994*, p. 41.
94. M. Fujishima, K. Asada, Y. Omura, and K. Izumi, "Low-Power 1/2 Frequency Dividers Using 0.1- μm CMOS Circuits Built with Ultrathin SIMOX Substrates," *IEEE J. Solid-State Circ.* **28**, 510 (1993).
95. Y. Kado, M. Suzuki, K. Koike, Y. Omura, and K. Izumi, "A 1-GHz/0.9-mW CMOS/SIMOX Divide-by-128/129 Dual-modulus Prescaler Using a Divide-by-2/3 Synchronous Counter," *IEEE J. Solid-State Circ.* **28**, 513 (1993).
96. Y. Kado, T. Ohno, M. Harada, K. Deguchi, and T. Tsuchiya, "Enhanced Performance of Multi-GHz PLL LSIs Using Sub-1/4-Micron Gate Ultrathin Film CMOS/SIMOX Technology with Synchrotron X-ray Lithography," *Tech. Digest IEDM*, Dec. 1993, p. 243.
97. G. G. Shahidi, T. H. Ning, T. I. Chappell, J. H. Comfort, B. A. Chappell, R. Franch, C. J. Anderson, P. W. Cook, S. E. Schuster, M. G. Rosenfield, M. R. Polcari, R. H. Dennard, and B. Davari, "SOI for a 1-Volt CMOS Technology and Applications to a 512 kb SRAM with 3.5 ns Access Time," *Tech. Digest IEDM*, Dec. 1993, p. 813.
98. T. Eimori, T. Oashi, H. Kimura, Y. Yamaguchi, T. Iwamatsu, T. Tsuruda, K. Suma, H. Hidaka, Y. Inoue, S. Satoh, and H. Miyoshi, "ULSI DRAM/SIMOX with Stacked Capacitor Cells for Low-Voltage Operation," *Tech. Digest IEDM*, 1993, p. 45.
99. W. M. Huang, K. Papworth, M. Racanelli, J. P. John, J. Foerstner, H. C. Shin, H. Park, B. Y. Hwang, T. Wetteroth, S. Hong, H. Shin, S. Wilson, and S. Cheng, "TFSOI CMOS Technology for Sub-1V Microcontroller Circuits," *Tech. Digest IEDM*, 1995, p. 59.
100. T. Douseki, S. Shigematsu, Y. Tanabe, M. Harada, H. Inokawa, and T. Tsuchiya, Digest of Technical Papers, "A 0.5V SIMOX-MTCMOS Circuit with 200ps Logic Gate," *IEEE Int. Solid-State Circuits Conf.* Jan. 1996, p. 84.
101. M. Ino, H. Sawada, K. Nishimura, M. Urano, H. Suto, S. Date, T. Ishiara, T. Takeda, Y. Kado, H. Inokawa, T. Tsuchiya, Y. Sakakibara, Y. Arita, K. Izumi, K. Takeya, and T. Sakai, "0.25 μm CMOS/SIMOX Gate Array LSI," *Digest of Technical Papers, IEEE Int. Solid-State Circuits Conf.* Jan. 1996, p. 86.
102. T. Fuse, Y. Oowaki, M. Terauchi, S. Watanabe, M. Yoshimi, K. Ohuchi, and J. Matsunaga, "0.5V SOI CMOS Pass-Gate Logic," *Digest of Technical Papers, IEEE Int. Solid-State Circuits Conf.* 1996, p. 88.
103. A. O. Adan, T. Naka, S. Kaneko, D. Urabe, K. Higashi, and A. Kasigawa, "Device Integration of a 0.35 μm CMOS on Shallow SIMOX Technology for High-Speed and Low-Power Applications," *Proc. IEEE Int. SOI Conf.* 1996, p. 116.
104. T. Fuse, Y. Oowaki, Y. Yamada, M. Kamoshida, M. Ohta, T. Shino, S. Kawanaka, M. Terauchi, T. Yoshida, G. Matsubara, S. Oshioka, S. Watanabe, M. Yoshimi, K. Ohuchi, and S. Manabe, "A 0.5V 200MHz 1-Stage 32b ALU Using a Body Bias Controlled SOI Pass-Gate logic," *Digest Tech. Papers, Int. Solid-State Circuits Conf.* 1997, p. 286.
105. Y. Ohtomo, S. Yasuda, M. Nogawa, J. Inoue, K. Yamakoshi, H. Sawada, M. Ino, S. Hino, Y. Sato, Y. Takei, T. Watanabe, and K. Takeya, "A 40 Gb/s 8 \times 8 ATM Switch LSI Chip Using 0.25 μm CMOS/SIMOX," *Digest Tech. Papers, Int. Solid-State Circuits Conf.* 1997, p. 154.

106. Semiconductor Business News, Aug. 3, 1998; <http://www.semibiznews.com/stories/8h03ibm.htm>.
107. Semiconductor Business News, Sept. 14, 1998; <http://www.semibiznews.com/stories98/sep98x/8i14sams.htm>.
108. F. Silveira, D. Flandre, and P. G. A. Jespers, "A g_m/I_D Based Methodology for the Design of CMOS Analog Circuits and its Application to the Synthesis of a Silicon-on-Insulator Micropower OTA," *IEEE J. Solid-State Circ.* **31**, 1314 (1996).
109. D. Flandre, B. Gentinne, J. P. Eggermont, and P. G. A. Jespers, "Design of Thin-Film Fully Depleted SOI CMOS Analog Circuits Significantly Outperforming Bulk Implementations," *Proc. IEEE Int. SOI Conf.* Oct. 1994, p. 99.
110. Y. Omura and K. Izumi, "A New Model of Switching Operation in Fully Depleted Ultrathin-Film CMOS/SIMOX," *IEEE Electron Device Lett.* **12**, 655 (1991).
111. J. Weyers, H. Vogt, M. Berger, W. Mach, B. Mütterlein, M. Raab, F. Richter, and F. Vogt, "Integration of Vertical/Quasivertical DMOS, CMOS and Bipolar Transistors in a 50 V SIMOX Process," *Microelectron. Eng.* **19**, 733 (1992).
112. C. Harendt, U. Apel, T. Ifström, H. G. Graf, and B. Höfflinger, "Bonded-Wafer SOI Smart Power Circuits in Automotive Applications," in M. A. Schmidt, C. E. Hunt, T. Abe, and H. Baumgart, eds., *Proc. 2nd Int. Symposium on "Semiconductor Wafer Bonding: Science, Technology, and Applications"*, The Electrochemical Society Proceedings, Vol. 93-29, 1993, p. 129.
113. H. Pein, E. Arnold, H. Baumgart, R. Egloff, T. Letavic, S. Merchant, and S. Mukherjee, "SOI High Voltage LDMOS and LIGBT Transistors with a Buried Diode and Surface p-Layer," *Proc. IEEE Int. SOI Conf.* Oct. 1992, p. 146.
114. F. Vogt, B. Mütterlein, and H. Vogt, "An Intelligent 500 V Power Vertical DMOS on SIMOX Substrate," in W. E. Bayley, ed., *Proc. 5th Int. Symp. Silicon-on-Insulator Technology and Devices*, Electrochemical Society Proceedings, Vol. 92-13, 1992, p. 77.
115. W. Wondrak, E. Stein, and R. Held, "Influence of the Back-gate Voltage on the Breakdown Voltage of SOI Power Devices," in U. Gösele, T. Abe, J. Haisma, and M. A. A Schmidt, eds., *Semiconductor Wafer Bonding: Science, Technology, and Applications*, Electrochemical Society Proceedings, Vol. 92-7, 1992, p. 427.
116. W. P. Maszara, D. Boyko, A. Caviglia, G. Goetz, J. B. McKitterick, and J. O'Connor, "Smart-Power Solenoid Driver for 300°C Operation," *Proc. IEEE Int. SOI Conf.* Oct. 1995, p. 131.
117. M. Radecker, H. L. Fiedler, F. P. Vogt, and H. Vogt, "Single-Chip Class-E Converter for Compact Fluorescent Lamp Ballast," *Digest Tech. Papers ISSCC*, Jan. 1997, p. 378.
118. J. Weyers, H. Vogt, M. Berger, W. Mach, B. Mütterlein, M. Raab, F. Richter, and F. Vogt, "Integration of Vertical/Quasivertical DMOS, CMOS and Bipolar Transistors in a 50V SIMOX Process," *Microelectron. Eng.* **19**, 733 (1992).
119. E. Arnold, H. Pein, and S. P. Herko, "Comparison of Self-Heating Effects in Bulk-Silicon and SOI High-Voltage Devices," *Tech. Digest IEDM*, Dec. 1994, p. 813.
120. K. Izumi, Y. Omura, M. Ishikawa, and E. Sano, "SIMOX Technology for CMOS LSIs," *Tech. Digest Symp. VLSI Technology*, June 1982, p. 10.
121. W. A. Krull and J. C. Lee, "Demonstration of the Benefits of SOI for High Temperature Operation," *Proc. IEEE SOS/SOI Technology Workshop*, Oct. 1988, p. 69.

122. T. W. Houston, H. Lu, P. Mei, T. G. W. Blake, L. R. Hite, R. Sundaresan, M. Matloubian, W. E. Bailey, J. Liu, A. Peterson, and G. Pollack, "A 1 μm CMOS/SOI 64 k SRAM with 10 nA Standby Current," *Proc. IEEE SOS/SOI Technology Workshop*, Oct. 1989, p. 137.
123. A. J. Auberton-Hervé, B. Giffard, and M. Bruel, "Performances of a 16 k SRAM Processed in a 150 nm SIMOX Film for High-Speed Applications," *Proc. IEEE SOS/SOI Technology Workshop*, Oct. 1989, p. 169.
124. W. E. Bayley, H. Lu, T. G. W. Blake, L. R. Hite, P. Mei, D. Hurta, T. W. Houston, and G. P. Pollack, "Processing and Transistor Characteristics of a 256 k SRAM Fabricated on SIMOX," *Proc. IEEE Int. SOI Conf.* Oct. 1991, p. 134.
125. L. K. Wang, J. Seliskar, T. Bucelot, A. Edenfeld, and N. Haddad, "Enhanced Performance of Accumulation-Mode 0.5 μm CMOS/SOI Operated at 300 K and 85 K," *Tech. Digest IEDM*, Dec. 1991, p. 679.
126. H. Lu, E. Yee, L. Hite, T. Houston, Y. D. Sheu, R. Rajgopal, C. C. Chen, J. M. Hwang, and G. Pollack, "A 1 Mbit SRAM on SIMOX Material," *Proc. IEEE Int. SOI Conf.* Oct. 1993, p. 182.
127. G. G. Shahidi, T. H. Ning, T. I. Chappell, J. H. Comfort, B. A. Chappell, R. Franch, C. J. Anderson, P. W. Cook, S. E. Schuster, M. G. Rosenfield, M. R. Polcari, R. H. Dennard, and B. Davari, "SOI for a 1-Volt CMOS Technology and Applications to a 512 kb SRAM with 3.5 ns Access Time," *Tech. Digest IEDM*, Dec. 1993, p. 813.
128. T. Eimori, T. Oashi, H. Kimura, Y. Yamaguchi, T. Iwamatsu, T. Tsuruda, K. Suma, H. Hidaka, Y. Inoue, T. Nishimura, S. Satoh, and M. Miyoshi, "ULSI DRAM/SIMOX with Stacked Capacitor Cells for Low-Voltage Operation," *Tech. Digest IEDM*, Dec. 1993, p. 45.
129. H.-S. Kim, S.-B. Lee, D.-U. Choi, J.-H. Shim, K.-C. Lee, K.-P. Lee, K.-N. Kim, and J.-W. Park, "A High-Performance 16M DRAM on a Thin-Film SOI," *Digest Tech. Papers of Symp. VLSI Technology*, June 1995, p. 143.
130. T. Oashi, T. Eimori, F. Morishita, T. Iwamatsu, Y. Yamaguchi, F. Okuda, K. Shimomura, H. Shimano, S. Sakashita, K. Arimoto, Y. Inoue, S. Komori, M. Inuishi, T. Nishimura, and H. Miyoshi, "16 Mb DRAM/SOI Technologies for Sub-1V Operation," *Tech. Digest IEDM*, Dec. 1996, p. 609.
131. K. Shimomura, H. Shimano, F. Okuda, N. Sakashita, T. Oashi, Y. Yamaguchi, T. Eimori, M. Inuishi, K. Arimoto, S. Maegawa, Y. Inoue, T. Nishimura, S. Komori, K. Kyuma, A. Yasuoka, and H. Abe, "A 1 V 46 ns 16 Mb SOI DRAM with Body Control Technique," *Digest Tech. Papers, ISSCC*, Jan. 1997, p. 68.
132. Y. H. Koh, M. R. Oh, J. W. Lee, J. W. Yang, W. C. Lee, C. K. Park, J. B. Park, Y. C. Heo, K. M. Rho, B. C. Lee, M. J. Chung, M. Huh, H. S. Kim, K. S. Choi, W. C. Lee, J. K. Lee, K. H. Ahn, K. W. Park, J. Y. Yang, H. K. Kim, D. H. Lee, and I. S. Hwang, "1 Giga Bit SOI DRAM with Fully Bulk Compatible Process and Body-Contacted SOI MOSFET Structure," *Tech. Digest IEDM*, Dec. 1997, p. 579.
133. Y. Hu, C. Teng, T. W. Houston, K. Joyner, and T. J. Aton, "Design and Performance of SOI Pass Transistors for 1Gbit DRAMs," *Digest Tech. Papers, Symp. VLSI Technology*, June 1996, p. 128.
134. A. Chatterjee, J. Liu, S. Aur, P. K. Mozumder, M. Rodder, and I. C. Chen, "Pass Transistor Designs Using Pocket Implant to Improve Manufacturability for 256 Mbit DRAM and Beyond," *Tech. Digest IEDM*, Dec. 1994, p. 87.

PROBLEMS

- 5.1 Calculate the body factor, and hence, the subthreshold slope (near threshold) in a bulk n-channel MOSFET, a fully depleted n-channel MOSFET, and a fully depleted n-channel MOSFET with accumulated back interface. The gate oxide thickness, doping density, box thickness and silicon film thickness are 15 nm, $7 \times 10^{16} \text{ cm}^{-3}$, 400 and 80 nm, respectively; $T = 300 \text{ K}$. There are no interface states.
- 5.2 The variation of threshold voltage with back-gate voltage of a fully depleted SOI n-channel MOSFET is -60 mV/V . The gate oxide thickness is 10 nm and the buried oxide thickness is 400 nm. What is the thickness of the silicon film?
- 5.3 Consider the bulk and fully depleted SOI transistors of Problem 5.1. The minimum of potential in the SOI MOSFET is close to the bottom of the device. Calculate the threshold voltage of each device, the saturation current for a supply voltage of 1 V ($V_G = V_{DS} = 1 \text{ V}$). Calculate the OFF current of the device at $V_G = 0 \text{ V}$ and $V_{DS} = 1 \text{ V}$, assuming $W = L = 1 \mu\text{m}$, assuming the current flowing through the devices at threshold, I_{Dth} , is equal to 100 nA, and assuming a perfect exponential dependence of the drain current on gate voltage below threshold. The workfunction difference between the gate material and the silicon is -0.7 V ; the electron channel mobility is $550 \text{ cm}^2/(\text{V} \cdot \text{s})$.
- 5.4 Consider a ring oscillator composed of 13 identical CMOS inverters. The n-channel transistors are identical to those of Problem 5.3. The p-channel transistors have a $W = 3 \mu\text{m}$ and a $L = 1 \mu\text{m}$ such that the saturation current of the p-channel devices is identical to that of the n-channel devices. The area of source and drain junctions is $W \times 3 \mu\text{m}$, the source and drain doping concentration is 10^{20} cm^{-3} , and $V_{DD} = 1 \text{ V}$. Neglecting the capacitance of the metal lines between the inverters, calculate the oscillation frequency.
- 5.5 The average power dissipated by a digital circuit is given by the sum of the static and dynamic power dissipation:

$$P_{\text{total}} = P_{\text{stat}} + P_{\text{dyn}} = I_{\text{OFF}} V_{DD} + f C_L V_{DD}^2$$

where V_{DD} is the supply voltage, f the input signal frequency, and C_L the load capacitance. Using data and result from Problems 5.3 and 5.4, calculate the power dissipation of a bulk and a fully depleted SOI CMOS inverter in the standby mode and switching at 10 MHz.

- 5.6 The CMOS analog switch (pass gate) provides a rough estimation of the lowest acceptable analog supply voltage, V_{DD} , that can be used in a circuit. In order to transmit a signal without alteration through a switch with the gate of the n-MOS transistor held at V_{DD} and that of the p-MOS transistor at 0 V, V_{DD} must be larger than a minimum value that can be expressed as a function

of the threshold voltage V_T and body factor n of the n- and p-MOSFETs. The following relationship must be satisfied:

$$V_{DD} \geq \frac{V_{Tp} \cdot n_n + V_{Tn} \times n_p}{n_p + n_n - n_p \times n_n}$$

Calculate the minimum power supply voltage at which a pass gate made using the transistors of Problem 5.3 will operate, assuming the p-channel transistors have the same threshold voltage and the same body-effect coefficient as the n-channel transistors.

- 5.7** The open-loop dc gain of a simple operational transconductance amplifier (OTA) is given by $A_{vo} = V_A(g_m/I_D)$, and the gain–bandwidth product is given by $GBW = g_m/2\pi C_L$. Assuming an Early voltage, V_A , of 50 V, and using device parameters of Problem 5.3, calculate the maximum dc gain and the gain–bandwidth product (best case) for bulk and fully depleted SOI OTAs having a load capacitance of 5 pF and a gate voltage of 1 V.
- 5.8** Up to what temperature can one expect to observe 2DEG quantization effects in GAA transistors having thicknesses of 8, 10, 20, 50, and 100 nm, respectively, when volume inversion is observed?
- 5.9** How long does it take to form a 150-mm SIMOX wafer (standard dose) using either a medium-current ($I = 1 \mu\text{A}$) oxygen implanter or a high-current oxygen implanter ($I = 40 \text{ mA}$)?
- 5.10** Calculate the variation of threshold voltage with temperature, dV_{th}/dT , in a bulk MOSFET and a fully depleted SOI device. The gate material is n^+ polysilicon, the gate oxide thickness is 19 nm, and the channel doping concentration is $1.6 \times 10^{17} \text{ cm}^{-3}$; $T = 300 \text{ K}$.

The Hot-Carrier Effect

BRIAN DOYLE

Intel Corporation
Hillsboro, OR

6.1 INTRODUCTION

The hot-carrier effect (HCE) and its consequences on transistor operation have been known and examined since the late 1970s^{1,2}. Because this effect leads to a gradual change in the drive current and threshold voltages of devices and in time, leads to potential circuit failure, considerable effort has been expended to understand, characterize, and reduce either the hot carrier, or the susceptibility of the transistor to this type of long-term reliability concern.

In this chapter we discuss the hot-carrier effect in n- and p-channel transistors. After a brief introduction to carrier heating (Section 6.1), the types of damage existing in the oxide/substrate interface are discussed (Section 6.2), followed by a description and characterization of the damage occurring for the different gate and drain voltage conditions of stress (Section 6.3).

Having explored the damage species generated for each of the different gate stress voltage conditions (and high drain voltage), we discuss the lifetimes of devices for each damage species (each gate voltage region of stress), and these are linked to transistors working in a circuit environment: ac stress (Section 6.4).

Following this, the different methods of characterizing hot-carrier damage that have been published are briefly explored in Section 6.5. Section 6.6 deals with the various “structural” aspects of the transistor, such as gate oxide thickness scaling, poly-length scaling, and junction design, while the processing aspects that have been found to affect the transistor’s hot-carrier hardness are dealt with in Section 6.7. This includes oxide processing (the various flavors of nitridation), plasma processing effects (antenna damage), and other processing steps that affect the HC properties of the device.

Finally some conclusion will be drawn in Section 6.8 regarding the hot-carrier effect, and the trends expected in the future.

6.1.1 Hot-Carrier Heating

With the reduction of gate lengths to deep submicrometer dimensions, the problem of the quality of the gate oxide and its interface with the silicon substrate has been of increasing concern in the fabrication of CMOS devices. This is due to what is called the hot-carrier effect (HCE). The HCE arises from the lateral field between the source and drain of the MOSFET (Fig. 6.1). At the pinchoff point of the transistor, toward the drain junction edge, the electric field starts to rise rapidly. This rise in electric field causes the carriers to gain energy from the field, thus increasing the average energy of electrons in n-MOSFET channels and of holes in p-MOSFET channels. In addition, scattering processes ensure not only that some carriers have energies lower than average but that some have higher than average energies. High-energy carriers in the tail of the energy distribution, which can be approximated by a Maxwell-Boltzmann distribution, are often referred to as “hot carriers.” When sufficient numbers of carriers gain sufficient energy to create impact-ionized electron-hole pairs, the fraction of the carriers that gain this energy is given by

$$\frac{I_b}{I_s} = K e^{-(\Phi_i/q\lambda E)} \quad (6.1)$$

where K is a constant, I_b and I_s are the substrate (base) and source currents respectively, Φ_i is the threshold energy to cause impact ionization in the silicon (1.6 eV), λ is the mean free path of the carriers, q is the charge of an electron, and E is

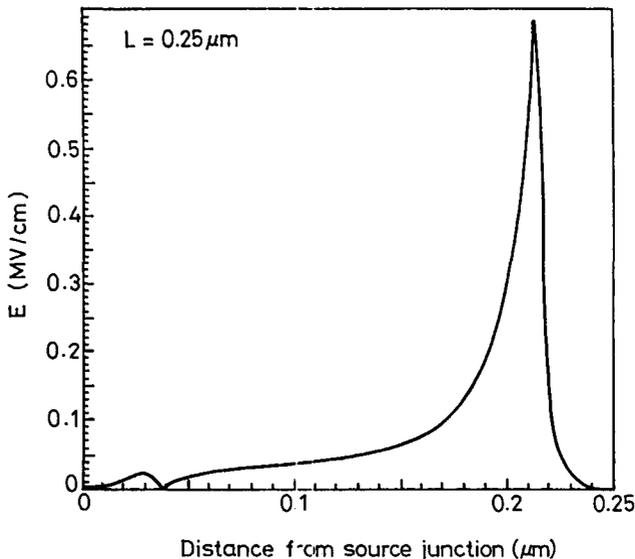


Figure 6.1. Lateral electric field versus distance between the source and drain of a 0.25- μm n-MOS transistor with 2.0 V applied to the drain, showing the high field region at the drain junction.

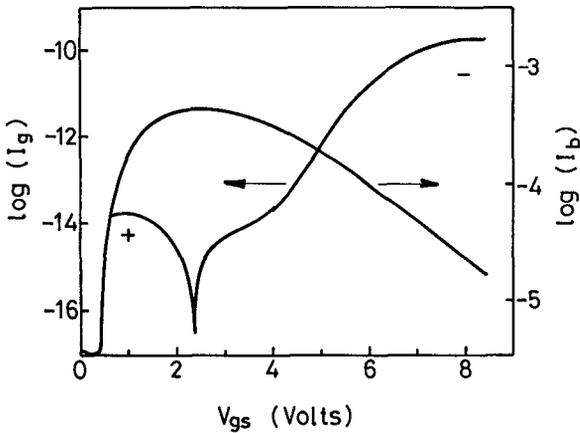


Figure 6.2 Substrate and gate currents as a function of gate voltage for an n-MOS transistor under high drain biases. (After Mistry and Doyle, Ref. 55, © 1993 IEEE.)

the effective electrical field. Since the transistor is biased from the saturated to the linear region (low gate voltages to high gate voltages) with a fixed drain voltage, the electric field maximum falls, at the same time the drain current increases. These two competing trends produce the well-known peak in the substrate current as a function of gate voltage that is shown in Figure 6.2.

A fraction of the channel carriers will also have sufficient energy to overcome the Si-SiO₂ potential barrier (3.1 eV), and may be injected into the gate oxide, leading to interface trap build-up and hot-carrier trapping in the oxide. The relationship linking the drain, substrate and gate currents is given by³:

$$\frac{I_g}{I_s} = \frac{I_b}{I_s} e^{-(\Phi_g/\Phi_i)} \quad (6.2)$$

where I_g is the gate current and Φ_g is the energy barrier for the injection of carriers into the gate oxide with a value of approximately 3.2 eV. Figure 6.2 also shows the gate current characteristics as a function of gate voltage. There are two components to this. At high gate voltages ($V_g = V_d$), although the lateral electric field is not high, the drain current is large, and the potential difference between drain and gate is such that all injected electron (that do not become trapped in the oxide) are collected at the gate electrode. As the gate voltage decreases, the gate field becomes repulsive to electrons, and although the flux of electrons entering the oxide continues to increase, the gate electronic current collected at the gate electrode decreases.⁴ At voltages a little below $V_g = V_d/2$, the net flux of carriers at the gate electrode is zero, and at even lower gate voltages, a hole current is seen, as the lateral electric fields are at their highest, and the oxide field is now attractive to gate hole injection.

The injected charge causes interface states and charge trapping in the oxides, and the net effect of this is to cause a gradual degradation of the MOSFET's I_d-V_d and I_d-V_g characteristics. This, in turn, causes the circuit switching characteristics to

change over time. The typical effect of hot-carrier degradation is to reduce the ON-state current in n-MOSFETs, while in p-MOSFETs, the OFF-state current is increased. This effect is discussed later in the chapter.

6.2 DAMAGE IDENTIFICATION

The types of damage that occur in HC stressing can be classified into two categories: interface states (denoted by the symbol N_{it}) and oxide-trapped charge (denoted by the symbol N_{ot}). There is also another sub-species of damage that shares some of the attributes of both interface states and oxide traps, and that is the slow, relaxable, or near-interface oxide traps (N_{niot}). Figure 6.3 sums up the types of damage that can occur in the hot-carrier stressing of transistors.

6.2.1 Interface States N_{it}

Figure 6.3 shows the types of damage that can occur in the interfacial region. In an unstressed transistor, all the bonds at the Si-SiO₂ interface are saturated (bonded to other silicon atoms, to oxygen atoms in the oxide, or to hydrogen atoms). However, when stressed under hot-carrier injection conditions, there is a transfer of energy from the hot carriers to the interface (more on these complicated mechanisms later) and a bond breaks at the interface. The atom can then have a dangling bond, broken

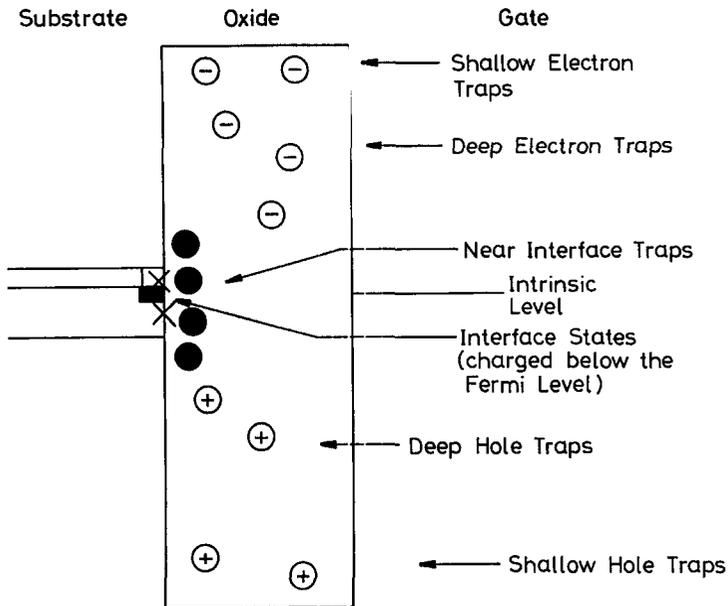


Figure 6.3 Representation of the oxide with the different types of damage that can exist in the oxide, and at the Si-SiO₂ interface.

bond, or interface state. The broken bond causes a disruption in the periodicity of the lattice at the point of the broken bond. The energy level occupied by this interface atom (interface state) slips from the continuum of states arising from the periodicity of the silicon lattice and now occupies a site in the so-called forbidden gap. This is illustrated in Figure 6.3. There are two types of interface state created. Donor states occupy the lower half of the forbidden gap give up their electron and become positively charged. They affect only p-MOS transistors. Acceptor states occupy only the upper half of the forbidden gap, accept an electron, and become negatively charged. They affect only n-MOS devices.

The effect of these interface states on the $I-V$ characteristics of the MOS transistor is to decrease the subthreshold gradient of the I_d-V_g characteristics. As the gate voltage increases in the region below threshold, the Fermi level at the Si-SiO² interface rises, causing the interface states whose energies are below the Fermi level to charge. A portion of the gate voltage now goes toward charging the interface states instead of increasing the band bending in the silicon, and the net effect is that a greater gate voltage is required to obtain the same drain current. This is true at each value of the Fermi level and the extra incremental increase in gate voltage required to charge the interface states as the Fermi level rises results in a degradation in the subthreshold gradient.

6.2.2 Oxide Traps N_{ot}

The other principal type of damage that can occur in the oxide and at the oxide-silicon interface is the oxide trap. When a charge is injected into the oxide, it can either fill preexisting traps (charging a neutral trap, or discharging a filled trap of opposite sign), or the injected charge can create a trap. In Figure 6.3, the different traps are shown as a function of energy and position within the oxide. As can be seen, there are three varieties of traps each for both hole and electron damage. The traps have, in effect, four basic characteristics:

Charge. Traps can be charged and discharged only by injection of electrons and holes from the channel or the gate. This is the principal difference between this type of damage and interface states. They are charged positively or negatively depending on the trap.

Position in Energy. Their position with respect to the oxide conduction and valence bands determines whether the traps are shallow or deep. Shallow traps can detrapp with thermal energy, while deep traps require an injected charge to change its occupancy. An exception is under Fowler-Nordheim conditions, where the voltage on the gate is sufficient to distort the conduction and valence bands of the oxide and allow the charge to tunnel out of the trap into the oxide valence or conduction band.

Position in the Oxide. Their position with respect to the oxide-silicon and oxide-gate interface—the proximity of either interface allows the charge to tunnel out (or to tunnel in). These near-interface oxide traps are discussed below.

Capture Cross-Section. This is the trap's ability to capture a charge. The traps can be thought of as having a region around them from which a carrier cannot escape, defined as the capture cross-section. Typical values run from 10^{-13} – 10^{-14} cm² for coulombic traps (charged traps that attract the free carriers being injected), to 10^{-20} – 10^{-21} cm² for repulsive traps, which gives a “characteristic” radius of the defect ranging from 100 Å (coulombic centers) to much less than 1 Å (repulsive centers) for the two ranges mentioned above.

The effect on the I_d – V_g characteristics is simply to shift the full transistor I – V characteristics to higher or lower voltages when either a negative or positive charge is trapped in the oxide.

6.2.3 Relaxable States (N_{niot})

Having described the two types of damage created in HC stressing, there is what might be considered a third type of damage, relaxable states, slow states, or near-interface oxide traps (N_{niot}). These states, as their names imply, occupy positions near the Si–SiO₂ interface (shown in Fig. 6.3). They have some of the attributes of both interface states and oxide traps in that they can charge either from the Fermi level through tunneling from the Fermi level or by injection of a charge into the oxide.

These states do not charge as rapidly as interface states, nor do they hold their charge indefinitely, as do oxide traps. Rather, the time constants for filling and emptying range from microseconds to millions of seconds. Discharging occurs through quantum-mechanical tunneling into and out of the oxide. Thus, their time constants depend exponentially on the distance of these states from the silicon-silicon dioxide interface.

It might be assumed, then, that the identification of the different types of damage would be a fairly simple task. However, the situation in short-channel transistors is more complicated than the simple picture presented above. This is because the problem is no longer one-dimensional and the position of the damage at the oxide/silicon interface between the source and drain is localized at the drain edge (where the high-field region is; see Fig. 6.1). The two-dimensional nature of the damage can cause the sub-threshold region to be completely unchanged after stress, but the postthreshold I_d – V_g characteristics can suffer significant degradation.^{6–8} This does not fit into the “classic” picture regarding damage creation that we have discussed above. The reason for this is the localization of the damage at the drain edge of the transistor. It has been shown that the type of degradation associated with interface states can even be obtained by oxide trapped charge if the damage is localized enough.⁶ Two further illustrations of the difficulties associated with damage identification are stresses at high drain currents (going toward and into snapback),⁷ and the effect of gate-edge upturn at the source and drain edges of the poly gate (due to the poly reoxidation step).⁸ Both these cases can cause the I – V characteristics to change from the classic subthreshold gradients change associated with interface states, to a change in the poststress I_d – V_g characteristics only, even though the

damage species does not change and the only effect that occurs is the localization of the damage.

With difficulties such as these in the identification of the types of damage, the role that gate voltage during stress plays was not initially appreciated. Consequently, the early steps to develop a full ac prediction model of hot-carrier stress were not fruitful. It was only with the understanding of the gate current and the role the injection species played that has led to an appreciation of the complexity of the hot-carrier stress process, and has led to a significant effort to understand, characterize, and eventually, resolve the HCE phenomenon. This will be discussed in the following sections.

6.3 GATE VOLTAGE DEPENDENCE OF STRESS

6.3.1 Intermediate Gate Voltage Stresses ($V_g = V_d/2$)

n-MOS

The two-dimensional nature of the hot-carrier effect gave rise to differences of opinion as to the nature of the damage. Opinion was divided early on between pure interface state creation,⁹ pure oxide charge trapping,^{1,2,11} and a combination of both¹² as being responsible for HC damage. Consequently, confusion reigned and it was only with the understanding of the dependence the gate voltage of stress that the present more complete picture of the hot-carrier effect was arrived at.

One of the first major observations was that the degradation behavior followed a time power law. When the damage, normally plotted as the degradation in transconductance (g_m), threshold voltage (V_t), drive current (I_{dsat}), and other variables is plotted as a function of time (Fig. 6.4a), it obeys the relation

$$\delta D = At^n \quad (6.3)$$

where D represents the degradation, n is the gradient in Figure 6.4, and has values usually between 0.5 and 0.7. The relation of Eq. 6.3 was found to hold for all drain voltages for the condition $V_g = V_d/2$. This behavior is fairly ubiquitous in hot-carrier stressing. A value of 0.5 suggests a diffusive process, and it was posited that hydrogen diffusion from the damage site might be responsible for the kinetics.³

A second observation was that the damage is gate-voltage-dependent. If a series of devices were stressed at a single drain voltage and different gate voltages, and the device characteristics measured after a given fixed time, the maximum damage tended to peak at the maximum of the substrate current. These characteristics are shown in Figure 6.5. The position of the damage peak coincides with the substrate current peak. Furthermore, this was seen to occur irrespective of the drain voltage of stress. The peak of the hot-carrier damage always occurs at the maximum of the impact-ionized hot hole substrate current.⁹ Although it has been assumed that interface states were responsible for the damage, the first definitive proof that the damage occurs at the maximum of the substrate current came using the charge-

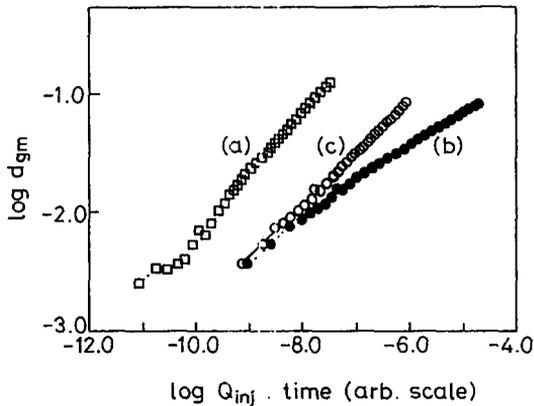


Figure 6.4 Transconductance change as a function of time of stress for 2 μm n-MOS transistors stressed at (a) $V_d = V_{d/2}$, (b) $V_g = V_d$, and (c) replotting the time dependence of degradation of curve b) as a function of injected gate current instead of time causes the gradient to change from 0.3 to 0.5 for stresses at $V_g = V_d$. (After Doyle et al., Ref. 35, © 1997 IEEE.)

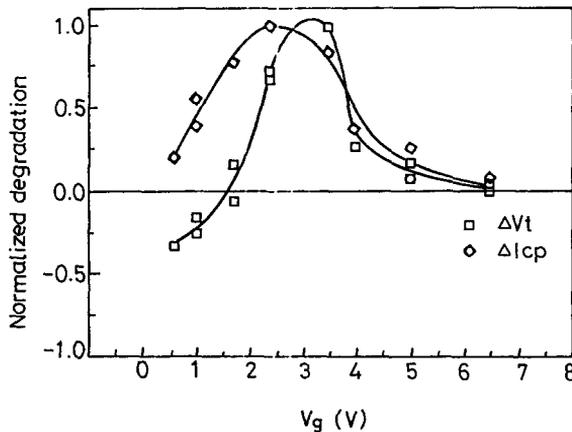


Figure 6.5 Dependence of degradation on gate voltage for transistors stressed at different gate voltages but the same drain voltage. Also plotted is the dependence of the substrate current under the same drain voltage conditions. (After Heremans et al., Ref. 15, © 1978 IEEE.)

pumping technique first reported by Jespers,¹³ and later developed by Groeseneken et al.¹⁴ This method allows the direct measurement of interface states, and measures the charging and discharging of interface states as the gate is pulsed (“pumped”) from accumulation to inversion and back. The methodology is discussed in Section 6.5. Figure 6.5 also shows the change in charge pumping current after stress, which is directly proportional to the interface-state density, as a function of gate voltage of stress. It can be seen that the correlation between interface-state density

and the degradation in device characteristics is excellent, confirming that interface states are indeed responsible for the damage at maximum substrate current conditions.

The reason for the $I_{b(\max)}$ peak is that these conditions correspond to high lateral fields (the lateral field is highest at the lowest gate voltage) as well as relatively large channel currents (the drain current increases with increasing gate voltage). However, an understanding of the reasons for the damage peak at $I_{b(\max)}$, and the mechanism of interface state creation comes from gate current measurements. Normally, the gate currents, especially the hole gate currents, are well below the 0.1–1-pA range of most ammeters. However, the floating-gate method⁴ allows the measurement of gate currents as low as 10^{-18} – 10^{-17} A (see Section 6.5). Figure 6.2 shows the gate current as a function of gate voltage. It can be seen that the conditions of maximum damage, paradoxically, occur when the overall gate current is very small, that is, when the gate electron and hole currents are equal and cancel each other. Although the gate current is low, the flux of electrons and holes is not zero. A model that explains such results is that the interface-state creation in MOS devices actually requires the injection of both electrons and holes into the oxide.^{16,17} The injected holes are first trapped in the oxide within 1.5–2.0 nm of the interface. The trap is subsequently neutralized by an injected electron and, by some as yet unknown mechanism, the hole trap is “transferred” to the interface, where it is transformed into an interface state. Hydrogen also appears to play a role in the interface-state creation. Measurements made after stress (see Section 6.7) link the presence of hydrogen to interface-state creation,¹⁸ as do radiation damage experiments on MOS transistors.¹⁹

6.3.2 Low Gate Voltage Stressing ($V_g = V_d/5$)

Hole Traps

The previous section has shown that the creation of interface states is intimately linked to the gate voltage conditions, the maximum damage arising under conditions of maximum substrate current. Examination of the gate currents during stress (Fig. 6.2) shows that at low gate voltages the preponderant gate current species is holes. Work on MOS capacitors under high gate voltage conditions¹⁷ has shown the presence of hole traps in the oxide. It is reasonable to assume that some hole trapping must be occurring also during hot-hole injection. That hot hole injection does indeed occur can be seen by examining the I_d – V_g data before and after hot hole injection. Figure 6.6 shows the characteristics of a conventional 1.5- μm effective-length transistor, before (a) and after stressing (b) at $V_d = 8.5$ V, $V_g = 1.5$ V (i.e., just above threshold). It can be seen that the effect of the stress is to cause the transistor I – V characteristics to improve. This increase in transconductance was first shown by Weber et al.²⁰ and Bellens et al.²² and identified as being due to localized hole traps ($N_{ot,h}$ —oxide traps created by holes) located at the drain junction edge of the transistor. Why this should happen has been shown by simulation to be due to localized hole trapping at the edge of the drain junction.²² Localized trapping causes a channel shortening effect—the local threshold of the transistor under the positive

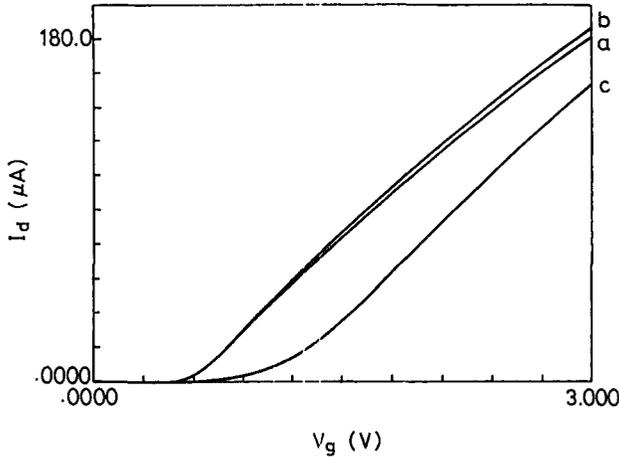


Figure 6.6 I_d - V_g characteristics of a transistor: (a) before stress; (b) after stress, at $V_g = V_d/4$; (c) following a short (6 s) electron injection pulse ($V_g = V_d = 7.0$ V). (After Doyle et al., Ref. 21, © 1990 IEEE.)

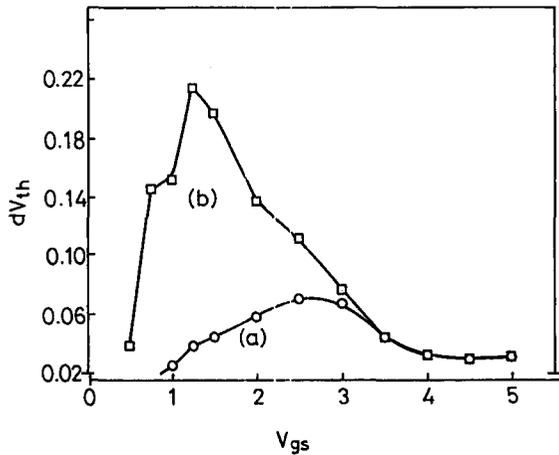


Figure 6.7 Dependence of degradation of gate voltage for transistors stressed at different gate voltages but the same drain voltage. The degradation was measured (a) before and (b) after a 6-second electron-injection pulse. (After Doyle et al., Ref. 21, © 1990 IEEE.)

charge has a greatly reduced threshold voltage. This part of the transistor consequently switches on before the rest of the channel. At voltages corresponding to the V_t of the undamaged part of the transistor, there is already considerable channel charge situated below the damaged region. The transistor thus appears, at threshold, as a device with a slightly shorter effective length. This results in an increase in the transconductance of the device (see Fig. 6.7b).

Electron Traps

It has been shown above that hot-hole injection causes hole trapping in the oxide, a not unexpected result. We shall now examine electron trap creation by hot holes, an unexpected, non-obvious consequence of hot-hole injection.

Returning to Figure 6.6 for a device stressed at $V_d = 8.5$ V, $V_g = 1.5$ V, curve (a) represents the unstressed measurements, and (b) represents the device after 10,000 s of stressing, which results in the trapping of a positive charge in the oxide. Now, if immediately after this stress, the device were to be further subjected to a short electron-injection stress phase for 6 s at $V_g = V_d = 8.5$ V (to inject negative charge into the oxide⁴), the $I_d - V_g$ curves would shift considerably (by 0.5 V!) to higher gate voltages (Fig. 6.6c).

Since changes of this nature do not arise from interface states (their occupancy depends simply on the Fermi level in the silicon, as mentioned earlier), the shift from curve (a) to curve (c) in Figure 6.6 can only be explained in terms of oxide traps. The injection of electronic charge following the stressing of a transistor under hot-hole injection conditions causes an extremely large degradation in the device characteristics resulting from a change in the occupancy of the traps in the oxide.

There are two interpretations of the oxide trap occupancy. The first is that there are only hole traps and interface states created under these conditions; the switching between Figure 6.6, curves (b) and (c) are due solely to neutralization of the positive hole traps, which were masking the effect of the interface states. The second is that the shift in the I–V characteristics is a result of emptying of the hole traps and the filling of what are called neutral electron traps.²¹

It has been shown that the second interpretation is true. Arguments are somewhat complex, and are elaborated elsewhere.^{24,25} Recently, further confirmation that neutral electron traps are indeed created has been made using impact-ionized substrate current and multiplication factor experiments.²⁶ Indeed, neutral electron traps have also been detected using other detection methods, which are discussed in Section 6.5. Similar neutral electron-trap generation has been seen in MOS capacitors subjected to high-field Fowler–Nordheim stressing,²⁷ where it was proposed that the electron trap is caused by hole trapping in the oxide and that a SiO₂ bond is broken in the process, leaving a trivalent silicon atom as the hole trap site and the nonbridging oxygen atom as the neutral electron trap. The trapped holes, thus, create equal quantities of electron and hole traps. A similar mechanism might take place under the low-voltage stressing conditions in these transistors.

There are, then, three types of damage occurring at low gate voltages (around $V_g = V_d/5$), interface states (in small quantities), oxide hole traps and oxide electron traps. When the damaged device is subjected to electron injection, the positive hole traps are neutralized, and the electron traps are now filled. The observed degradation (in Figure 6.6c) is due to both electron traps and interface states. If a series of transistors are stressed as a function of gate voltage (for a given drain voltage), and an electron pulse is then applied to them (Figure 6.7b), the degradation peaks at low gate voltages, voltages corresponding to the hole gate-current peak (see Figure 6.2). This strongly links these types of damage to hot-hole injection into the oxide. Figure 6.7a also shows the magnitude of the damage before electron injection. At this

particular drain voltage, the electron trap damage far outweighs the damage peak for interface state creation at $V_g = V_d/2$.

The hole and electron traps were further investigated to obtain their capture cross section,²¹ and values of capture cross sections $\sigma_1 = 3 \times 10^{-15} \text{ cm}^{-2}$ and $\sigma_2 = 3 \times 10^{-16} \text{ cm}^{-2}$ were obtained. The $3 \times 10^{-15} \text{ cm}^{-2}$ was assumed to be the hole trap, and the $3 \times 10^{-16} \text{ cm}^{-2}$ cross section corresponds to the charging of the neutral trap.²¹

Relaxable (Slow) States N_{not}

There is another type of damage that, although it does not share all the characteristics of electron and hole traps, is closely linked. This type of damage can best be manifested under the following conditions. If a transistor is stressed under hot-hole conditions (around $V_g = V_d/5$), and is then subjected to an electron-injection pulse (V_d high, $V_g = V_d$, to charge the hot-hole-generated electron traps), the threshold voltage of the device slowly decreases with time following electron injection.²⁹ This decrease can be monitored by measuring the drain current at gate voltages just above threshold, and $V_d = 0.1 \text{ V}$, as can be seen in Figure 6.8. The increase in current is the result of a gradual decrease in the post-stress threshold voltage as a function of time, due to the discharging of electron traps in the oxide²⁹ into the substrate.

To trace the origins of the effect, the size of the transients were measured as a function of gate voltage after a fixed period of relaxation.²⁹ The maximum transient was obtained at very low gate voltages, the approximate conditions for the maximum hole gate current, indicating that hot holes are responsible of this type of damage, as they were for the creation neutral electron traps.

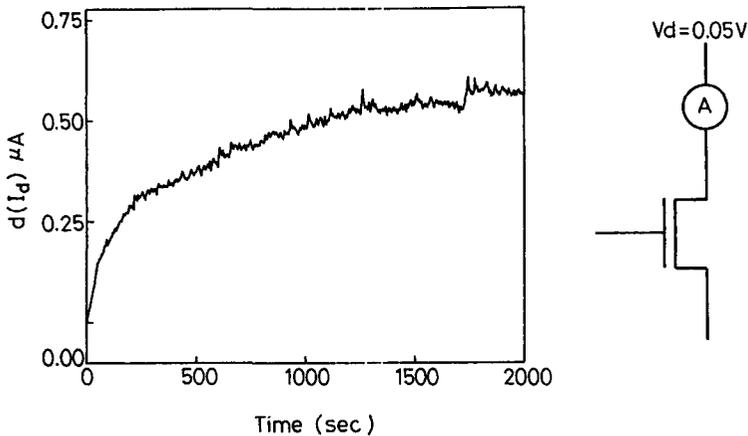


Figure 6.8 Typical curves of drain current as a function of time measured at a fixed gate voltage and $V_d = 0.05 \text{ V}$ following a hole injection stress, and an electron injection trap-filling pulse. The measurement conditions are shown in the schematic. (After Bourcierie et al., Ref. 29, © 1990 IEEE.)

The fact that the transients are not thermally activated led to the conclusion that this behavior was in fact due to the quantum mechanical tunneling of a charge from traps located close to the interface (illustrated in Fig. 6.3). What occurs here is that immediately after the electron-filling step, the traps located close enough to the Si–SiO₂ interface tunnel out into the substrate, resulting in a gradually decreasing negative oxide charge and, consequently, a gradually decreasing threshold voltage. It is this decreasing threshold voltage that causes the transients seen in Figure 6.8. Analysis of the transient shows that it does not consist of a single time constant, nor a continuum of time constants, but rather several distinct values well separated in time. This suggests that the traps are situated at discrete distances into the oxide.

Before leaving this type of damage, several further points should be noted:

- *Capture Cross Section.* The measurements of these near–interface oxide traps gave values of approximately 10^{-15} cm^{-2} , which corresponds to values between hole and electron traps. The experimental error in each capture cross-sectional measurement case was large.
- *Slow Hole Traps.* Were found also to exist, and had the same V_g dependence of creation as the electron N_{tot} .
- *Time Constants.* Figure 6.8 shows time constants of the order of 10 s and longer. This interval was used because these traps are far easier to characterize. The tunneling-out time from these states are long enough to easily study the charging/or discharging and so on. It should be noted that there is a considerable body of work that exists on these near-interface oxide traps; many of them have considerably shorter time constants, up to the limit where a relaxable trap is indistinguishable from an interface state.^{30–33} Much of these have been detected through either $1/f$ noise measurements³¹ or random telegraph signals.³⁰ The charge pumping method can also detect these states, when the device is pumped at low frequencies.³²
- *Localization.* Paulsen et al.³² used an expression for trap-to-trap tunneling to calculate the distance into the oxide from the interface for traps with different time constants.³² They showed that the traps are situated at certain distances from the interface, from approximately 1.0 nm for the short-time-constant traps (10^{-6} s), to between 2.0 and 3.0 nm for the traps²⁹ with time constants longer than 10 s.

6.3.3 High Gate Voltage Stressing ($V_g = V_d$): Electron Trapping

Given the dependence of gate current on gate voltage in Figure 6.2, it might be expected that the high gate currents passing through the oxide at simultaneously high gate and drain voltages might have some deleterious effect on the oxide. If the injection distance from the drain junction is considered to be $0.1 \mu\text{m}$ or less, the current densities in this injection area can be as high as 10^{-3} A/cm^2 . Oxide trap damage under such conditions has been reported.^{34,35} In some systems, there is seen not only the peak at $I_{b(\text{max})}$ conditions, as shown in Figure 6.5, but also significant

damage at higher gate voltages. This damage does not correlate with charge-pumping current. Analysis of the time dependence shows that the damage at these high gate voltages of stress does follow a time power law (as in Eq. 6.3), but with a reduced gradient of 0.2–0.3 (Fig. 6.4*b*). The absence of a charge pumping signal following stresses indicates that the damage is due to electron trapping in the oxide ($N_{ot,e}$ -oxide traps created by electrons). That this trapping occurs in the presence of significant gate electronic currents further indicates that the damage is created by injected electrons.

An explanation for the lower gradient obtained for these types of stress as compared to those at $I_{b(max)}$ conditions has been suggested.³⁵ As electrons are injected into the oxide, the trapped charge in the oxide builds up a repulsive potential field that causes a decrease in the amount of charge injected into the oxide (this is discussed further in the next section on p-channel stressing). Therefore, the amount of injected charge per unit time decreases as the stress time increases. However, plotting degradation as a function of injected charge instead of time (Fig. 6.4*c*) shows that the ubiquitous 0.5 gradient to the power law.

Relaxable Damage

Following arguments developed for hot hole damage, it would be expected that the oxide traps are located at all distances into the oxide from the interface and, thus, show relaxable or slow state properties. The presence of slow electron states, has proved much more difficult to detect than would be thought. The presence of slow states has been seen in this gate voltage range. However, it is in the reoxidized nitrated oxides, which show much larger electron trapping, that evidence of slow states are seen unambiguously.³⁶ The maximum amount of relaxable damage was found to peak at the same gate voltage as the gate current, as would be expected for hot-electron-generated relaxable traps.

6.3.4 Gate Voltage Dependence of Stress: p-MOS

In the previous sections, the types of damage found in n-MOS devices have been described. Four types of damage have been found: interface states, hole traps, and two types of electron traps (not counting slow states). These types of damage are intimately linked to gate current during hot-carrier stressing. In this section, we shall show that similar damage exists for the p-channel transistors. They show the same types of damage, for the same species of charge injected. The gate currents are reversed for p-MOS compared with n-MOS, and consequently the gate voltage range for each type of damage is reversed—linking the types of damage created in p- and n-MOS stressing to the gate current species independent of whether the damage is in n-channel or p-channel transistors.

Figure 6.9 shows the gate current–gate voltage spectrum. At low gate voltages, the current is electrons and large—several orders of magnitude larger than in the n-channel transistors. The reason for this is that the secondary impact-ionized electrons have a much lower barrier (3.1 eV) for injection into the oxide than do holes in n-MOS devices (4.8 eV). The current is, consequently, much larger, even if

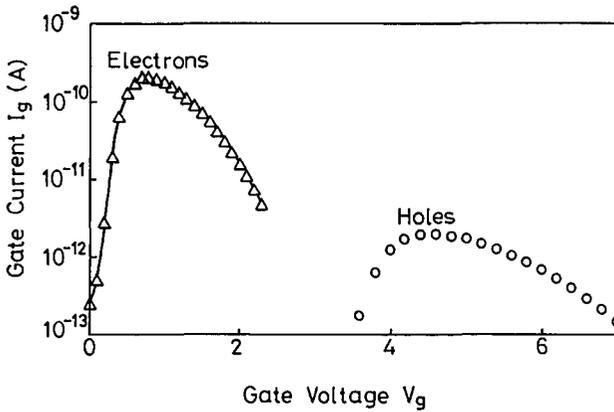


Figure 6.9 Log $I_g - V_g$ characteristics of a p-MOS transistor measured at high drain voltage, showing the large gate electronic current at low V_g voltages. The high gate voltage hole current is a representation of the current at this gate voltage range.

the secondary impact-ionized current is lower, due to the shorter mean free path of the holes. At medium gate voltage, the flux is both electrons and holes, and the gate electron current drops to zero and changes sign. At high gate currents, hot holes can be injected into the oxide. However, the conditions for hot-hole detection are difficult as these conditions also correspond to Fowler–Nordheim injection between the source and the gate.

Low Gate Voltage Range

Electron Traps At low gate voltages, under conditions of gate-electron injection, the damage suffered from hot-electron injection manifests itself as an increase in the transconductance of the transistor (Fig. 6.10a)^{37,38}. This increase in transconductance is associated with trapping of the injected electrons in the oxide. The trapped negative charge leads to a local decrease in the threshold voltage and an earlier turnon of the channel in hot-carrier-damaged region.^{37,38} The device appears at threshold as a transistor with a shorter effective length, as in Figure 6.6b. This results in an increase in the transconductance of the device, and the size of the damaged region determines the amount of the current gain of the transistor after stress.

Figure 6.11 shows the transconductance change as a function of time on a log–log scale, while Figure 6.12a shows the dependence of lifetime on gate voltage of stress. Most damage (shortest lifetimes) occurs at low gate voltages, indicating that the gate electron current is responsible for the damage (see the gate current dependence on gate voltage in Fig. 6.9). Returning to Figure 6.11, the devices in this figure were stressed under maximum gate current conditions (V_g just above threshold voltage) for different drain voltages. It can be seen that, unlike the case of the hot-carrier stress of n-MOS transistors, the transconductance (g_m) change for the p-MOS transistors does not follow a time power law, but has a tendency to saturate at

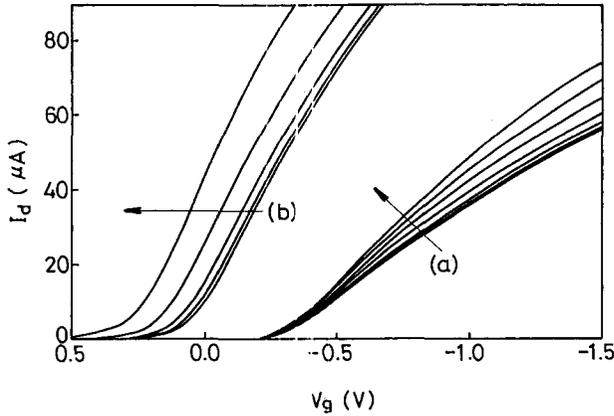


Figure 6.10 $I_d - V_g$ characteristics of a p-channel transistors stressed at low gate voltages: (a) longer channel ($> 0.3 \mu\text{m}$); (b) short channel ($< 0.3 \mu\text{m}$) for various times of stress. The arrows indicate the shifts in the characteristics after stress.

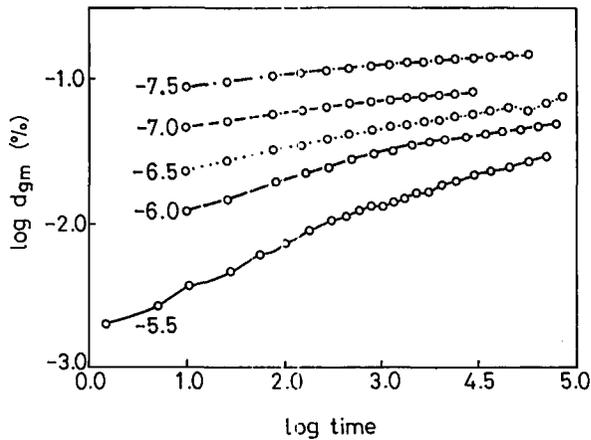


Figure 6.11 Degradation (g_m) percentage as a function of log time for a series of devices stressed at low gate voltages. In contrast to Figure 6.4, the data not show power-law behavior, but do reveal a tendency to saturate at long times.

long times and high drain voltages.^{37,38} There are several reasons for this apparent saturation.

Once the trapped charge in the localized damage region has sufficiently shifted the local threshold, the charge above the drain junction no longer significantly affects the characteristics, the channel shortening effect stops, and the device current saturates. However, the device current does not saturate completely. Brox et al.³⁹ showed that if the damage were to be plotted on a linear-logarithmic scale, the resultant plot would show a straight-line behavior. This straight-line behavior was

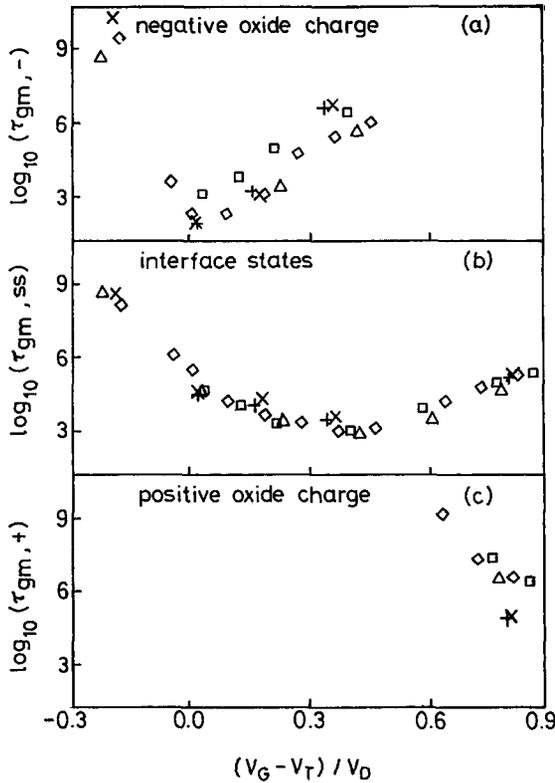


Figure 6.12 Lifetime curves for a series of transistors stressed as a function of voltage: (a) negative oxide traps, (b) interface states, and (c) positive oxide charge at high gate voltage damage, as detected by charge pumping current. The charge pumping current peaks at intermediate-gate voltage in a manner similar to that for interface state damage in n-MOS transistors (Figure 6.2). Note that shorter lifetimes indicate longer hot carrier damage. (After Woltjer et al., Ref. 43, © 1994 IEEE.)

explained in terms of the growth of the damage region towards the source as a function of the stress time. Working from the model shows that the trapping kinetics follows a first-order rate equation. The authors showed that there is a trapping front established and that the edge of the trapping front moves toward the source junction linearly as the stress time progresses on a log scale. Brox et al. modeled this system, and found very good agreement with experimental results.³⁹

Another contributing factor to the saturation in the plots of Figure 6.11 could arise from the gate current behavior as a function of time. The gate current does not remain constant as a function of time, but gradually decreases⁴⁰—at least at high drain voltages and/or long times. This decrease is a consequence of the trapping of electrons in the oxide, which alters the vertical and horizontal fields and causes fewer electrons to be injected over the Si-SiO₂ barrier as a function of time (i.e., as the amount of trapping increases).

It might be assumed that the damage species arising from electron injection in the oxide for both n-MOS and p-MOS transistors should be identical. While it is not possible to prove conclusively that this is so, measurements of the capture cross section made on p-MOS transistors give a value in the mid- 10^{-18} cm $^{-2}$ range, a value within the experimental error of the values found in electron-generated electron traps in n-MOS transistors.

Relaxable (Slow) States N_{nit} Given the correlation between hot-electron-created oxide electron traps in n-channel transistors, and their counterparts in the p-channel transistors, these traps would be expected to show some evidence of electron detrapping also. Brox et al. and others have shown that these states do, indeed, detrapp.^{39,41} The detrapping follows a logarithmic time dependence and that the stronger the negative field on the gate during the detrapping phase (the drain, source, and substrate being grounded), the more rapidly the damage detraps.

Intermediate Gate Voltage Stressing ($V_g = V_d/2$)

Interface States Initially it was thought that electron trapping was the only (or overwhelming) damage mechanism. However, the symmetry in the gate oxide injection species (see Figs. 6.2 and 6.9) suggest that the damage in p-channel transistors should be reflected in the behavior of the gate current, that is, electron trapping at low gate voltages in p-MOS due to the large electron gate current, and interface states and oxide hole traps at high gate voltages, where gate hole injection occurs. The first indications that mechanisms other than electron trapping might be occurring was shown by Tsuchiya et al.⁴² in shorter-channel transistors (0.25 μm) at high gate voltages, where he showed that there was a contribution to hot-carrier damage from interface states. This damage occurs at high gate voltages and peaks at voltages corresponding approximately to $V_g = V_d/2$, as can be seen in Fig. 6.12, curve *b*).^{43,44}

High Gate Voltage Stressing ($V_g = V_d$)

Hole Traps Positive oxide charge is also trapped in the oxide when stressing at high gate voltage^{43,44} through the use both of the charge-pumping technique and by using an electron injection phase mentioned in the n-MOS low gate voltage stressing. Figure 6.12*c* shows that this occurs at high gate voltages of stress—at large gate hole current conditions.

Relaxable Damage As with the electron traps at low gate voltage, it would be expected that the hole trapping should detrapp, depending on their position from the Si interface and their position in the oxide forbidden gap. There appears to be no literature alluding to hole detrapping after stressing at $V_g = V_d$. Woltjer et al. did examine this, and were unable to detect any detrapping of holes.⁴¹ Given the correlation between the damage in n- and p-MOS seen so far, the absence of hole detrapping is anomalous and curious.

Negative-Bias Temperature Instability (NBTI)

The final type of damage that p-MOS transistors suffer is called *negative bias temperature instability* (NBTI). This damage can be seen on both n-MOS and p-MOS devices, but is far more prevalent in p-MOS transistors, and is usually tested only on these devices. It has been shown that NBTI does have hot-carrier repercussions. Consequently, this damage mechanism is discussed here. First we give a brief introduction to NBTI followed by a discussion of the hot-carrier effect on NBTI.

NBTI is a phenomenon that occurs at high temperatures, usually over 100°C, under the influence of high negative gate fields (see Ref. 45 and references cited therein). The effect manifests itself in an increase in the n-channel threshold and a decrease in the subthreshold gradient, indicating N_{it} and hole N_{ot} creation. Many different mechanisms are thought to be responsible, including oxygen vacancy formation, strained bond reaction, hole or electron tunneling in the oxide, and electrochemical reactions.⁴⁶

Although this is a general reliability concern, the interest here is what occurs when simultaneous high drain and high gate voltage stresses are applied to the transistor. This effect is called *hot carrier NBTI*, or HC-NBTI.⁴⁵ The effect of applying voltages to the gate and drain is to accelerate the phenomenon by a full order of magnitude. This degradation *increases* with increasing temperature in contrast to the other mechanisms that have been discussed. Although this effect is seldom considered because simultaneous high gate and high drain voltages do not occur for very long in a clock cycle, it is important for nitrided gate oxides discussed in Section 6.7, since these oxides show increased susceptibility to NBTI and HC-NBTI.

Note that this HC-NBTI effect is not the summation of hot-carrier damage and NBTI damage (as has been suggested⁴⁶), since the drain voltages can be so low as to not suffer HC damage (at room temperature), and still show HC-NBTI.

6.3.5 Conclusion

Summing up this subsection on the different mechanisms that occur in HC stressing of n- and p-MOS transistors, Table 6.1 shows the various types of damage arising during stress. Interface states, hole traps, several types of electron traps, and their slow trap counterparts are seen to be created in both transistor types. Each type of damage is closely linked to the gate current species that is injected into the oxide. It can be seen that most types of damage that occur in n-MOS also occur in p-MOS. The conditions for damage creation are very similar, the gate current species are also identical, and the capture cross-section values are also similar enough to suggest that the same type of damage is being created in n-MOS as in p-MOS transistors.

There are, however, some exceptions. For instance, hole-generated electron traps and their slow-state counterparts have not, as yet, been detected in p-MOS—the injection conditions give rise to ambiguities in the results. Furthermore, not as much research has been carried out on the p-MOS transistor and it is possible that further work will reveal that these are, indeed, created. In addition, HC-NBTI damage in n-MOS has not been detected.

TABLE 6.1 Different Types of Damage Created in Oxide and Characteristics of that Damage

Damage Species		Gate Voltage Range	Damage Species	Capture Cross Section (cm ²)
Interface states	n-MOS	Intermediate: gate voltages	Combined holes + electrons	1×10^{-14}
	p-MOS	Intermediate: gate voltages	Combined holes + electrons	1×10^{-14}
Hole traps	n-MOS	Low gate voltages	Holes	1×10^{-14}
	p-MOS	High gate voltages	Holes	1×10^{-14}
Slow hole traps (slow states)	n-MOS	Low gate voltages	Holes	?
	p-MOS	Not observed	Not observed	?
Electron traps (hot electron created)	n-MOS	High gate voltages	Electrons	1×10^{-18}
	p-MOS	Low gate voltages	Electrons	1×10^{-18}
Slow electron traps	n-MOS	High gate voltages	Electrons	1×10^{-18}
	p-MOS	Low gate voltages	Electrons	1×10^{-18}
Electron traps (hot hole created)	n-MOS	Low gate voltages	Holes	?
	p-MOS	Not observed	Not observed	?
Slow electron traps (hot hole created)	n-MOS	High gate voltages	Electrons	?
	p-MOS	Not observed	Not observed	?

The next section shows that each type of mechanism is necessary for a full description of hot-carrier damage during circuit functioning and that, considering the quasistatic contributions of each of the damage components, the ac lifetimes can be accurately predicted. This is certainly true of n-channel transistors. For the p-channel counterpart, research has not yet established that this is so.

6.4 HOT-CARRIER LIFETIME ESTIMATION: AC AND DC

6.4.1 Introduction

The field of n-channel transistor hot-carrier ac lifetime is one that has generated a considerable difference of opinion. Early reports showed that gate voltage transients (with constant drain voltage) lead to increased degradation, whereas drain voltage

transients (with constant gate voltage) do not.⁴⁷ It was further claimed that the falling, not the rising, gate voltage edge, is responsible for increased degradation and that faster fall times lead to larger degradation.^{47,48} Higher frequencies, accordingly, also led to faster degradation.⁴⁸ This degradation was also linked to a transient effect: enhanced substrate currents during falling gate voltage edges.^{47,48} It was suggested that the enhancement in substrate current was larger at shorter transitions times; the phenomenological link between substrate current and hot-carrier degradation then explained the enhanced degradation effect.

However, several reports failed to confirm any substrate current enhancement, at least for rise/and fall times in the range 1 ms to 3 ns,⁴⁹ and the earlier substrate current enhancement was then linked to measurement difficulties.^{50,51} With the measurement difficulties removed, no difference was found in the degradation rate between the falling and rising gate voltage edge. Thus, the transient effects that had been postulated as an explanation for enhanced hot-carrier degradation was discarded. The most accurate ac model involves the inclusion of the different gate voltage degradation mechanisms discussed above.^{50,54} However, in order to discuss this, the dc lifetime extraction techniques need to be examined first, because these dc lifetime techniques will form the basis of the ac lifetime extraction method.

6.4.2 n-MOS Static Hot-Carrier Damage Modes

Interface States

If a series of devices are stressed at different drain and gate voltages (for gate voltages around $V_d/2$), a series of straight lines are obtained if the device degradation is plotted as a function of time of stress on a log–log plot, as has been seen in Section 6.3 (Fig. 6.4a). Early on, Hu et al.³ found that if the substrate and drain currents during stress are monitored and the time for damage to reach a certain critical level (normally taken as a 10% degradation in transconductance) is noted, plotting these two variables on a log–log plot gives the following relation (Fig. 6.13a):

$$\tau = K \cdot \frac{(I_b/I_d)^{-m}}{I_d} \quad (6.4)$$

where τ is the hot-carrier lifetime, I_b is the substrate current, I_d is the drain current, m is an empirical constant approximately equal to 3.0, and K is an empirical constant. Equation 6.4 was found to hold irrespective of drain voltage. From plots such as these, we can extrapolate to working circuit voltages by simply noting the maximum substrate current at the working drain voltage and extrapolating to the point on the graph of Figure 6.13a.

Hot-Hole-Generated Electron Traps (Low V_g)

In the low gate voltage range, where the gate current is hole-dominated, three damage species are found; trapped holes, neutral electron traps, and interface states. Equation 6.4 still holds true, except that interface states are not the predominant

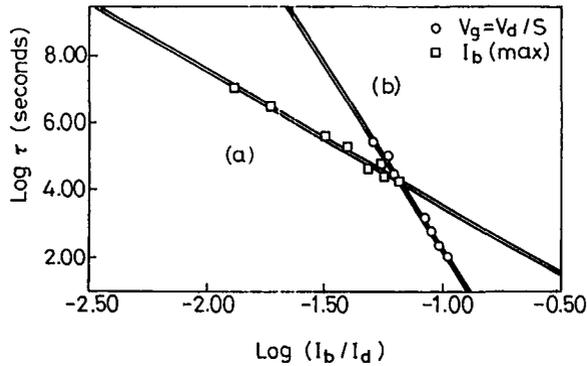


Figure 6.13 Lifetime of a series of device stressed at (a) maximum substrate voltage conditions and (b) $V_g = V_d/5$, followed by an electron injection pulse. The results are plotted as the log of the lifetime as a function of the log of substrate/drain currents. (After Doyle et al., Ref. 21, © 1990 IEEE.)

damage species at high drain and low gate voltages, as has been discussed previously. However, most of the low V_{gs} damage is not readily apparent from MOSFET I_d - V_g curves immediately after stress, as has been discussed earlier, and as was shown in Figure 6.6. This is because the electron traps are neutral and the trapped holes mask the effect of the interface states. The positive charge cancels the negative interface-state charge and leads to negligible g_m or V_t shifts. When stress at low V_{gs} is followed by a brief electron-injection phase (stress at a high V_{gs} condition of electron injection into the oxide), the neutral electron traps are filled and the trapped holes are neutralized, effectively exposing the damage electrostatically.²¹ It has been shown that low V_{gs} damage can occur for gate voltages below the threshold voltage of the MOSFET, and that under certain conditions, damage for $V_{gs} < V_t$ can even predominate.⁵²

If a series of devices are stressed at high drain and low gate voltages, followed after each stress by a short electron-injection phase (to fill the electron traps), estimates for damage at low gate voltage can be made from plots like that of Figure 6.4. Plotting the lifetime for each stress as a function of drain and substrate currents according to the Eq. 6.4 gives a straight-line behavior (Fig. 6.13b), but here m has a value between 5 and 15,^{21,22} as opposed to the value of 3.1 obtained for interface state generation.

Hot-Electron-Generated Electron Traps (High V_g)

Under conditions of high gate bias, where the gate current is predominantly an electron current, significant damage can occur in the gate bias region in the form of electron traps. The substrate current under the stress conditions $V_{gs} = V_{ds}$ is about an order of magnitude smaller than the peak I_b . Were it not for electron trap damage, one would expect from Eq. 6.4 that the lifetime at $V_{gs} = V_{ds}$ would be three orders of magnitude smaller than that at peak I_b conditions. $N_{ot,e}$ damage can be

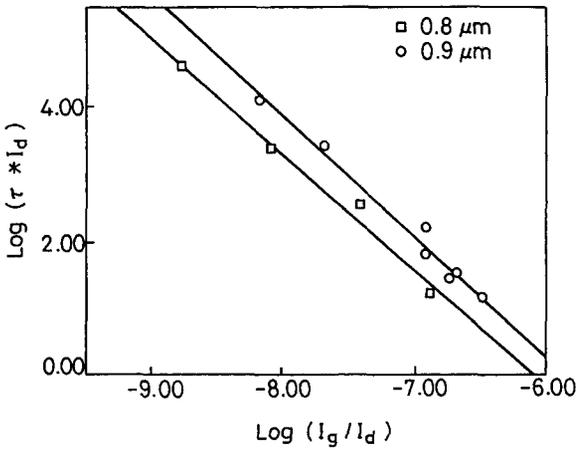


Figure 6.14 Lifetime of a series of devices stressed at different drain voltages at gate voltage conditions corresponding to maximum gate current. The results are plotted as the log of the lifetime as a function of the log of the gate current over the drain current. (After Doyle and Mistry, Ref. 53, © 1991 IEEE.)

modeled by⁵³

$$\tau = C \cdot \frac{(I_g/I_d)^{-p}}{I_d} \quad (6.5)$$

where C is an empirical constant, with p taking values close to 1.5. Note that this equation is valid only for V_{gs} values such that the gate current is negative (i.e., consists primarily of electrons). As an approximation, it is valid for $V_{gs} > V_{ds}/2$. Figure 6.14 shows the plot of log lifetime versus log I_g/I_d . It can be seen that a straight line is obtained with a gradient of -1.5 , which is consistent with Eq. 6.5.

In summary, Figure 6.15 shows the different types of damage occurring in the gate voltage spectrum. At low gate voltages (even below V_t), electron and interface states are created and hole trapping occurs; in the intermediate gate voltage region, interface states are the primary damage mode; and at high gate voltages, electron traps and a small amount of interface states are created. All of these damage modes occur during ac stress, as is illustrated by the waveform of Figure 6.15, and all three modes are necessary to explain the enhanced degradation during dynamic stressing of n-channel MOS devices, as is shown in the following sections.

6.4.3 n-MOS AC Stress Lifetimes

Given that there are three distinct damage regions occurring in the gate voltage spectrum, it would be expected that ac stress of transistors should encompass all three damage mechanisms. It has been shown that ac stress can indeed be explained by taking into account the three different mechanisms discussed above.^{50,54} The

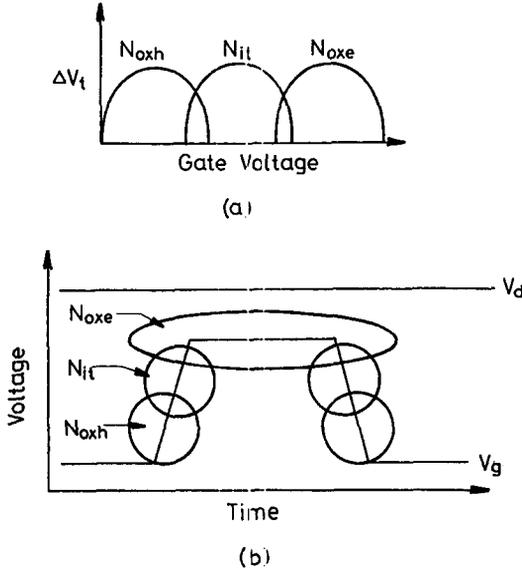


Figure 6.15 (a) Representation of the different types of damage occurring in the gate voltage spectrum during dc hot-carrier stress at different gate voltages. (b) Regions during a gate voltage pulse in which the three damage modes are seen during ac stressing for 50% duty-cycle stresses. Inset. All three types of damage occur during ac stress.

approach is to examine the waveform of the transistor under ac operation and, knowing the instantaneous values of I_d , I_b and I_g for the stress, calculating the contributions of the three damage mechanisms during ac by integrating Eqs. 6.4–6.6:

$$\frac{1}{t_{N_{es}}} = \frac{1}{A} \int_C^{i_r} \left(\frac{I_b}{I_d}\right)^m \cdot dt \tag{6.6}$$

$$\frac{1}{t_{N_{ot,h}}} = \frac{1}{B} \int_C^{i_r} \left(\frac{I_b}{I_d}\right)^n \cdot I_d \cdot dt \tag{6.7}$$

$$\frac{1}{t_{N_{ot,e}}} = \frac{1}{C} \int_C^{i_r} \left(\frac{I_g}{I_d}\right)^l \cdot \frac{I_g}{I_d} \cdot dt \tag{6.8}$$

In these integrals, we treat $1/t$ as a damage function, which is integrated over the time period T of the stress waveform.

To compare the quasistatic contributions of the three damage modes to the stress results, we still need some method to combine the quasistatic damage contributions. It has been proposed that the following Matthiessenlike rule be used for that purpose.⁵⁰

$$\frac{1}{t_{ac}} = \frac{1}{t_{N_{it}}} + \frac{1}{t_{N_{ot,h}}} + \frac{1}{t_{N_{ot,e}}} \tag{6.9}$$

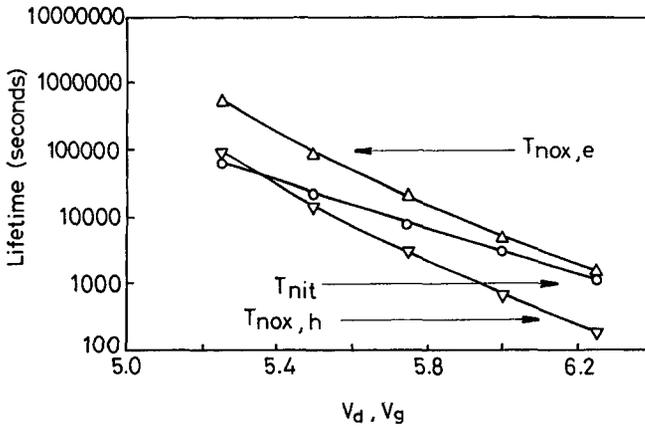


Figure 6.16 Quasistatic contributions from N_{oxh} (inverted triangles), N_{it} (circles) and N_{oxe} (triangles) to the ac lifetime for 50% duty cycle stresses. (After Mistry and Doyle, Ref. 55, © 1993 IEEE.)

where the terms on the right-hand side have been calculated above. Equation 6.9 also is equivalent to viewing $1/t$ as a damage function. The damage functions for the three damage modes are added together to calculate the total damage.

Using a waveform such as that shown in Figure 6.15, where the regions in which the three different damage modes contribute to the lifetime are shown schematically the a.c. lifetime of a device can be obtained using Eq 6.9.

The quasistatic components for each damage mechanism for the waveform of Figure 6.15 is shown in Figure 6.16. It can be seen that at highest voltages, the hot-hole-generated electron trap damage is the most important mechanism, leading to the shortest lifetimes, while at lower drain and gate voltages, the damage caused by interface state creation starts to become the dominant mechanism governing stressing at this drain voltage.

The results shown in Figure 6.17 were obtained by calculating the lifetime based on Eqs. 6.6–6.8, and applying the Matthiessenlike rule of Eq. 6.9 for the waveform of Figure 6.15. The results show that this formulation leads to an extremely good match between the measured ac data and the modeling of the ac data based on the dc data and the stress waveform. Thus, the quasistatic contributions of $N_{ot,h}$, N_{it} , and $N_{ot,e}$, taken together, fully account for the so-called enhanced ac hot-carrier degradation for this type of C stress waveform.

To conclude this section, the ac behavior of a transistor in a circuit is completely described by taking the quasistatic contributions of the three damage mechanisms and integrating them over the waveform applied to the transistor.

6.4.4 p-MOS Static Hot-Carrier Damage Modes

Low Gate Voltage

Figure 6.11 shows the results of stresses performed on p-MOS transistors, for variable V_d and for V_g corresponding to maximum gate currents. As discussed

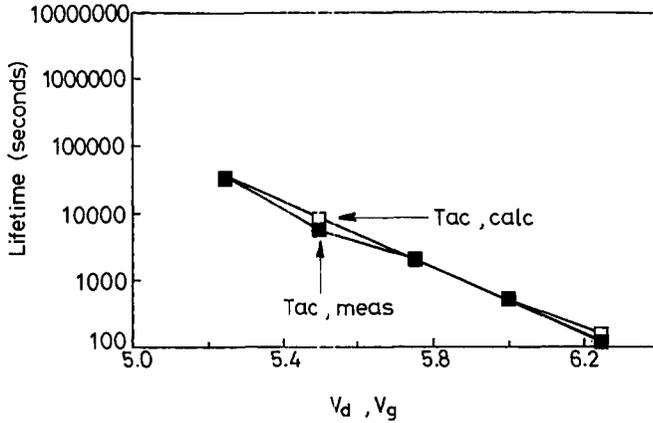


Figure 6.17 AC stress lifetimes compared to the combined quasistatic contributions of N_{oxh} , N_{it} , and N_{oxc} from Figure 6.16. (After Mistry and Doyle, Ref. 55, © 1993 IEEE.)

earlier, the time evolution here does not show a time power-law behavior (as do n-MOS devices; see Fig. 6.4a), but does show variable gradients.

However, several methods can be used to estimate hot-carrier lifetimes:

- The first is to plot the data as dg_m versus log time, as has been discussed in Section 6.3. The different curves are then “normalized,”³⁸ which involves the lateral displacement of the $\log I_{d(sat)}$ -log time curves for different drain voltages such that a single curve is obtained. Figure 6.18a illustrates this effect. Here, the different stress curves have been shifted in time with respect to the lowest drain stress voltage data, to form a single normalized curve. From this curve, the lifetime for the unshifted stress voltage curve can be read directly off as the time for a 10% change in this figure.
- The second approach is to take the stress data and plot them on a linear-log scale. The straight-line behavior allows for the extrapolation of the data to a certain critical level of stress. The problem with this approach is that significant amounts of damage are required to establish the gradient at low drain voltages, as each stress drain voltage has its own gradient.

The third method uses ΔL_{eff} , the change in the effective length of the transistor following stress, instead of Δg_m . Woltjer et al.^{56,57} established that the drain resistance was directly proportional to the L_{eff} change. Plotting the L_{eff} damage represented by $(\Delta L_{eff}/T_{ox})^{0.5}$ as a function of the log of time gave a series of straight lines for different drain voltages. The relationship is

$$\Delta N_{ot,0} = \left(\frac{\Delta L_{eff}}{T_{ox}} \right)^{0.5} = 0.33 \log_{10}(1 + Kt) \quad (6.10)$$

where K depends on the technology and the degradation conditions. From this, it is possible to establish the characteristic gradient for the technology at

high drain voltages. Then, with a measure of the damage at the drain voltage of interest, the data can be extrapolated to critical damage levels, and the lifetime read directly from the figure.

From the quasistatic point of view, a model is needed that would allow the prediction of the lifetime based on the MOSFET output currents, as was done in the case of the n-MOS device (Eqs 6.4 and 6.5). This point is discussed later in the section on p-MOS ac lifetime.

Intermediate Gate Voltages

In the case of damage at higher V_g values, where interface states are created, the damage follows the relationship⁵⁷

$$\frac{\Delta L_{\text{eff}}}{T_{\text{ox}}} = L t^{0.45} \quad (6.11)$$

where L is degradation-dependent parameter. Figure 6.12b shows that the peak of the damage occurs in the intermediate gate voltage range. This suggests that impact-ionized substrate current would be a good indicator to use in a lifetime-prediction model having the same form as Eq. 6.4. Thus far, no published data exist for this type of damage in p-channel devices. It should be noted that whereas the first damage mechanism causes an *increase* in drive current (a shortening of L_{eff}), the intermediate gate voltage damage (as well as hole trapping discussed below) causes a *decrease* in drive current. Consequently, there exists a range in which these two mechanisms compete. This gives rise to the crossover of the transconductance characteristics seen in going from low gate voltage stresses (g_m increases) to higher gate voltage stresses (g_m decreases).⁵⁶

High Gate Voltages

Hole trapping occurs at high gate voltages. Woltjer et al have also developed a model to track the time dependencies of trapping.⁵⁶ They use an expression similar to that for negative-charge trapping:

$$\Delta N_{\text{ot},0} = \left(\frac{\Delta L_{\text{eff}}}{T_{\text{ox}}} \right)^{0.5} = G \cdot \{\log_{10}(1 + K t)\} \quad (6.12)$$

where the symbols have their usual meanings.

AC Model With respect to circuit lifetime extraction, very little has been published about any ac model. This is because the n-MOS, with its decrease in drive current with time has worried circuit designers more than p-MOS, with its increase in I_d . There is currently no ac model that would allow for the prediction of the p-MOS lifetime, given the ac waveform of the device in the circuit under examination.

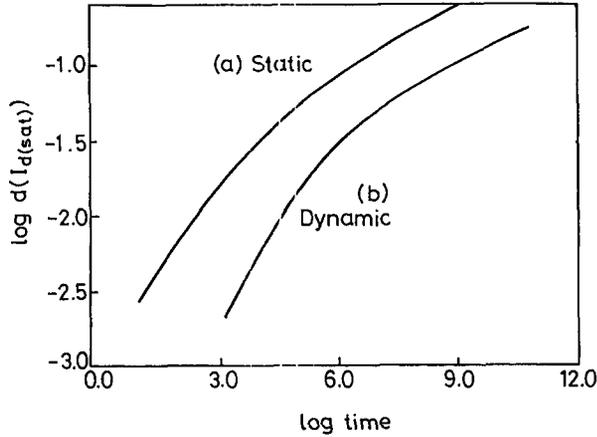


Figure 6.18 (a) Normalization of the dc stress data of results such as in Figure 6.11 using the data shifting technique discussed in Ref. 38. The data have been normalized with respect to the lowest stress voltage. (b) A similar shift technique applied to ac stress data.

The most advanced work in this domain involves only the modeling of the dc behavior of Figure 6.18 from Eqs 6.10--6.12.⁵⁷

However, work on ac stressing of p-MOS transistors under inverter waveform shows transconductance changes that lend themselves to the normalization method of Ref. 37. Figure 6.18b shows the results of a series of devices stressed at different drain and gate voltages under inverter stressing conditions⁵⁸ and normalized according to Ref. 38. This results in a smooth curve very similar to that obtained for dc stresses. This method allows quantifying the hot-carrier lifetime for ac stressing, but does not allow a predictive method, and does not include the higher gate voltage contributions (indicating that, in this case, they did not play a dominant role).

In conclusion, although the dc lifetimes of p-MOS transistors have been characterized, an ac model has not yet been published for the p-channel transistor. This is becoming important because, as transistor lengths are reduced to the $\leq 0.1 \mu\text{m}$ gate length regime and as the oxides are thinned to the point where the electron and hole traps are too close to the interface to remain trapped, interface state creation will become the predominant damage mechanism in p-channel transistors. Consequently, a model that includes all stress components, including interface state damage and hole trapping, is necessary for ac lifetime prediction.

6.5 HOT-CARRIER MEASUREMENTS

The problem of the identification of the types of damage and the localization of the damage (for both process amelioration and modeling the device degradation) has led to a plethora of different characterization techniques that allow for the identification, quantification, and localization of the damage.

6.5.1 Charge-Pumping Technique

Interface State Detection

The most widely used technique and the richest in terms of information that it can provide is the *charge pumping technique*. This technique was proposed initially in 1969;¹³ However, it came of age as a technique in 1984, when Groeseneken et al. explored it fully.¹⁴ The technique involves the application of pulses to the gate of a transistor (with source and drain at a fixed potential, usually 0 V; see Figure 6.19a). These pulses cause the interface states to charge and discharge during the pulses, and the measured substrate current measures interface states directly. To understand the technique, a quick explanation is given here. A fuller explanation can be found in several references.^{14,59,60}

Looking at the band diagram, the states that are available in terms of charge pumping current are those that are not in equilibrium with the charge pumping voltage pulses. This is shown in Figure 6.19b. The states in equilibrium (hatched

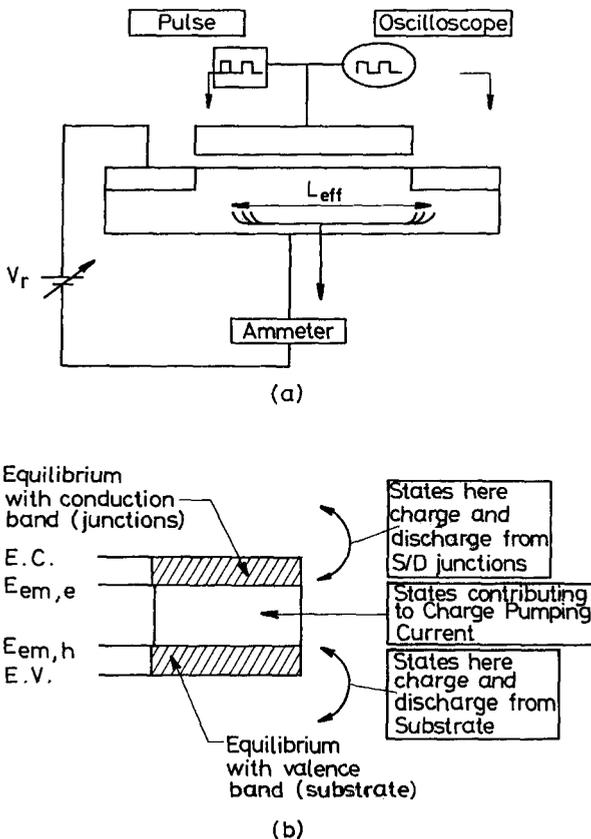


Figure 6.19 Schematics for the (a) charge pumping technique and (b) energy gap, and the states that contribute to the charge pumping current.

area in Figure 6.19*b*) fill with electrons during CP gate pulsing from the conduction band (the upper states in diagram). During a full cycle, the net contribution to the drain current from these states is thus zero. Similarly, the states close to accumulation (lower states in the diagram) both fill and empty into the substrate (by holes), and the net substrate current is also zero.

For states between the two levels, the time constants for emission and capture of a charge are longer than the voltage transients, and these states are in non equilibrium during the transitions. They consequently change occupancy only when large numbers of holes (as in accumulation), or large numbers of electrons (in inversion) are available to fill or empty the interface state by recombination. There is a net current of electrons from the source drain regions to the substrate, or hole current in the opposite direction, and the charge transfer is proportional to the number of interface states available to be pumped per cycle. The expression for the charge pumping current is given by¹⁴

$$I_{cp} = 2 q f A_g k T D_{it} \ln(v_{th} n_i (\sigma_n \sigma_p)^{1/2} (t_{em,e} t_{em,h})^{1/2}) \quad (6.13)$$

where q is the charge of the electron, f is the frequency of the charge pumping signal, A_g is the area of the gate being pumped, k is the Boltzmann constant, T is temperature, v_{th} is the thermal velocity of the electrons, n_i is the intrinsic concentration, σ_n and σ_p are the capture cross sections for electrons and holes respectively, and $t_{em,e}$ ($t_{em,h}$) are the times available for emission (capture) of electrons during the gate cycle.

Typical curves obtained using the charge pumping method are shown in Figure 6.20, along with the gate voltage pulses corresponding to each region on the charge pumping curve. The charge pumping curve is obtained by keeping the amplitude of the voltage pulse constant and varying only the base voltage. At point *a*, the top of the pulse is not yet in inversion, and no current is seen. At point *b*, the top of pulse is in inversion, and the current increases dramatically. As the base voltage increases, the transistor continues to be pulsed between inversion and accumulation. However, at point *d*, the base level moves above accumulation, and there is no longer the supply of holes to discharge the interface states charged in inversion, and consequently the charge pumping substrate current decays to zero.

Oxide Trap Detection

Figure 6.20 shows that the charge pumping technique can be extended to detect oxide traps. If stressing transistors creates interface states and oxide-trapped charge, the local threshold of the part of the transistor adjacent to the oxide traps will effectively be changed (along with the local flat-band voltage). Figure 6.21 shows the effect of trapped charge. In Figure 6.21*b*, the CP signal from the created interface states is moved to lower base voltages as a result of the local shift in the threshold and flat-band voltages. The effect is that the CP curve shows a shoulder at low base voltages. Similarly, the presence of negative trapped charge shifts the threshold of the stressed part of the transistor to higher voltages and shifts the point at which the

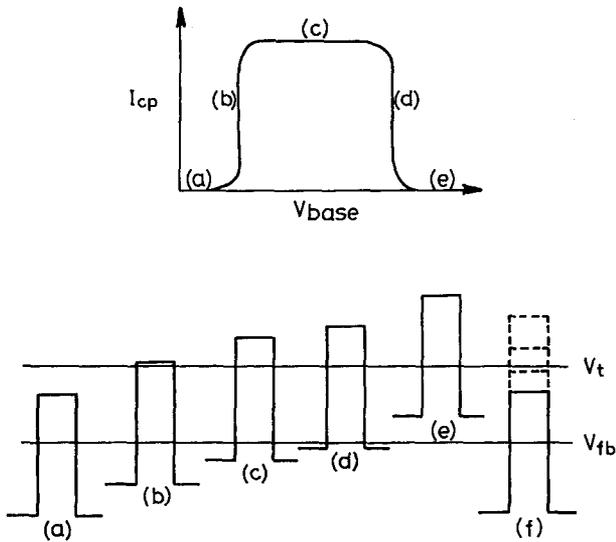


Figure 6.20 Typical curves obtained from the charge pumping technique. The gate pulse conditions leading to the various sections of the charge pumping curve are shown in the lower diagram. Point *f* refers to the charge pumping technique where the base level is left at a constant value, and the amplitude of the pulse is varied.

CP signal is seen from the stress damage to higher base voltages. The net effect is the presence of a shoulder on the I_{cp} curves at high base voltages, as is illustrated in Figure 6.21c.

The illustrations of the effect of damage on the charge pumping characteristics have been seen for both positive-charge trapping and interface states in n-MOS transistors⁵⁹. The behavior in the case of negative-charge trapping has also been seen for p-MOS transistors.⁵⁹ However, in the case of electron damage in n-MOS devices, we do not see this simple picture of oxide trapped charge shown in Figure 6.21c. The absence of a striking, easily defined charge pumping signal such as that shown in Figure 6.21c led to some controversy regarding the presence of hot-hole-created electron traps.^{24,25} However, careful analysis of the charge pumping current levels in conjunction with the drain current degradation showed conclusively that such states exist.²⁵ These states have since been confirmed using other techniques.⁶¹ In fact, the charge-pumping technique does indeed detect hot-hole-generated electron traps in n-MOS transistors. This is discussed below.

The method of extracting the amount of oxide damage from charge pumping curves was proposed by Vuillaume et al.⁶² They showed that if the charge-pumping current at the turnon point was examined with high sensitivity, there is a shift in the negative tail of the CP current curve (Fig. 6.22). The amount of the shift in the tail is directly proportional to the amount of oxide charge. However, no change in the high gate voltage portion in the characteristics are observed.

Another method of obtaining information on positive oxide traps from charge-pumping is a method in which the base level of the CP pulse is set at a constant value

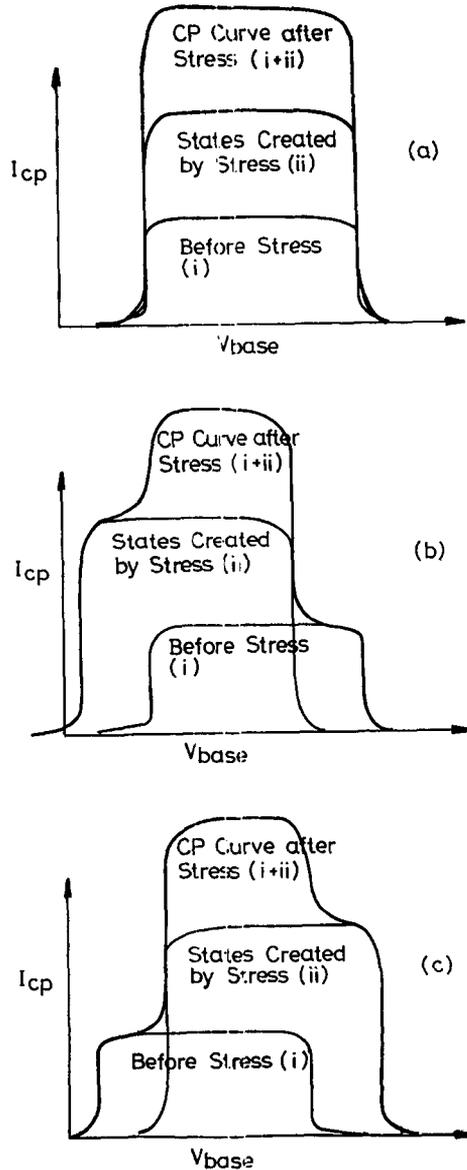


Figure 6.21 Representation of the charge pumping current as a function of base voltage level, with and without gate oxide charge: (a) interface states alone; (b) interface states in the presence of hole traps; (c) interface in the presence of electron traps.

and the amplitude of the pulse is varied (see Fig. 6.20f). Taking the derivative of I_{cp} with respect to the high level of the pulse, Chen and Ma⁶³ showed that the peak in the signal versus voltage shifted in voltage, indicating oxide trapping adjacent to the created interface states.

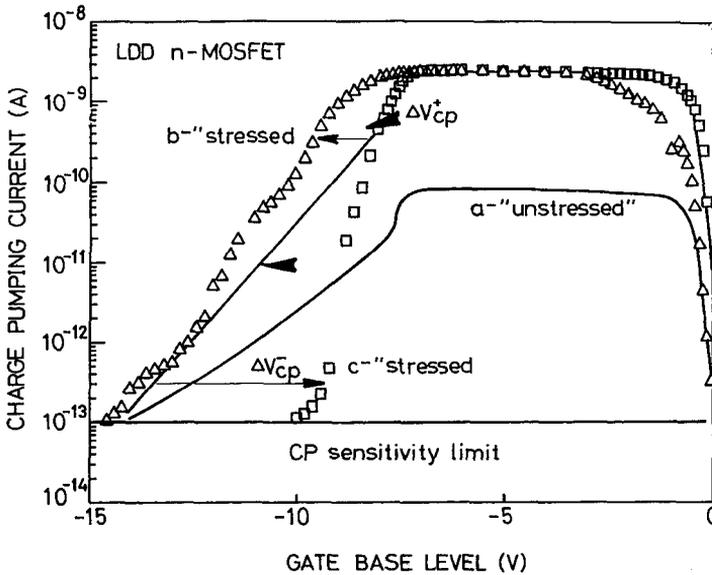


Figure 6.22 CP characteristics (a) of an unstressed device, (b) following hot-hole stressing, and (c) following a short electron-injection pulse to fill the neutral electron traps. The effect of the pulse is to move the tail of the CP current to higher base voltages (ΔV_{cp}^+). (After Vuillaume et al., Ref. 62, © 1993 IEEE.)

Interface-state Capture Cross Section

The charge pumping technique can also be used to give information about capture cross sections.⁶⁴ Saks and Ancona used three voltage levels for the charge pumping pulse (V_b , V_i , V_h). The intermediate level, V_i , allowed the traps between V_j and V_i to come to equilibrium and emit their charge back to the conduction band. This differs from the results obtained from the two-level (V_b , V_h) technique in which the trap captures a hole from the valence band and produces a charge pumping current.

By varying the height of V_i and comparing I_{cp} signals, a slice of the energy bandgap can be examined, and the states in this gap characterized and quantified. Saks and Ancona took this one step further. If the time t_i that the pulse remains at the level V_i is varied, the point at which the states go from equilibrium to non-equilibrium (t_e) can be found. Since this point is related to the capture cross section σ through the expression

$$t_e = \frac{1}{\sigma_e v_{th} n_i} e^{-\phi_i/kT} \quad (6.14)$$

They found values of approx 10^{-16} for σ_h and 10^{-13} – 10^{-15} cm^{-2} for σ_e .

Localization of Damage

Localization of the damage has also proved possible to detect using the CP technique. Heremans et al.⁵⁹ showed that it is possible to detect the presence of

interface states and oxide traps by modulating the junction depletion layer. This they did by applying a voltage to the substrate to change the depletion region around the junction edges (the body effect). With a negative voltage applied to the substrate in Figure 6.19, the junction depletion region spreads out from the junction edge, resulting in a space charge region, which shields the interface states in this region. Following stress, the interface states and trapped oxide charge can be characterized with respect to the proximity of the junction edge using this technique.

The weakness with this technique is that charge trapping in this area during hot-carrier stressing changes the local flat-band voltage and threshold voltage, and that these effects are not taken into account. However, using the substrate bias technique⁶⁵ and 2D simulation, we can compensate for the local variation in threshold and flat-band voltages. A similar technique has also been proposed by Li et al.⁶⁶ A similar type of CP profiling has also been used to look at damage under the oxide spacer of LDD structures.⁶⁶ In this particular approach, the progressive decrease of V_{base} allows for the profiling into the accumulated region under the spacer.

On a final note with respect to the modeling of pure charge pumping approach, although Cilingiroglu⁶⁷ succeeded early on in using Shockley–Reed–Hall (SRH) recombination theory to fairly successfully model the charge pumping approach, the turnon and turnoff points were not well modeled. In 1994 a more successful approach has been developed using SRH theory,⁶⁸ that shows very good agreement with experiment.

6.5.2 Floating-Gate Technique

The floating-gate method also allows for separation of interface states from trapped oxide charge.⁶⁹ The floating-gate technique⁴ consists of applying a voltage to the isolated gate of a transistor, floating the gate node, applying a drain voltage such that charge is injected into the oxide, and monitoring the gate voltage as the injected hot carriers discharge the gate node. Because the drain current is a function of gate voltage, the gate current that discharges the floating gate node can be calculated. In the case of a stressed device, the damage causes a change in the injection of hot carriers into the oxide, and it is this change in the oxide current that is used as a damage monitor. This technique is a particularly sensitive monitor since the point of injection of the gate current is also the point at which the damage occurs.

Figure 6.23*b* shows the I_g-V_g behavior of a device before and after stress in which hot-hole-generated electron traps are present. The effect of the damage is to shift the whole gamut of the I_g-V_g characteristics. This shift occurs as a result of the repulsive field in the oxide caused by the trapped charge. Figure 6.23*a* shows the behavior following interface state creation.⁷⁰ The decrease in gate current can be seen to occur mostly at high gate voltages, and has a signature very distinct from that seen for oxide trap damage, in the low gate voltage region, allowing the identification of oxide trap and interface-state damage.⁷⁰ The reason for the difference is that the quasi-Fermi level varies with gate bias. The quasi-Fermi level

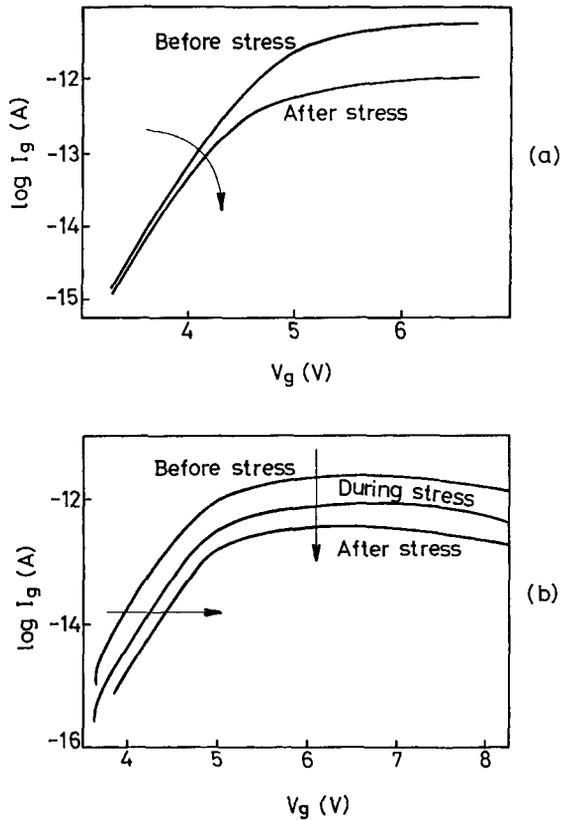


Figure 6.23 I_g - V_g behaviors for an n-MOS: (a) stressed to create interface states ($V_g = V_d/2$); (b) stressed to create hot-hole-generated electron traps ($V_g = V_d/5 +$ short electron injection phase at $V_g = V_d$). (After Doyle et al., Ref. 70, © 1993 IEEE.)

at low gate biases lies below the energy level of the (acceptor) interface states. These created states are thus not charged, and consequently do not influence the injection of electrons into the oxide at low gate biases.

6.5.3 Gated-Diode Technique

Another interesting technique called the gated-diode method⁷¹ involves biasing a MOS transistor gate into accumulation. Under these conditions, the quasi-Fermi levels for holes and electron is swept toward (and into) the junction since the gate is biased further into accumulation. Interface states existing in this region act as generation-recombination centers giving rise to a measurable current whose magnitude depends on the quantity of interface states located around midgap. The proximity of the conduction and valence bands in strong accumulation (high negative gate biases for n-MOS devices) causes the conduction band and valence

band of the Si to approach one another, causing tunneling between valence and conduction bands via the interface state.

Two other approaches that allow for the separation of interface states and oxide traps will briefly be mentioned.

6.5.4 Gate Overlap Capacitance

The gate overlap capacitance technique also offers a method of measuring and differentiating between interface states and oxide charge.⁷² This method involves applying an ac signal to the drain of a transistor, and measuring the gate-to-drain capacitance as a function of gate voltage bias. In sweeping the gate bias, the capacitance goes from the overlap plus the outer fringing capacitance at $V_g < V_t$, to $C = 0.5 C_{\text{gate}}$ plus the overlap and the outer fringing capacitance.

The extension of the junction edge varies according to whether positive or negative charge is trapped in this region, and this can be obtained from C_{gd} .

- Hole trapping decreases V_t , and results in a larger overlap capacitance. Trapping negative charge depletes the edge of the junction, resulting in a decrease of the amount of channel that is effectively in inversion.
- Interface states also have a unique signature, stretching out the transition from accumulation to inversion, in a manner analogous to the stretchout effect due to interface states seen in MOS capacitors.

6.5.5 DCIV Method

The final method of separating oxide traps from interface states is the c current–voltage (DCIV) method developed by Neugroschel and co-workers.⁷³ By using the well contact of the MOS (substrate contact) as the base with the drain as the collector and source as the emitter, they showed that the two species can be identified as follows: The base current of the vertical BJT is used to measure recombination due to the interface states at the Si–SiO₂ interface, and the collector current of the vertical BJT is used to measure the oxide trap concentration, since the collector current increases when the gate voltage passes from flat band toward depletion and inversion.

6.6 STRUCTURE DEPENDENCE

This section discusses the effects on the hot carriers of what will be called the structural components of the transistor, in contrast to the next section where the effects of processing are discussed. This involves the various components related to the scaling of the devices, such as gate oxide thickness, channel length, and the use of LDDs. The section finishes with a short discussion of mechanical stress on the hot-carrier properties of transistors.

6.6.1 Length

There are several components to the effect of length on the hot-carrier effect, each of which plays a role in the sensitivity of the transistor being stressed. The most obvious is that the lateral field is intimately linked to length—keeping the drain voltage constant increases the lateral field, increasing the damage for a given stress time. Thus, the shorter the effective length of the transistor, the greater the damage.⁷⁵ This has been discussed previously.

The second effect arises due to the length of the damage region. It has been shown that the length of the damage region, l , is independent of the effective length of the transistor. Thus, even if transistors of different lengths were stressed under the same field conditions—with drain and gate voltages corresponding to the same lateral field—decreased lifetime for the shorter channel length devices suffering the same amount of damage⁷⁵ would result.

Third, transistors that are stressed in the pass-transistor mode (drain and source are interchanged during the stress) suffer damage at both ends of the channel. In n-channel transistors, the decrease in lifetime can be as much as an order of magnitude shorter in the transistor stressed at both source and drain ends than in a device stressed only at the drain end.⁷⁶ The case of the p-MOS transistor stressed under electron-injection conditions is more marked. As the transistor effective length decreases, the damage size becomes a bigger and bigger fraction of the transistor length. Stressing at the source and drain ends under hot-electron injection accentuates the short-channel effect, resulting in a greatly decreased lifetime for devices suffering damage at both source and drain edges. The lifetimes of such transistors are reduced by at least two orders of magnitude.³⁸ Furthermore, when the damage length l approaches half the length of the transistor, the trapping in the transistors stressed at both junctions can no longer be considered localized, and the I_d-V_g characteristics shift from the behavior seen in Figure 6.10a to that of Figure 6.10b. In this case, the damage is no longer dominated by transconductance shifts, but by threshold voltage changes, which become the predominant degradation effect.

Finally, as gate lengths are reduced to the <100-nm range, velocity overshoot becomes increasingly important. Carrier overshoot arises when the high-field region of the channel approaches the mean-free-path length of the electrons and holes (5–10 nm).⁷⁷ At these dimensions, the carriers pass through the high-field region without suffering a collision and are consequently not injected into the oxide (nonstationary transport regime). It would be expected that transistors in this regime would suffer less hot-carrier damage. It has been seen that very short-length transistors show less impact-ionized substrate current,⁷⁸ but definitive data showing decreased hot-carrier damage have yet to be published.

6.6.2 Drain Engineering

LDD Engineering

The LDD has been used for some time to alleviate the effects of hot carriers. The LDD region itself can be thought of as a resistive region between the channel and the

junctions that drops some of the lateral field across this region. Consequently the field in this region is not as great as in the conventional transistor and the lifetimes are longer.⁷⁹

Two points need to be added here: (1) for performance reasons (drive currents), the LDD region doping concentration in scaled MOSFETs is becoming comparable to that of the source/drain junction, and it is becoming more difficult to alleviate hot-carrier effects; and (2) in p-MOS devices, the trapping in the LDD region can result in a larger change in effective length for devices stressed under electron trapping conditions, and the hot-carrier properties can sometimes be worse with an LDD than without one.⁸⁰ Careful engineering is thus required to optimize this region from the hot-carrier standpoint.

Elevated Source–Drains

Elevated source–drains have also been suggested as a potential method of decreasing hot-carrier susceptibility. Tasch et al. suggested that by building in an n-field reduction region through selective silicon epitaxy of the junction regions, the hot-carrier effect could be alleviated.⁸¹ Other similar structures have also been built.⁸² However, as noted in the LDD discussion above, any “field-reducing” region introduced into the transistor acts to lower the drive current of the device. Consequently, careful engineering is required in this junction edge region.

6.6.3 Oxide Thickness

One of the few scaling tendencies that reduces the hot-carrier problem is oxide scaling. The effect of decreasing oxide thickness is to decrease a transistor's sensitivity to the hot-carrier effect.^{83–85} Aur et al.⁸³ measured about a factor of about 3–4 increase in hardness for n-MOS transistors stressed at $I_{b(\max)}$ conditions when the oxide is reduced from 6.5 to 4.5 nm, even though the thinner oxides have higher lateral fields (the substrate doping was scaled for constant V_t). They attribute the increased hot-carrier hardness to the increased inversion-layer charge at a given gate voltage, since the inversion-layer charge scales as $Q_{\text{inv}} = C_{\text{ox}}(V_g - V_t)$. Thus, the inversion-layer charge tends to screen the interface states more effectively with thinner oxide. This result was supported by 2D simulations.

In p-MOS transistors, measurements of buried channel devices show that this effect also occurs here,⁸⁴ with a factor of almost 5 decrease in hot-carrier damage in going from a 20-nm oxide to a 7-nm oxide. More strikingly, in the case of surface channel transistors, the lifetimes have been found to decrease by up to four orders of magnitude, as measured by threshold voltage shifts (Fig. 6.24) for 0.25- μm L_{eff} devices stressed in forward and reverse modes (worst-case conditions) under electron trapping conditions.⁸⁵ Analysis of the damage pointed to localization as being responsible for this increased hot-carrier hardness, gate-to-channel capacitance measurements confirming that the trapping becomes increasingly localized over the drain junction when the oxides become thinner.

The preceding section discusses oxides that have thicknesses down to 4.5 nm (6.5 nm for p-MOS). Tsuchiya et al.⁴² have measured p-channel transistors whose

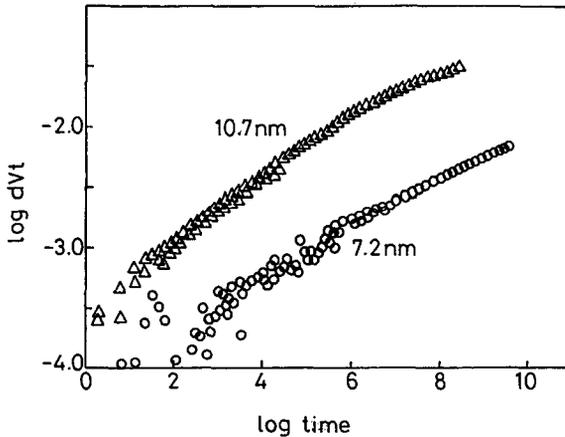


Figure 6.24 ΔV_t versus time for p-channel MOSFET's, for 10.7 and 7.2-nm oxides, stressed in both forward and reverse directions. The thinner oxides show a four orders-of-magnitude decrease in lifetime. (After Doyle et al., Ref. 85, © 1995 IEEE.)

gate oxide thicknesses were 3.5 nm, and have shown that the nature of the hot-carrier damage changes at these oxide thicknesses, going from predominantly electron trapping to interface-state creation. This can be understood in terms of the tunneling distance of the charge from the oxide. Cohen et al.³³ have shown that near-interface traps begin to behave like interface states when these states are situated between 1.2 and 1.8 nm from either interface. Thus, for oxide thicknesses of 4.0 nm and thinner, the problem of oxide trapping diminishes and the interface state creation mechanism becomes the predominant factor, for both n- and p-channel transistors. Finally, it has been shown that oxides at the tunneling limit are indeed more hot-carrier-resistant, although these oxides were not fully analyzed.⁸⁶

6.6.4 Mechanical Stress

Another structural parameter that plays an important role in hot-carrier susceptibility of a MOS structure is mechanical stress. Hamada et al.⁸⁷ showed, by using a 4-point bending jig, that compressive stress on both n- and p-MOS resulted in larger threshold voltage shifts than in mechanically unstressed devices. It was originally thought that this was due to enhanced trapping/interface-state generation, but measurements performed by Degraeve et al.⁸⁸ showed that the multiplication factor (I_b/I_d) was strongly influenced by mechanical stress, with tensile (compressive) stress giving increases (decreases) in multiplication factor in n-channel transistors. In this case, tensile stress resulted in more degradation than compressive stress, although in all cases, the increased susceptibility is not great. They also showed, using *substrate hot-electron injection* experiments, that the trap creation rate itself was unchanged.

Finally, the effects of stress are also seen from a variety of sources, such as gate electrodes (W in this case⁸⁹), passivation processing,⁹⁰ and transistor isolation (in

SOI devices).⁹¹ These sources have to be considered in analyzing the susceptibility of a particular transistor structure to hot carriers.

6.7 PROCESS DEPENDENCE

This section deals mainly with three recent additions to the hot-carrier damage portfolio: plasma damage, the different techniques that result in the introduction of bonded nitrogen within the gate dielectric, and the effect of back-end processing, and passivation anneals.

6.7.1 Plasma Damage

Another aspect of reliability is plasma charging.⁹² Plasma charging results from the presence of a nonuniform plasma in the various dry etching and ashing steps during processing. Plasma damage has been reported for poly (silicon) etch,⁹³ metal etch,⁹⁴ and photoresist ashing.⁹²

The nonuniform plasma causes potential imbalances at the wafer surface and these imbalances neutralize themselves. The gate oxide offers such a path and for thin oxides (<200 Å), the local imbalance of the charge causes the gate electrode potential to build up to a point that Fowler–Nordheim current flows through the oxide, which lowers the gate potential. The passage of a charge through the oxide weakens the oxide and renders the devices more susceptible to failure. As the gate oxide thickness decreases, the field in the oxide for a given plasma voltage increases, as does the Fowler–Nordheim current. Consequently the situation worsens as the gate oxide thickness decreases.⁹⁶

The potential at the gate is determined by the degree to which the charge can be collected. The charge that gathers on the poly electrode is determined by the size of the metal antenna connecting the gate to the interconnect system. This charge collects on the antenna and is funneled down to the thin gate oxide, building up the potential on the gate. The size of the metal (or polysilicon on thick gate oxide) antenna determines the amount of charge that gathers at the gate oxide level. This effect is referred to as the *antenna effect*, and the antenna ratio is defined as the ratio of the poly gate area on thin gate oxide to the poly/metal area on thick oxide/interlevel dielectric (ILD). Figure 6.25 shows an example of the effect of antenna ratio on the charge to breakdown of an oxide.

That the plasma charging mechanism is caused by Fowler–Nordheim ($F-N$) injection has been shown by Shin and Hu.⁹² They showed through quasistatic $C-V$ that the oxide damage caused by the antenna effect was similar to that caused by an oxide that underwent constant current stress. This similarity allows for the quantification of the charge collected and short-circuited through the oxide during the plasma process.

However, it is not so much the antenna area as the periphery that is important in determining plasma damage.⁹⁷ The reason for this, and the understanding of plasma damage, can be described as follows. Take the example of the plasma etching of a

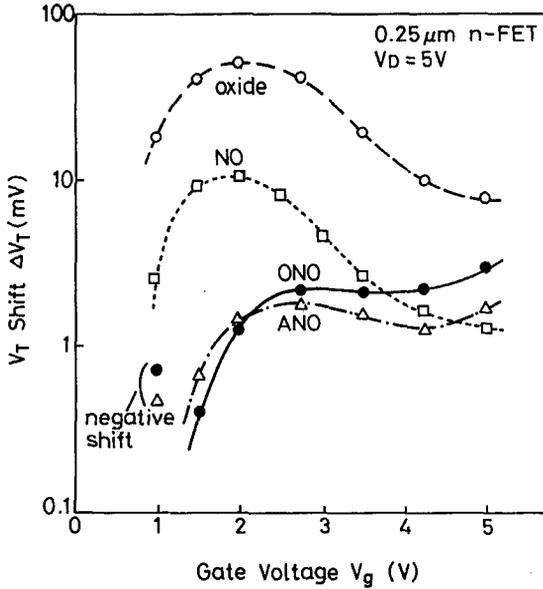


Figure 6.25 Hot-carrier degradation as a function of gate voltage. The different curves represent the different gate oxide splits. The nitrated oxides (NO) and reoxidized nitrated oxide (ONO) greatly reduce the peak of the hot-carrier damage. (After Hori et al., Ref. 120, © 1992 IEEE.)

metal pad. The pad, covered with resist, is in contact with the unprotected deposited metal until the exposed metal has been completely etched. No charge builds up on the pad until this point. Once the unprotected metal is completely removed, the charge begins to build up on the exposed metal sides at the edge of the resist mask. The longer the overetch, the greater the buildup of charge.⁹⁴

The damage itself takes the form of interface states,⁹² slow interface states,⁹⁸ and positive hole traps.⁹² This damage appears not to be annealed out—interface states reappear after an alloy anneal,⁹⁹ and the hole traps remain, even after anneals as high as 900°C.¹⁰⁰

Plasma damage has a significant effect on the hot-carrier hardness of the oxide.^{99,101} This added susceptibility comes from the prestressing of the gate oxide by Fowler–Nordheim injection that arises during the plasma processing. Initial measurements appeared to show that n-MOS transistors were affected by antenna ratios. However, later work⁹⁹ showed that the n-MOS was much less susceptible to interface state generation but that it had increased electron trapping (at $V_g = V_d$). More striking was the decrease in p-MOS hot-carrier susceptibility.⁹⁸ The p-channel device shows a very marked increase in damage under electron trapping conditions (V_g just above threshold).

In summary, the damage in both n- and p-MOS under hot-carrier stress for unprotected device with large antenna ratios takes the form of increased electron trapping susceptibility.

Even though the phenomenon is quite simple to perceive, plasma damage has subtleties that are not immediately obvious:

- *Electron Shading Effect.* In dense fingerlike structures, antenna structures suffer damage while similar sparse-line structures show no damage for the same etch conditions.⁹⁵ This damage arises due to the trajectory of the ions and electrons. Electron trajectory tends to be more scattered than ion trajectory, which have a more vertical incidence on the surface of the wafer. Consequently, just before the exposed metal at the bottom of the spacing between dense lines is cleared, the ions can impinge on the metal, whereas the more easily scattered electrons are blocked by the shading effect of the photoresist.
- *Resist Damage.* It might appear counterintuitive that resist might play a negative role in plasma damage. However, Chien et al. have shown that the presence of resist on large antenna ratios (AR) during resist etching can have a detrimental effect on the plasma damage compared to similar devices whose resist is wet-etched.¹⁰¹ They explain this in terms of plasma self-adjustment. The potential between the wafer surface and substrate is lowered by losing charge to the substrate through the substrate contacts. However, when resist is covering the metal, there is no path for the charge to the substrate, and this potential adjustment step is severely retarded, resulting in higher surface potentials at the surface. Using a simple equivalent capacitor structure model, they showed that the potential dropped across the gate oxide with an aspect ratio of 10,000 could be up to 100 times that of a capacitor with an aspect ratio of 1. Further indication that this occurs is found if the gate polysilicon is covered with a dielectric before gate definition.¹⁰² The deposited oxide acts as a capacitor in series with the photoresist, decreasing the capacitance of the photoresist stack.
- *In Situ Annealing.* Although plasma damage has been presented in terms of charging effects, there is also a significant amount of UV photons generated by the process.⁹³ UV photons above 9 eV can cause damage in the oxide.¹⁰³ Lai et al. showed that periphery antenna structures show less damage for multifingered structures, than for single fingered structures with the same AR, when exposed to a capacitively coupled O₂ plasma.⁹³ This is explained in terms of the different wavelengths of the vacuum UV photons given off during the plasma etch process. Above 9 eV, the photons cause damage. However, photons between 4.3 and 9 eV generate photoelectrons, which, in turn, neutralize trapped positive charge. Since it is only the photons at the polysilicon edges that are not absorbed by the polysilicon itself, the structures with the greater periphery are photoannealed preferentially. The differences with other results (e.g., Ref. 97) are explained in terms of difference in intensity and wavelength of the emitted photons from one plasma etch system to another.

Finally, before leaving the subject, two points should be made with respect to alleviating plasma damage. SOI transistors show virtually no plasma damage

compared to their bulk counterparts.¹⁰⁴ This is attributed to the fact that the devices themselves are isolated from the substrate by a buried oxide. It is, consequently, not possible for the gates to charge up to a different potential from the isolated source–drain areas. Thus, using an SOI technology should alleviate the plasma damage effect almost completely.

Second and more important, at some point in gate oxide scaling, the gate oxide begins to suffer leakage due to direct tunneling. When the gate oxide begins to leak sufficiently, it will no longer be completely capacitive, but will have a resistive component. Charge building up on the gate during plasma processing will start to leak away. At a certain oxide thickness, it will no longer be possible to maintain sufficient potential across the oxide to cause plasma damage and the plasma charging issue will be largely solved through gate oxide scaling. This point, at which the oxide begins to suffer direct tunneling, is around 3 nm. Somewhere below this point we can expect that the plasma damage will decrease, although the thickness at which the oxide begins to improve from the plasma damage will depend very much on the tunneling properties of the oxide. Oxides with less leakage current for a given oxide thickness will suffer more oxide damage.¹⁰⁵

6.7.2 Oxidation

The introduction of nitrogen bonded into the gate oxide dielectric has been one of the more fruitful areas of research into making hot-carrier hard oxides. The techniques were developed using what has been called reoxidized nitrided oxides (ROXNOs, or RNOs),^{106–111} but this work has spawned a full range of different process techniques which place nitrogen at the Si–SiO₂ interface. For the purposes of this discussion, the techniques have been broken down into the following:

- *Pre-oxidation.* This consists of implants of nitrogen into the silicon before gate oxidation. During oxidation, the nitrogen becomes incorporated into the oxide.¹¹²
- *During Oxidation.* These methods are based on the slight nitridation of the oxide (up to 10 at.% N incorporated into the oxide), by NH₃ after the oxidation,^{106–111} followed by a reoxidation step. Other well-researched techniques include either growing the oxide in an N₂O ambient^{111,113} or annealing after growth in N₂O. Similarly, NO anneals have been used,¹¹⁴ as have nitrogen plasma treatment of the preoxidized wafer surface.¹¹⁵
- *Post-Oxidation.* This method consists of the implant of N into the gate poly followed by an anneal to diffuse the nitrogen into the oxide.¹¹⁶

Each method has its own pros and cons. However, since much of the understanding of what happens to nitrided oxides in general can be obtained by a discussion of RNO, we discuss only the NH₃ RNO approach in detail. The interested reader is referred to the references for more information on the other techniques.

Nitridation during Oxidation

Reoxidized Nitrided Oxides (RNO) From the historical perspective, we begin with the techniques that were developed to introduce nitrogen into the oxide during oxidation. The technique consists in growing the oxide in the conventional manner followed by a nitridation in which NH_3 is introduced to the oxide for a certain length of time. Finally, an anneal step (also called *reoxidation*) completes the process. This last step is mostly performed in an oxygen ambient. The net result is a buildup of nitrogen, which peaks at the Si– SiO_2 interface. Typically, the reoxidation step is twice as long as the nitridation step. The presence of nitrogen at the interface prevents further oxidation, at least until the reoxidation step is at least 4 times that of the nitridation.¹¹⁷ This reoxidation step is important in ridding the oxide of its trapping properties.

The first reports of the effect of nitridation on hot-carrier hardness were in the early 1980s. These reports showed that the interface state creation was suppressed following the exposure of the newly grown oxide to an ammonia anneal, with lifetimes as much as a factor of 100 longer than conventional oxides. However, this effect was mitigated by the presence of large quantities of electron traps in the oxides. Later, it was discovered that a reoxidation following the nitridation step would greatly decrease the propensity of the oxide to trap electrons¹⁰⁶ and this led to considerable research into RNO oxides^{106–111}. The initial nitriding approach to avoid too heavily nitriding the oxide was to use diluted gases (low-pressure oxidation systems were used¹⁰⁸). Later RTN was also used.¹⁰⁷

Figure 6.25¹⁰⁷ summarizes the effect of RNO on the hot-carrier properties of n-MOS transistors. In this Figure, the familiar peak of the stress at $I_{b(\text{max})}$ conditions⁹ can be seen in the pure oxide samples. In the various reoxidized nitrided-oxide curves, the peak of damage at $I_{b(\text{max})}$ has been completely suppressed, and there is a rising tail in the damage at high gate voltages. Longer nitridation times result in larger gains in hot-carrier hardness, as the percentage of nitrogen at the interface is increased.¹¹¹ This figure also shows that the propensity of the oxide to trap electrons is somewhat suppressed by anneals in oxygen or nitrogen ambients.

The mechanism behind the hot-carrier stress hardness is the presence of nitrogen in the interfacial region, which is responsible for the suppression of interface states. Carr and Buhrman¹¹⁸ have shown that the bonded nitrogen is substituted for the oxygen in the oxide, becoming bonded to two silicon atoms. However, in the interfacial region, the nitrogen becomes bonded to three silicon atoms. It is this nitrogen that passivates the interface against N_{it} generation, possibly by bonding to hydrogen-terminated interface states.

Another explanation is based on the Lai hole trap-interface state creation model.¹⁶ Ma et al.¹¹⁹ suggest that interfacial strain caused by nitrogen incorporation inhibits the conversion of holes trapped in the interfacial region from transforming to interface states. A further explanation based on gate current measurements by Hori et al.,¹²⁰ is that the hole gate current is suppressed by the nitridation process. The decrease in the hole concentration would also result in less interfacial holes being

trapped and would, consequently, decrease the interface state creation. The decrease in the gate voltage damage at very low gate voltage stressing^{52,121} would also support this. No conclusive experiment has been performed to indicate which of the approaches is correct. It is possible that all three contribute.

In the case of electron trapping, the presence of hydrogen in the NH_3 nitridation process is thought to be responsible for the increased electron trapping seen in both n-MOS and p-MOS transistors. Studies of the p-MOS hot-carrier hardness for RNOs show that although there is considerable loss of hot-carrier hardness (to electron trapping),¹²² the reoxidation step reduces the susceptibility of the dielectric to trapping. However, considerable reoxidation times are required to achieve this, and the oxide is still inferior to pure oxides for low gate voltage stressing in p-channel devices.

The nitridation of the oxide does not come without a price. Table 6.2 shows the effects on the transistor performance of the degree of nitridation (based on a similar table by Momose et al.¹¹⁰). It can be seen that, to choose the best RNO conditions, a balance of the various deleterious effects has to be made. This balance consists in trading off increased mobility degradation, increased fixed-hole traps (Q_{ox}), and increased initial interface state density against hot-carrier hardness. From the hot-carrier point of view, the gain in n-MOS HC hardness has to be balanced by the decreased hot-carrier hardness of the p-MOS transistors. Consequently, a nitrogen concentration exists in which the process is optimized for both performance and hot-carrier hardness. Momose et al. suggest that this is at approximately 1 at.%,¹¹⁰ represented by the ellipse in Table 6.2.

N_2O A solution to the problem of the presence of hydrogen can be found in the oxidation in a nitrous oxide ambient (N_2O). N_2O -grown oxides offer a method of adding nitrogen to the oxide, without adding hydrogen. This method, developed initially by Kwong and co-workers,¹¹³ involves the addition of nitrogen to the interface during the oxidation process, as opposed to after the oxide has been grown. The result is an oxide that has nitrogen distributed throughout the oxide,¹¹⁸ although growing oxide in N_2O through an RTO process results in an oxide with nitrogen peaked at the Si-SiO₂ interface.¹¹⁸ This process also results in lower nitrogen concentrations than with RNO but does not suffer the increased electron trapping that the RNOs show. The effect is to increase the hot-carrier hardness of both the n-MOS and p-MOS hot-carrier properties (Fig. 6.26), although the n-MOS increase in hot-carrier hardness is not as large as is seen with NH_3 nitridation.

Following the work of Okada et al.,¹¹⁴ who showed that the active ingredient in the N_2O oxidation process was nitric oxide (NO), NO itself is now also used as the gaseous species. The result was an increase in nitrogen at the interface plus the increased beneficial hot-carrier hardness resulting from a higher N dose in the oxide.¹¹⁴

Pre- and Postoxidation Introduction of Nitrogen

Three further methods are briefly mentioned here. The first is the introduction of nitrogen into the silicon before oxidation. In this method, an implant of nitrogen is

TABLE 6.2 Effect of Nitrogen Concentration in Oxide of RNO Oxides on Different Properties of MOS Transistors

Nitrogen concentration (at%)	0	0.1	1	10
V_t	Equal to		Much worse	worst
N_{it}				
<u>Mobility</u>				
<u>n-MOS</u>				
μ -low field				
μ -high field				
<u>p-MOS</u>				
μ -low field				
μ -high field				
<u>B-Penetration</u>	Worst	Better		Best
<u>HC Damage</u>				
<u>n-MOS</u>				
N_{it}				
N_{oxh}				
N_{oxe}				
<u>p-MOS</u>				
N_{oxe}				
N_{it}				
N_{oxh}	/	?	?	?

performed before oxidation.¹¹² The nitrogen species diffuses to the interface during the oxidation phase and becomes incorporated into the oxide. The result is an oxide that is both more hot-carrier-resistant, but also effectively stops boron penetration. The downside of this approach is that the yield of oxides made in this way is not high, possibly due to the varying degree of nitrogen incorporation across the wafer, and the possibility of local oxide thinning.¹²³

The second method of introducing N into the oxide is to implant it into the poly following the oxide growth. This method successfully adds nitrogen to the oxide but has a tendency to pile up at the oxide-polysilicon interface rather than the oxide-silicon interface. Although this method is highly effective in preventing boron penetration, it is less effective in increasing hot-carrier hardness.¹¹⁶

The final method of creating hot-carrier-hard oxides involves the incorporation of fluorine into the oxide.¹²⁴ This method involves the implant of fluorine into the gate oxide followed by an anneal. The effect is to increase the hot-carrier hardness of the oxides, but by a factor of only 2 or so. The breakdown properties of the oxide were also greatly improved.

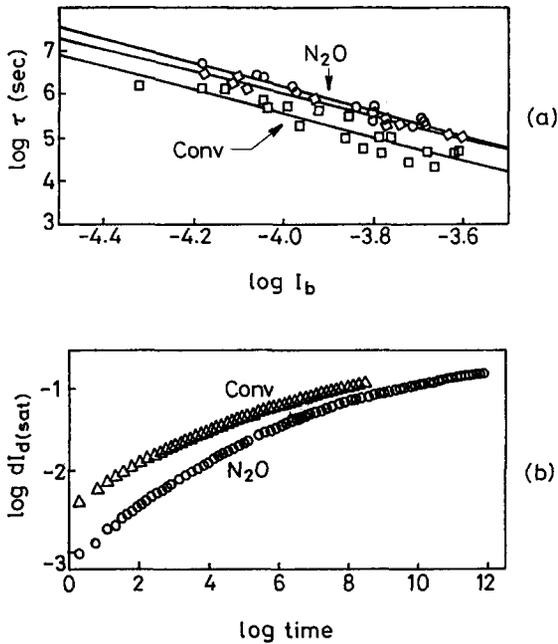


Figure 6.26 (a) Log lifetime versus log substrate current for n-MOS transistors for conventional (squares) and two N_2O oxides (diamonds, circles); (b) for p-MOS transistors for conventional (triangles) and N_2O -nitrided oxides (circles). The effect of N_2O oxidation is to increase the hot-carrier lifetime.

6.7.3 Back-End Processing

Alloy Anneal

It is well known that hydrogen plays a role in the passivation of interface states—alloy anneals at the end of the process play an important role in improving the subthreshold characteristics of a MOS transistor. Kizilyalli et al.¹²⁵ have shown that hydrogen is also involved in interface state creation during hot-carrier stressing, and, more surprisingly, that alloy annealing in deuterium increases the lifetime of an n-MOS transistor stressed under interface state creation conditions, by an order of magnitude over the standard alloy or sinter anneals. This confirms the role of deuterium seen in experiments performed on irradiated transistors by Saks and Rendell.¹⁹ The proposed mechanism is that, in the formation of an interface state, the Si-H (or in this case, Si-deuterium) bond is broken. The larger mass of the deuterium atom slows down the velocity at which the deuterium leaves the broken bond site, giving the bond the necessary time to reform. The beneficial effect of deuterium depends intimately on the sinter annealing conditions and the ability of deuterium to diffuse to the Si-SiO₂ interface to passivate the dangling bonds. The presence of a SiN capping layer, for instance, is known to prevent hydrogen (deuterium) from getting to this interface. Nitride spacers could have a similar effect.

Linked to the role of hydrogen in the processing is what has been called poststress interface trap creation. Bellens et al.¹¹⁸ showed that if a transistor is grounded after stress, the interface state concentration continued to grow even after the voltages have been removed from the device, at a rate that depends on the log of the time the device is left grounded. Through a series of experiments, they established that hydrogen played a role in this generation and that the presence of trapped holes in the oxide was a prerequisite for this phenomenon. A silicon nitride passivation layer was found to play an important role in the process, either as a source of hydrogen, or as a barrier preventing hydrogen from diffusing outward.

The effect of back-end processing has also been found to be detrimental to hot-carrier hardness and these effects have been linked to the presence of water in the ILD film. Takayanagi et al.¹²⁶ report that SOG and TEOS-O₃ showed that the damage created during hot-carrier stress [$I_b(\text{max})$ conditions] showed a steady increase in HC damage with time of anneal and that the increased damage scaled linearly with the anneal time. Fourier transform-infrared FTIR measurements showed that these films were losing water during the anneal. This was also confirmed by Shimokawa et al.,¹²⁷ who also showed that capping the ILD below these SOG and TEOS-O₃ films with plasma CVD SiO₂ (P-SiO) effectively prevents the diffusion of water to the Si-SiO₂ interface, preventing the increased HC effect seen with these water-rich films. Another study of the role of the ILD in hot-carrier reliability by Machida et al.¹²⁸ showed that these films not only showed the presence of H₂O, but also showed strong hydrogen signals (using TDS). The presence of molecular hydrogen suggests that the reaction might be



In other words, hydrogen is being introduced to the interface by water, and during hot-carrier stressing, this hydrogen reacts with the interface, resulting in an interface state and releasing molecular hydrogen.

6.8 SUMMARY AND FUTURE TRENDS

Several recent and projected future trends (at the time of writing) affecting the hot-carrier effect are summarized below.

A description of the hot-carrier effect under all drain and gate biases now exists for both n-channel and p-channel transistors. This complete description includes interface states, positive and negative traps in the oxide, and relaxable damage, due to near-interface states. This complete description has enabled ac hot-carrier stressing models to be developed with a high degree of accuracy for the n-MOS transistors. For p-channel devices, work on the link of the damage to the different device voltage conditions is still needed before a complete ac model is possible.

As the effective lengths of the transistors are reduced to the <100-nm range, velocity overshoot becomes increasingly important, as discussed briefly in Section 6.6. At these dimensions, the carriers can pass through the high-field region without

suffering a collision and without being scattered and are injected into the oxide (non-stationary transport regime). Transistors in this regime would be expected to suffer less hot-carrier damage. It has been seen that very short-gate-length transistors show less impact-ionized substrate current, but definitive data showing decreased hot-carrier damage have yet to be published.

Decreasing oxide thickness plays an important role in the hot-carrier effect in two ways:

1. As the thickness of the gates approaches the tunneling distance, charge trapping tends to become less important and the principal HC damage goes from predominantly electron trapping to interface state creation.⁴² This can be understood in terms of the tunneling distance of the charge from the oxide. Cohen et al. have shown³³ that oxide traps near the Si–SiO₂ begin to behave like interface states, when these states are situated between 1.2 and 1.8 nm from the interface. Thus, for oxides of thicknesses of the order of 4.0 nm and thinner, the problem of oxide trapping steadily diminishes, and the interface state creation mechanism becomes the predominant factor, for both n- and p-channel transistors. Beyond this, the advent of high dielectric-constant dielectrics to replace SiO₂ (if such a search is successful) will spawn a further chapter in the study of the hot-carrier effect.

2. Equally important, at some point in gate oxide scaling, the gate oxide begins to suffer significant leakage due to direct tunneling. This point is reached around 3 nm, although the thickness at which the oxide begins to improve from the plasma damage depends very much on the quality of the oxide—oxides with less leakage current for a given oxide thickness will suffer more oxide damage. When the gate oxide begins to leak sufficiently, it will no longer be completely capacitive, but will have a resistive component. Charge building up on the gate during plasma processing will start to leak away. At a certain oxide thickness, it will no longer be possible to maintain sufficient potential across the oxide to cause plasma damage and the plasma charging issue will be largely solved through gate oxide scaling—if the circuits are found to support these high levels of leakage current.

Finally, with respect to voltage, Figure 6.27 shows the lateral field (approximated by the supply voltage over the effective length of the shortest transistor) taken from the Silicon Industry Association's *National Technology Roadmap* (NTR).¹²⁹ There are several trends evident in this figure. For the longer transistor lengths, the lateral fields have increased with each succeeding generation. This was because the power supply was held constant in order to preserve compatibility of voltages between the chips on an integrated circuit board. It is these increases in the lateral field that have led to the interest in the HCE in the 1990s. However, recent generations have dropped power supply voltages, leading to decreasing lateral fields in the channel region over the past few lithography generations and, consequently, there has been less concern with the issue of hot carriers than there was before.

Looking into the future, however, two forces are driving the power supply: channel-length scaling, or the need to reduce the power that chips output; and the need to increase the circuit speed. These two forces and the application of the

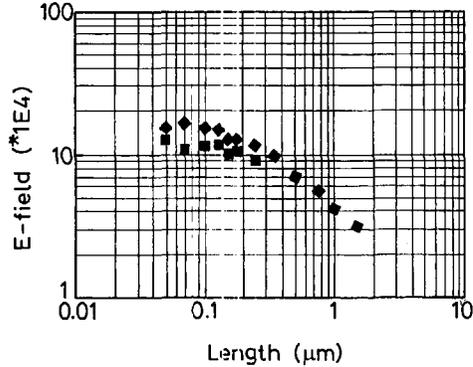


Figure 6.27 Lateral field versus technology generation. (From *National Silicon Technology Roadmap*, 1997.)

particular IC, drive the choice of the power supply. In Figure 6.27, this is captured by the two curves for generations below 200 nm. The higher curve represents the high-speed-high-power needs, while the lower curve represents the applications where low power is needed (from the *National Technology Roadmap*¹²⁹). It can be seen from these curves that, although the low power curve tends to saturate with decreasing transistor gate length, the lateral fields tend to increase for the more power-intensive applications, suggesting that the hot-carrier effect will be with us for some time to come.

REFERENCES

1. S. A. Abbas and R. C. Dockerty, "N-Channel IGFET Design Limitations due to Hot-Electron Trapping," *IEDM Tech. Digest* 35, (1975).
2. T. H. Ning, C. M. Osburn, and H. N. Yu, "Effect of Electron Trapping in IGFET Characteristics," *J. Electron. Mat.* **6**, 93 (1977).
3. C. Hu, S. C. Tam, F. C. Hsu, P. K. Ko, T.-Y. Chan and K. W. Terrell, "Hot-Electron-Induced MOSFET Degradation—Model, Monitor and Improvement," *IEEE Trans. Electron Devices* **32**, 375 (1985).
4. N. S. Saks, P. L. Heremans, L. Van Den Hove, R. F. de Keersmaecker, and G. J. Declerck, "Observation of Hot-Hole Injection in MOS Transistor Using a Modified Floating Gate Technique," *IEEE Trans. Electron Devices* **33**, 1529 (1986).
5. G. A. Scoggan and T. P. Ma, "Effects of Electron-Beam Radiation on MOS Structures as Influenced by the Silicon Dopant," *J. Appl. Phys.* **48**(1), 294 (1977).
6. H. Haddara and S. Cristoloveanu, "Two-Dimensional Modeling of Locally Damaged Short-Channel MOSFET's Operating in the Linear Region," *IEEE Trans. Electron Devices* **ED-34**, 378 (1987).
7. B. S. Doyle, D. K. Krakauer, and K. R. Mistry, "High Current (Hot-Carrier, Snap-Back, ESD) Stress Damage in n-MOS Transistors," *IEEE Trans. Electron Devices*, **ED-40**, 980 (1993).

8. B. Doyle, C. Bergonzoni, R. Benecchi, and A. Boudou, "The Influence of Gate Edge Shape on the Hot Carrier Stress Damage in n-MOS Transistors" *IEEE Electron Device Lett.* **EDL-12**, 363 (1991).
9. E. Takeda, A. Shimizu, and T. Hagiwara, "Role of Hot-Hole Injection in Hot-Carrier Effect and Small Degraded Channel Region in MOSFET's," *IEEE Electron Device Lett.* **EDL-4**, 3291 (1983).
10. T. Poorter and P. Zoestbergen, "Hot Carrier Effects in MOS Transistors," *IEDM Tech. Digest* 100 (1984).
11. T. Tsuchiya, T. Kobayashi, and S. Nakajima, "Hot Carrier Degradation Mechanism in Si nMOSFET's," *7th Conf. Solid State Device Mat.* 21 (1985).
12. C. Lombardi, P. Olivio, B. Ricco, E. Sangiorgi, and M. Vanzi, "Hot Electrons in MOS Transistors: Lateral Distribution of the Trapped Oxide Charge," *IEEE Electron Device Lett.* **EDL-3**, 215 (1982).
13. J. S. Brugler and P. G. A. Jespers, "Charge Pumping in MOS Devices," *IEEE Trans. Electron Devices* **ED-16**, 297 (1969).
14. G. Groeseneken, H. E. Maes, N. Beltran, and R. D. de Keersmaecker, "A Reliable Approach to Charge-Pumping Measurements in MOS Transistors," *IEEE Trans. Electron Devices* **ED-31**, 42 (1984).
15. P. Heremans, R. Bellens, G. Groeseneken, and H. E. Maes, "A Consistent Model for the Hot-Carrier Degradation in n-Channel and p-Channel MOSFET's," *IEEE Trans. Electron Devices* **35**, 2194 (1988).
16. S. K. Lai, "Two Carrier Nature of Interface State Generation in Hole Trapping and Radiation Damage," *Appl. Phys. Lett.* **38**, 58 (1981).
17. G. Hu and W. C. Johnson, "Relationship between Trapped Holes and Interface States in MOS Capacitors," *Appl. Phys Lett.* **36**, 590 (1980).
18. R. Bellens, E. de Schrijver, G. Van den bosch, G. Groesenenen, P. Heremans, and H. E. Maes, "On the Hot-Carrier-Induced Post-Stress Interface Trap Generation in n-Channel MOS Transistors," *IEEE Trans. Electron Devices.* **ED-41**, 413–419 (1994).
19. N. S. Saks and R. W. Rendell, "The Time-Dependence of Post-Irradiation Interface Trap Build-up in Deuterium-Annealed Oxides," *IEEE Trans. Nucl. Sci.* **39**, 2220 (1992).
20. W. Weber, C. Werner, and A. Schwerin, "Degradation in n-MOS Transistors after Pulsed Stress," *IEEE Electron Device Lett.* **EDL-5**, 518 (1994).
21. B. S. Doyle, M. Bourcerie, C. Bergonzoni, R. Benecchi, A. Bravais K. R. Mistry, and A. Boudou, "The Generation and Characterization of Electrons and Hole Traps Created by Hot Hole Injection during Low Gate Voltage Hot Carrier Stressing of n-MOS Transistors," *IEEE Trans Electron Devices* **ED-37**, 1869 (1990).
22. R. Bellens, P. Heremans, G. Groeseneken, and H. E. Maes, "Hot-Carrier Effects in n-Channel MOS Transistors under Alternate Stress Conditions," *IEEE Electron Device Lett.* **EDL-9**, 232 (1988).
23. C. Bergonzoni and B. S. Doyle, "Simulations of the Aging Effects in MOS Transistors," *Proc. ESSDERC Conf.* 1987, p. 721.
24. P. Heremans, R. Bellens, G. Groeseneken and H. E. Maes, "Comments on 'The Generation and Characterization of Electrons and Hole Traps Created by Hot Hole Injection During Low Gate Voltage Hot Carrier Stressing of n-MOS Transistors,'" *IEEE Trans. Electron Devices* **ED-39**, 458–460 (1992).

25. B. S. Doyle, K. R. Mistry, M. Bourcier, C. Bergonzoni, R. Benecchi, A. Bravaix, and A. Boudou, "Reply to" Comments on "The Generation and Characterization of Electrons and Hole Traps Created by Hot Hole Injection During Low Gate Voltage Hot Carrier Stressing of n-MOS Transistors'," *IEEE Trans. Electron Devices* **ED-39**, 460 (1992).
26. W. Weber, M. Brox, R. Thewes, and N. S. Saks, "Hot-Hole-Induced Negative Oxide Charge in n-MOSFET's," *IEEE Trans. Electron Devices* **ED-42**, 1473–1480 (1995).
27. I.-C. Chen, S. Holland, and C. Hu, "Electron Trap Generation by Recombination of Electrons and Holes in SiO₂," *J. Appl. Phys.* **61**, 4544 (1987).
28. A. R. Stivers and C. T. Sah, "A Study of Oxide Traps and Interface States of the Silicon-Silicon Dioxide Interface," *J. Appl. Phys.* **51**, 6292 (1980).
29. M. Bourcier, B. S. Doyle, J.-C. Marchetaux, J.-C. Soret, and A. Boudou, "Relaxable Damage in Hot Carrier Stressing of n-MOS Transistors—Oxide Traps in the Near Interfacial Region of the Gate Oxide," *IEEE Trans. Electron. Devices* **ED-37**, 708 (1990).
30. H. Muto, M.-H. Tsai, and T.-P. Ma, "Random Telegraph Signals in Small MOSFETs after X-ray Irradiation," *IEEE Trans. Nucl. Sci.* **38**(6) (Part 1), 1116 (1991).
31. M.-H. Tsai and T.-P. Ma, "Effect of Radiation-Induced Interface Traps on 1/f Noise in MOSFET's," *IEEE Trans. Nucl. Sci.* **39**(6) (Part 1), 2178 (1992).
32. R. E. Paulsen, R. R. Siergiej, M. L. French, and M. H. White, "Observation of Near-Interface Oxide Traps with the Charge-Pumping Technique," *IEEE Electron Device Lett.* **EDL-13**, 627 (1992).
33. N. L. Cohen, R. E. Paulsen, and M. H. White, "Observation and Characterization of Near-Interface Oxide Traps with C-V Techniques," *IEEE Trans. Electron Devices* **ED-42**, 2004 (1995).
34. B. Doyle, M. Bourcier, J.-C. Marchetaux, and A. Boudou, "Interface State Creation and Charge Trapping in the Medium-to-High Gate Voltage Range ($V_d/2 > V_g > V_d$) during Hot Carrier Stressing of n-MOS Transistors," *IEEE Trans. Electron. Devices* **ED-37**, 744 (1990).
35. B. S. Doyle, K. R. Mistry, and J. Faricelli, "Examination of the Time Power Law Dependencies in Hot Carrier Stressing of n-MOS Transistors," *IEEE Electron Device Lett.* **EDL-18**, 51 (1997).
36. B. S. Doyle and G. J. Dunn, "Recovery of Hot-Carrier Damage in Reoxidized Nitrided Oxide MOSFET's," *IEEE Electron Device Lett.* **EDL-13**, 38–40 (1992).
37. B. Doyle and K. Mistry, "Examination of Gradual Junction P-MOS Structures for Hot Carrier Control Using a New Lifetime Extraction Method," *IEEE Trans. Electron Devices* **ED-39**, 2290 (1992).
38. B. Doyle and K. Mistry, "The Characterization of Hot-Carrier Damage in p-Channel Transistors," *IEEE Trans. Electron Devices* **ED-40**, 152 (1993).
39. M. Brox, A. Schwerin, Q. Wang, and W. Weber, "A Model For the Time- and Bias-Dependence of p-MOSFET Degradation," *IEEE Trans Electron Devices* **ED-41**, 1184 (1994).
40. B. S. Doyle and K. R. Mistry, "A Lifetime Prediction Method for Hot-Carrier Degradation in Surface-Channel p-MOS Devices," *IEEE Trans. Electron Devices* **ED-37**, 1301–1307 (1990).

41. R. Woltjer, A. Hamada, and E. Takeda, "Time Degradation of p-MOSFET Hot-Carrier Degradation Measured and Interpreted Consistently over Ten Orders of Magnitude," *IEEE Trans. Electron Devices* **ED-40**, 392 (1993).
42. T. Tsuchiya et al., "Hot Carrier Degradation Mode and Prediction method of D.C. lifetime in Sub-Micron PMOSFETs," *Ext. Abst. SSDM Conf Symp. Sendai* 291 (1990).
43. R. Woltjer, G. M. Paulzen, H. Lifka, and P. Woerlee, "Positive Oxide-Charge Generation During 0.25 μm PMOSFET Hot-Carrier Degradation," *IEEE Electron Device Lett.* **EDL-15**, 427 (1994).
44. R. Woltjer, G. M. Paulzen, H. G. Pomp, H. Lifka, and P. H. Woerlee, "New Hot-Carrier Degradation Mechanisms in 0.25 μm PMOSFET's," *Symp. VLSI Technol.* 141 (1994).
45. B. S. Doyle, B. J. Fishbein, and K. R. Mistry, "NBTI-Enhanced Hot Carrier Damage in p-Channel MOSFET's," *IEDM Proc.* 529 (1991).
46. G. La Rosa, F. Guarin, S. Rauch, A. Acovic, J. Lukaitis, and E. Crabbe, "NBTI-Channel Hot Carrier Effects in PMOSFETs in Advanced CMOS Technologies," *IEEE Intl. Rel. Phys. Symp.* 282 (1997).
47. T. Choi, P. K. Ko, and C. Hu, "Hot-Carrier-Induced MOSFET Degradation under AC Stress," *IEEE Electron Device Lett.* **EDL-8**, 333 (1987).
48. R. Bellens, P. Heremans, G. Groeseneken, and H. E. Maes, "Analysis of the Mechanisms for the Enhanced Degradation during AC Hot Carrier Stress of MOSFETs," *IEDM Tech. Digest* 212 (1988).
49. S. Subrahmaniam, J. Y. Chen, and A. H. Johnston, "MOSFET Degradation Due to Hot-Carrier Effect at High Frequencies," *IEEE Electron Device Lett.* **EDL-11**, 21 (1990).
50. K. R. Mistry and B. S. Doyle, "The Role of Electron Trap Creation in Enhanced Hot-Carrier Degradation During AC Stress," *IEEE Electron Device Lett.* **EDL-11**, 267 (1990).
51. R. Bellens et al., "The Influence of Measurement Setup on Enhanced AC Hot Carrier Degradation of MOSFETs," *IEEE Trans. Electron Devices* **ED-37**, 310 (1990).
52. B. S. Doyle and K. R. Mistry, "Hot Carrier Stress Damage in the Gate 'OFF' State in N-Channel Transistors," *IEEE Trans. Electron Devices* **ED-39**, 1774 (1992).
53. B. S. Doyle and K. R. Mistry, "A Lifetime Prediction Method for Oxide Electron Trap Damage Created During Hot Electron Stressing of n-MOS Transistors," *IEEE Electron Device Lett.* **EDL-12**, 178 (1991).
54. K. R. Mistry, B. S. Doyle, A. Philipossian, and D. B. Jackson, "AC Hot Carrier Lifetimes in Oxide and ROXNOX n-Channel MOSFET's," *IEDM Proc.* 727 (1991).
55. K. R. Mistry and B. S. Doyle, "AC Versus DC Hot-Carrier Degradation in n-Channel MOSFET's," *IEEE Trans. Electron Devices* **ED-40**, 96 (1993).
56. R. Woltjer and G. M. Paulzen, "Oxide-Charge Generation during Hot-Carrier Degradation of PMOST's," *Proc. IEDM* 713 (1993).
57. R. Woltjer, G. M. Paulzen, H. G. Pomp, H. Lifka, and P. H. Woerlee, "Three Hot-Carrier Mechanisms in Deep-Submicronic PMOSFET's," *IEEE Trans. Electron Devices* **42**, 109 (1995).
58. B. S. Doyle, unpublished results.
59. P. Heremans, J. Witters, G. Groeseneken, and H. E. Maes, "Analysis of the Charge Pumping Technique and Its Applications for the Evaluation of MOSFET Degradation," *IEEE Trans. Electron Devices* **ED-36**, 1318 (1989).

60. An overview and extensive references to the charge pumping technique can be found in Chapter 1 of C. T. Wang, ed., *Hot Carrier Design Considerations for MOS Devices and Circuits*, Van Nostrand-Reinhold, 1993.
61. W. Weber, M. Brox, R. Thewes, and N. S. Saks, "Hot-Hole-Induced Negative Oxide Charges in n-MOSFET's," *IEEE Trans. Electron Devices* **ED-42**, 1473 (1995).
62. D. Vuillaume, J.-C. Marchetaux, P.-E. Lippens, A. Bravaix, and A. Boudou, "A Coupled Study by Floating Gate and Charge-Pumping Techniques of Hot-Carrier-Induced Defects in Submicrometer LDD n-MOSFETs," *IEEE Trans Electron Devices* **ED-40**, 773 (1993).
63. W. Chan, and T.-P. Ma, "Oxide Charge Buildup and Spread-out during Channel-Hot-Carrier Injection in NMOSFET's," *IEEE Electron Device Lett.* **EDL-13**, 319 (1992).
64. N. S. Saks, and M. G. Ancona, "Determination of Interface Trap Capture Cross Sections Using Three-Level Charge Pumping," *IEEE Electron Device Lett.* **EDL-11**, 339 (1990).
65. R. G.-H. Lee, J.-S. Su, and S. S. Chung, "A New Method for Characterizing the Spatial Distribution of Interface States and Oxide-Trapped Charges in LDD n-MOSFET's," *IEEE Trans. Electron Devices* **ED-43**, 81 (1997).
66. H.-H. Li, Y.-L. Chu, and C.-Y. Wu, "A Novel Charge-Pumping Method for Extracting the Lateral Distribution of Interface-Trap and Effective Oxide-Trapped Charge Densities in MOSFET Devices," *IEEE Trans. Electron Devices* **ED-44**, 782 (1997).
67. U. Cilingiroglu, "A General Model for Interface Trap Charge-Pumping Effects in MOS Devices," *Solid State Electron* **28**, 1127 (1985).
68. H.-H. Li, Y.-L. Chu, and C.-Y. Wu, "A New Simplified Charge Pumping Current Model and its Model Parameter Extraction," *IEEE Trans Electron Devices* **ED-43**, 1857 (1996).
69. D. Vuillaume, and B. Doyle, "Properties of Hot Carrier Induced Traps in MOSFET's Characterized by the Floating-Gate Technique," *Solid State Electron.* **35**, 1099 (1994).
70. B. Doyle, J. Faricelli, K. Mistry, and D. Vuillaume, "Characterization of Oxide Trap and Interface State Creation during Hot Carrier Stressing of n-MOS Transistors Using the Floating Gate Technique," *IEEE Electron Device Lett.* **EDL-14**, 63 (1993).
71. P. Speckbacher, J. Berger, A. Asenov, F. Koch, and W. Weber, "The 'Gated-Diode' Configuration in MOSFETs, a Sensitive Tool for Characterizing Hot-Carrier Degradation," *IEEE Trans. Electron Devices* **ED-42**, 1287 (1995).
72. C. H. Ling, S. E. Tan, and D. S. Ang, "A Study of Hot Carrier Degradation in NMOSFET's by Gate Capacitance and Charge Pumping Current," *IEEE Trans. Electron Devices* **ED-42**, 1321 (1995).
73. A. Neugroschel, C.-T. Sah, K. M. Han, M. S. Carroll, T. Nishida, J. T. Kavalieros, and Y. Lu, "Direct-Current Measurements of Oxide and Interface Traps on Oxidized Silicon," *IEEE Trans. Electron Devices* **ED-42**, 1657 (1995).
74. Y. Tososhima, F. Matsuoka, H. Hayashida, H. Iwai, and K. Kanzaki, "A Study of Gate Oxide Thickness Dependence of Hot Carrier-Induced Degradation for n-MOSFET's," *Proc. Symp. VLSI Technol.* 39 (1988).
75. K. Mistry and B. S. Doyle, "An Empirical Model for The L_{eff} Dependence of Hot Carrier Lifetimes of n-Channel MOSFET's," *IEEE Electron Device Lett.* **EDL-10**, 500 (1989).
76. K. Mistry, and B. Doyle, "Hot Carrier Degradation in n-MOSFET's Used as Pass Transistors," *IEEE Trans. Electron Devices* **EDL-37**, 2415 (1990).

77. L. Selma, E. Sangiorgi, R. Bez, and B. Ricco, "Measurement of the Hot Hole Injection Probability from Si into SiO₂ in p-MOSFETs," *Proc. IEDM* 333 (1993).
78. S.-I. Takagi, and A. Toriumi, "New Experimental Findings on Hot Carrier Transport under Velocity Saturation Regime in Si MOSFETs," *Proc. IEDM* 711 (1992).
79. J. J. Sanchez, K. K. Hsueh, and T. A. DeMassa, "Drain-Engineered Hot-Electron-Resistant Device Structures: A Review," *IEEE Trans. Electron. Devices* **ED-36**, 1125 (1989).
80. B. S. Doyle, and K. R. Mistry, "Anomalous Hot Carrier Behavior for LDD p-Channel Transistors," *IEEE Electron Device Lett.* **EDL-14**, 536 (1993).
81. A. F. Tasch, H. Shin, and C. M. Maziar, "A New Vertical Structural Approach for Reducing Hot Carrier Generation in Deep Sub-micron MOSFET's," *Proc. VLSI Symp.* 43-44 (1990).
82. J. E. Moon, C. Galewski, T. Garfinkel, M. Wong, W. G. Oldham, P. K. Ko, and C. Hu, "A Deep-Submicrometer Raised Source/Drain LEE Structure Fabricated Using Hot-Wall Epitaxy," *Proc. VLSITSA* 117 (1991).
83. S. Aur and R. A. Chapman, "Gate Oxide Thickness Effect on Hot Carrier Reliability in 0.35 μ m NMOS Devices," *IEEE Intl. Rel. Phys. Symp.* 48 (1994).
84. A. Hiroki, and S. Odanaka, "Gate-Oxide Thickness Dependence of Hot-Carrier-Induced Degradation in Buried p-MOSFET's," *IEEE Trans. Electron Devices* **ED-39**, 1223 (1992).
85. B. S. Doyle, K. R. Mistry, and C.-L. Huang, "Analysis of Gate Oxide Thickness Hot Carrier Effects in Surface Channel P-MOSFET's," *IEEE Trans. Electron Device* **ED-42**, 1 (1995).
86. H. S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S.-I. Nakamura, M. Aaito, and H. Iwai, "Tunnelling Gate Oxide Approach to Ultra-High Current Drive in Small-Geometry MOSFETS," *Proc. IEDM* 593-576 (1994).
87. A. Hamada, T. Furusawa, and E. Takeda, "A New Aspect on Mechanical Stress Effects in Scaled NOS Devices" *Symp. VLSI Technol.* 113 (1990).
88. R. Degraeve, I. De Wolf, G. Groeseneken, and H. E. Maes, "Analysis of Externally Imposed Mechanical Stress Effects on the Hot-Carrier-Induced Degradation in MOSFET's," *Proc. IRPS* 29 (1994).
89. N. Yamamoto, S. Iwata, and H. Kume, "The Influence of Internal Stresses in Tungsten-Gate Electrodes on the Degradation of MOSFET Characteristics Caused by Hot Carriers," *IEEE Trans. Electron Dev.* **ED-34**, 607 (1987).
90. B. S. Doyle, and D. Lau, "The Properties and Annealing of Gate Oxide Damage of Oxynitride Passivated CMOS Transistors Arising From Mechanical Stresses During Packaging," *IEDM Conf. Proc.* 91 (1989).
91. K. R. Mistry, Private Communication.
92. H. Shin, and C. Hu, "Monitoring Plasma-Process Induced Damage in Thin Oxides," *IEEE Trans. Semiconduct. Mfg.* **6**, 96 (1993).
93. K. Lai, K. Kumar, A. Chou, and J. Lee, "Plasma Damage and Photo-Annealing Effects of Thin Gate Oxide and Oxynitrides During O₂ Plasma Exposure," *IEEE Electron Device Lett.* **EDL-17**, 82 (1996).

94. H. Shin, C.-C. King, R. Moazzami, T. Horiuchi, and C. Hu, "Characterization of Thin Oxide Damage During Aluminium Etching and Photoresist Ashing Processes," *VLSITSA Proc.* 210–213 (1991).
95. K. Hashimoto, "Charge Damage Caused by Electron Shading Effects," *Jpn. J. Appl. Phys.* **33**, 6013–6018 (1994).
96. Y. Uraoka, K. Eriguchi, T. Tamaki, and K. Tsuji, "Evaluation Technique of Gate Oxide Damage," *IEEE Trans. Semiconduct. Mfg.* **7**, 293 (1994).
97. H. Shin, C.-C. King, and C. Hu, "Thin Oxide Damage by Plasma Etching and Ashing Processes," *IEEE Proc. IRPS* **37** (1992).
98. P. Tanner, S. Dimitrijević, Y.-T. Yeow, and H. B. Harrison, "Measurement of Plasma Etch Damage by a New Slow Trap Profiling Technique," *IEEE Electron Device Lett.* **EDL-17**, 515 (1996).
99. K. R. Mistry, B. J. Fishbein, and B. S. Doyle, "Effect of Plasma-Induced Charging Damage on n-Channel and p-Channel MOSFET Hot Carrier Reliability," *IRPS Proc.* **42** (1994).
100. J. C. King, and C. Hu, "Effect of Low and High Temperature Anneal on Process-Induced Damage of Gate Oxides," *IEEE Electron Device Lett.* **EDL-15**, 475 (1994).
101. R. Rakkhit, F. P. Heiler, P. Fang, and C. Sander, "Process Induced Oxide Damage and Its Implications to Device Reliability of Submicron Transistors," *IEEE IRPS Proc.* 293 (1993).
102. C. T. Gabriel, and M. J. Weling, "Gate Oxide Damage Reduction Using a Protective Dielectric Layer," *IEEE Electron Device Letters* **EDL-15**, 269 (1994).
103. Y. Yunogama, T. Mizutani, K. Suzuki, and S. Nishimatsu, "Radiation Damage in SiO₂/Si Induced by VUV Photons," *Jpn. J. Appl. Phys.* **28**, 2172 (1988).
104. M. Sherony, A. J. Chen, K. R. Mistry, D. A. Antoniadis, and B. S. Doyle, "Comparison of Plasma-Induced Charging Damage in Bulk and SOI MOSFETs," *IEEE Int. SOI Conf.* **21** (1993).
105. S. Ma, private communication.
106. T. Kaga, and T. Hagiwara, "Short- and Long Term Reliability of Nitrided Oxide MOSFET's," *IEEE Trans. Electron Devices* **ED-35**, 929 (1988).
107. T. Hori, and H. Iwasaki, "Improved Hot-Carrier Immunity in Submicrometer MOSFET's with Reoxidized Nitrided Oxides Prepared by Rapid Thermal Processing," *IEEE Electron Device Lett.* **EDL-12**, 64 (1989).
108. K. R. Mistry, and B. S. Doyle, "AC Hot Carrier Lifetimes in Oxide and ROXNOX n-Channel MOSFETs," *IEDM Proc.* 727 (1991).
109. B. J. Gross, K. S. Krisch, and C. G. Sodini, "An Optimized 850 C Low-Pressure-Furnace Reoxidized Nitrided Oxide (ROXNOX) Process," *IEEE Trans. Electron Devices* **ED-38**, 2036 (1991).
110. H. S. Momose, T. Morimoto, Y. Ozawa, K. Yamage, and H. Iwai, "Electrical Characteristics of Rapid Thermal Nitrided-Oxide Gate n- and p-MOSFET's with Less than 1 Atom% Nitrogen Concentration," *IEEE Trans. Electron Devices* **ED-41**, 546 (1994).
111. T. Hori, T. Yasui, and S. Akamatsu, "Hot-Carrier Effects in MOSFET's with Nitrided-Oxide Gate-Dielectrics Prepared by Rapid Thermal Processing," *IEEE Trans. Electron Devices* **ED-39**, 134 (1992).

112. C. T. Liu, E. J. Lloyd, Y. Ma, M. Du, R. L. Opila, and S. J. Hillenius "High Performance 0.2 mm CMOS with 25 A Gate Oxide Grown on Nitrogen-Implanted Si Substrates," *Proc. IEDM* 499 (1996).
113. H. Huang, W. Ting, D.-L. Kwong, and J. Lee, "Electrical and Reliability Characteristics of Ultrathin Oxynitride Gate Dielectric Prepared by Rapid Thermal Processing in N_2O ," *IEDM Proc.* 421 (1990).
114. Y. Okada, P. J. Tobin, K. G. Reid, R. I. Hegde, B. Maiti, and S. A. Ajuria, "Furnace Grown Gate Oxynitride Using Nitric Oxide (NO)," *IEEE Trans. Electron Devices* **ED-41**, 1608 (1994).
115. S. V. Hattangady, R. Kraft, D. T. Grider, M. A. Douglas, G. A. Brown, P. A. Tiner, J. W. Kuehne, P. E. Nicollian, and M. F. Pas, "Ultrathin Nitrogen-Profile Engineered Gate Dielectric Film," *Proc. IEDM* 495 (1996).
116. Y. Okazaki, S. Nakayama, M. Miyake, and T. Kobayashi, "Characteristics of Sub-1/4 mm Gate Surface Channel PMOSFET's Using a Multilayer Gate Structure of Boron-Doped Poly-Si on Thin Nitrogen-Doped Poly-Si," *IEEE Trans. Electron Devices* **ED-41**, 2369 (1994).
117. A. Philipossian, private communication.
118. E. C. Carr, and R. H. Buhrman, "Role of Interfacial Nitrogen in Improving Thin Silicon Oxides Grown in N_2O ," *Appl. Phys. Lett.* **63**, 54 (1993).
119. Z.-J. Ma, P. T. Lai, Z. H. Liu, P. K. Ko, and Y. C. Cheng, "Mechanisms for Hot-Carrier-Induced Degradation in Reoxidized-Nitrided-Oxide n-MOSFET's under Combined AC/DC Stressing," *IEEE Trans. Electron Devices* **ED-40**, 1112 (1993).
120. T. Hori, S. Akamatsu, and Y. Odake, "Deep-Submicrometer CMOS Technology with Reoxidized or Annealed Nitrided-Oxide Gate Dielectrics Prepared by Thermal Processing," *IEEE Trans. Electron Devices* **ED-39**, 118 (1992).
121. A. B. Joshi and D. L. Kwong, "Effects of AC Hot Carrier Stress on n- and p-MOSFET's with Oxynitride Gate Dielectrics," *IEEE Trans. Electron Devices* **ED-41**, 671 (1991).
122. B. Doyle, and A. Philipossian, "P-Channel Hot Carrier Optimization of RNO Gate Dielectrics through the Reoxidation Step," *IEEE Electron Device Lett.* **EDL-14**, 161 (1993).
123. M. Rennie, private communication.
124. Y. Nishioka et al. "Hot-Electron Hardened Si-Gate MOSFET Utilizing F Implantation," *IEEE Electron Device Lett.* **EDL-10**, 141 (1989).
125. I. C. Kizilyalli, J. W. Lyding, and H. Hess, "Deuterium Post-Metal Annealing of MOSFET's for Improved Hot Carrier Reliability," *IEEE Electron Device Lett.* **EDL-18**, 81 (1997).
126. M. Takayanagi, S. V. Takagi, I. Yoshii, and K. Hashimoto, "Characterization of Hot-Carrier-Induced Degradation of MOSFETs Enhanced by H_2O Diffusion for Multilevel Interconnection Processing," *IEDM Proc.* 703 (1992).
127. K. Shimokawa, T. Usami, S. Tokitou, N. Hirashita, M. Yoshimaru, and M. Ino, "Suppression of the MOS Transistor Hot Carrier Degradation Caused by Water Desorbed from Intermetal Dielectric," *Symp. VLSI Technol.* 96 (1992).
128. K. Machida, N. Shimoyama, J.-I. Takahashi, Y. Takahashi, N. Yabumoto, and E. Arai, "Improvement of Water-Related Hot-Carrier Reliability by Using ECR Plasma- SiO_2 ," *IEEE Trans. Electron Devices* **ED-41**, 709 (1994).
129. *National Silicon Technology Roadmap*, Semiconductor Industry Association, 1997.

PROBLEMS

- 6.1** A wet-oxide device suffers significantly greater high gate voltage-induced electron trapping ($N_{ot,e}$) than do dry-oxide devices. Determine the most significant damage mechanism under dc stressing conditions for a given wet-gate oxide transistor by estimating the lifetime for both interface state creation (N_{it}) and electron trapping (for a degradation of V_t of 0.1 V). Assume square-root and cube-root degradation rates for N_{it} and $N_{ot,e}$ respectively, and an oxide quality factor A factor of 1×10^{-5} and 1×10^{-3} for N_{it} and $N_{ot,e}$, respectively.
- 6.2** In Problem 6.1, how much would the oxide trap quality of the oxide have to be adjusted (as measured by the electron trapping prefactorial term) for the interface state and oxide trap mechanism to have equal lifetimes at a degradation level of 0.1 V.
- 6.3** An n-MOS transistor is found to have an order-of-magnitude shorter lifetime under interface state generation conditions than needed. Estimate the extent to which the LDD region has to be adjusted in terms of substrate current if the device is to have an acceptable lifetime (assume that the substrate current is one-tenth of the drain current initially and that the optimal LDD implant adjustment decreases the drive current by 10%).
- 6.4** A device ($\delta V_t = 100$ mV) stressed under electron-trap-creation conditions ($V_g = V_d$) has the following terminal currents: $I_d = 300 \mu\text{A}/\mu\text{m}$, $I_b = I_d/10$. Assuming that the impact ionization energy is 1.6 eV, the oxide barrier energy is 3.2 eV, the trapping efficiency is 1×10^{-6} /injected electron, and that 1×10^{11} traps/cm² give a 1-mV V_t shift, calculate the threshold voltage shift after 1 s, assuming linearity between injection and trapping (assume the length of the damage region between source and drain to be 0.1 μm).
- 6.5** Using the threshold voltage shift after 1 s as the oxide quality factor A in Eq. 6.3, calculate the lifetime of this device, assuming $n = 0.3$.
- 6.6** A device has a drain current of $300 \mu\text{A}/\mu\text{m}$ of gate width at lb (max) (interface state creation— N_{it}) conditions ($I_b/I_d = .01$), and $100 \mu\text{A}/\mu\text{m}$ at $V_g = V_d/5$ (neutral electron trap creation— $N_{ct,n}$) conditions ($I_b/I_d = 0.2$). Estimate the circuit lifetime of a transistor assuming a 10% duty cycle for *each* of the two mechanisms, and assuming K values of 10 and 1×10^{-3} , and m values of 3 and 10 for N_{it} and $N_{ot,h}$ creation, respectively.

DRAM and SRAM

HISASHI (SAM) SHICHIJO

Texas Instruments
Dallas, TX

7.1 INTRODUCTION

Dynamic random access memory (DRAM) and static random access memory (SRAM) are the most widely manufactured and used semiconductor memories. Figure 7.1 shows the memory hierarchy of a typical computer system including the personal computer (PC). The memory access speed is fastest at the top (first level cache) and the slowest at the bottom (hard disk). Most of the program codes and data used by a central processing unit (CPU) are stored in main memory. DRAMs are usually used as a main memory because of their low cost and high density capability. Approximately 65% of all the DRAMs manufactured in the world are now used as the main memory of PC systems. The first-level and second-level cache memories are inserted between the CPU and the main memory to store the most often and recently used data to speed up the overall data access. If the data needed by the CPU is not available in the cache memory (cache miss), the CPU must obtain the data from the slower main memory with a considerable speed penalty. SRAMs are used as the first and second level cache memories because of their faster access speed compared to DRAMs. Furthermore, the first level cache is now almost always integrated on the same chip as the CPU for even faster access.

7.2 DRAM CELL STRUCTURES

7.2.1 Memory Cell Concept

Except for some of the embedded applications (DRAM integrated with the logic chip), all the current DRAMs use a MOS transistor and a capacitor as a memory unit (cell). This is illustrated in Figure 7.2. The word-line is used to switch the pass

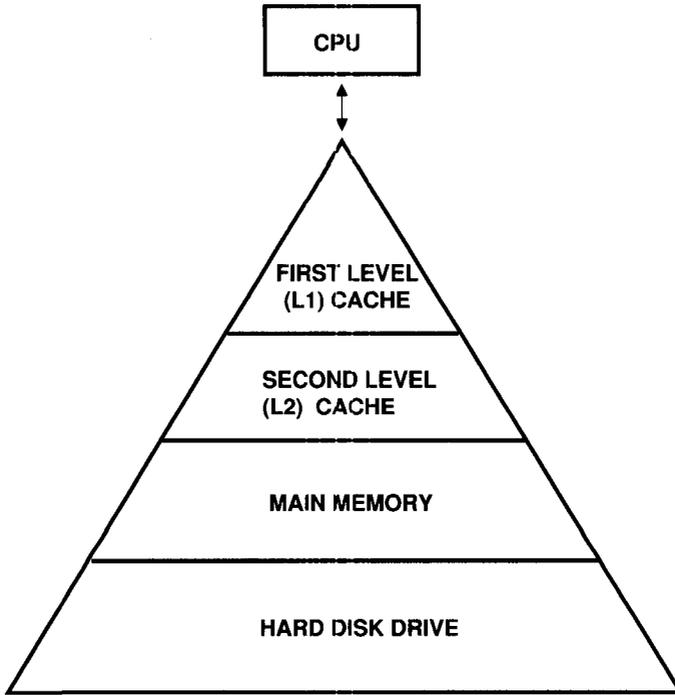


Figure 7.1 Memory hierarchy for a typical computer system.

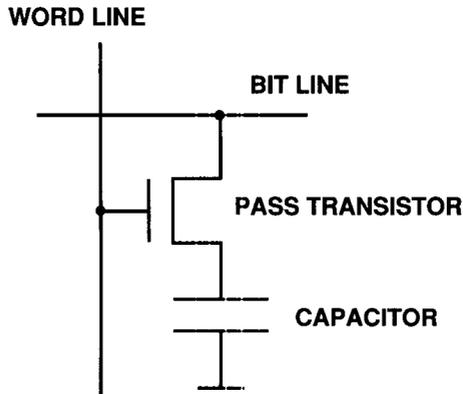


Figure 7.2 Schematic of a one-transistor, one-capacitor DRAM memory cell.

transistor on or off to connect the bit line to the capacitor or to isolate the capacitor from the bit line. The bit line is used to sense the stored charge in the memory cell during the read operation and to supply the data to the memory cell during the write operation.

When the capacitor is charged to a “high” voltage, which is usually the power supply voltage, the memory cell stores a “1” state. When the capacitor is

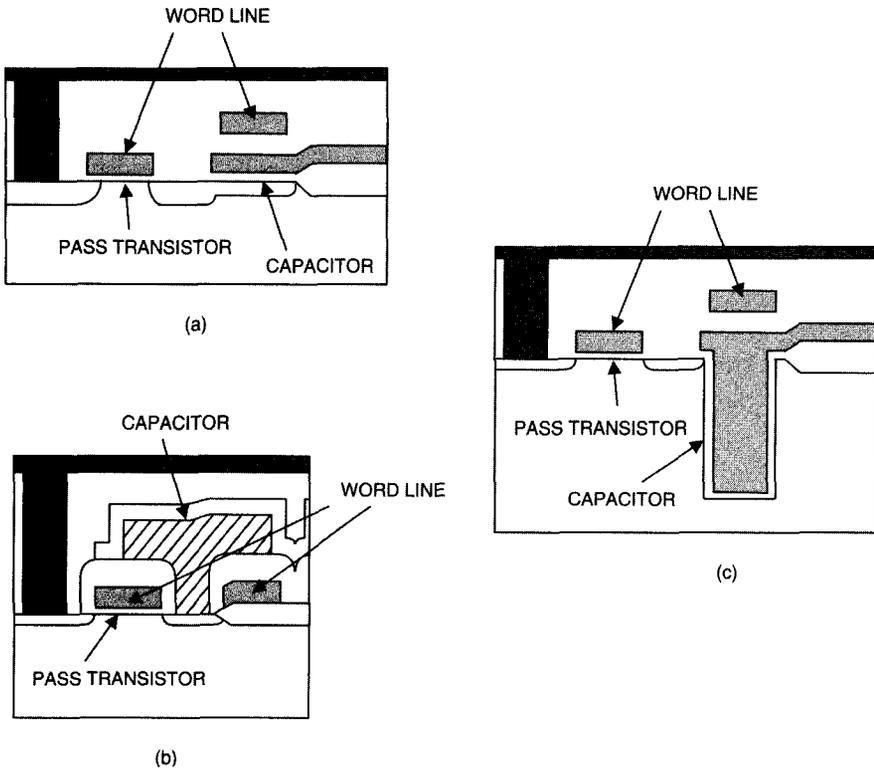


Figure 7.3 Three different DRAM memory cell structures: (a) planar capacitor DRAM cell; (b) stack capacitor DRAM cell; (c) trench capacitor DRAM cell.

charged to a “low” voltage, which is usually the ground, the memory cell stores a “0” state.

Figure 7.3 illustrates the physical implementation of the three types of DRAM cell structure. Figure 7.3a shows a “planar” cell where the capacitor dielectric film is formed between the planar n^+ -doped silicon surface and the upper heavily doped polysilicon layer. This structure was used in early generations of DRAMs with a density of up to 1 Mb, but is no longer used because the capacitor occupies too much area to permit high density. Figure 7.3b shows a “stack”-type structure where the capacitor dielectric is formed between the two heavily doped polysilicon layers and is stacked over the MOS pass transistor. Figure 7.3c shows a “trench”-type structure where the capacitor dielectric is formed inside a trench etched into the silicon substrate and between the n^+ -doped silicon and the upper, heavily doped polysilicon layer. The details of these structures are further described in the next section.

7.2.2 Cell Scaling

The key element of DRAM technology evolution has been the scaling down of the minimum feature size that allows a smaller memory cell size and higher density

integration for a given chip area and, therefore, a lower cost. Since the minimum feature size is going down by approximately 70% for each generation, the memory cell area can go down by $0.5 \times$ just by this scaling ($0.7 \times 0.7 = 0.49$). However, looking at the past trend of the minimum feature size and the DRAM memory cell size as shown in Figure 7.4, the memory cell size has actually gone down by a factor of $0.4 \times$ per generation, which is more than expected from the simple size scaling. The additional scaling has been made possible by reducing the capacitor area for each generation while keeping the cell capacitance basically constant. The reduced capacitor area can be compensated by reducing the capacitor dielectric thickness and providing three-dimensional structural enhancement such as a trench or stack structure.

This can be better understood by looking at the equation for the memory cell capacitance, C_s , as

$$C_s = \frac{\epsilon \epsilon_0 A}{t} \tag{7.1}$$

where ϵ is the relative dielectric constant of the capacitor film material, ϵ_0 is the dielectric constant of free space, A is the capacitor area, and t is the thickness of the capacitor film. Historically, because of the signal sensing and soft-error requirements (described in later sections), the minimum cell capacitance has been kept at

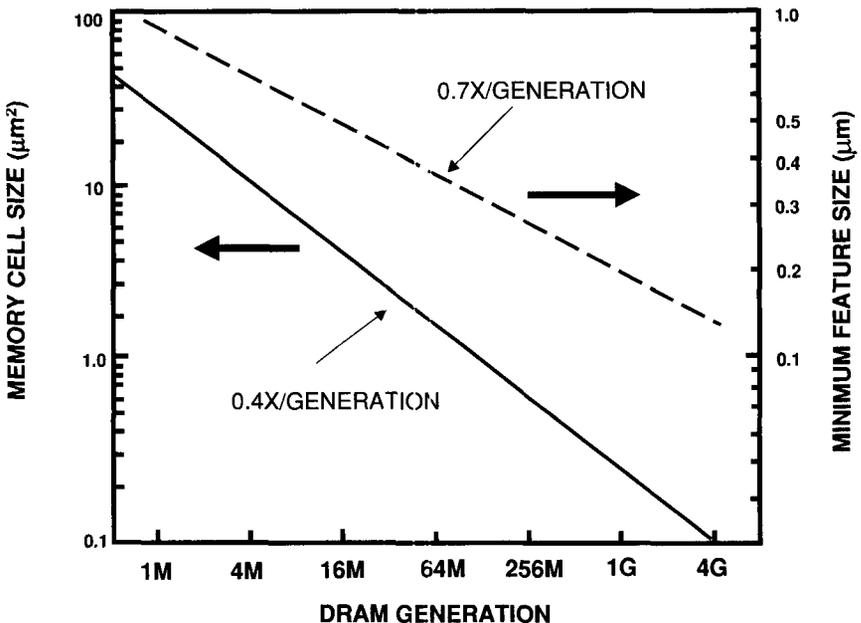


Figure 7.4 Historic trend of memory cell size and minimum feature size.

TABLE 7.1 Evolution of DRAM Cell Technology

DRAM Generation	1M	4M	16M	64M	256M	1G	4G
Minimum feature size (μm)	1.0	0.7	0.5	0.35	0.25	0.18	0.13
Capacitor equivalent oxide thickness (nm)	10	8	5	4	3	2	1
Dielectric material (bulk dielectric constant)	Nitride/oxide (7)				Ta ₂ O ₅ (25)		BST (500)
Cell structure	Planar	Stack or Trench					

25–30 fF for each generation of DRAMs. To maintain the same cell capacitance while the memory cell size and capacitor area are reduced, three approaches have been used.

1. Increase ϵ , that is, use a high-dielectric-constant material.
2. Reduce t , the dielectric film thickness.
3. Increase A , that is, use a three-dimensional capacitor structure (stack or trench).

Table 7.1 summarizes how these approaches have been applied to each generation of DRAMs.

First, the capacitor dielectric film has changed from pure oxide with a dielectric constant of 4 for 256 K (256-kilobits) DRAM to a nitride film with a dielectric constant of 7 for 1–64 Mb (megabits) DRAMs in the form of oxide/nitride/oxide (ONO) or nitride/oxide (NO) composite film. It is expected that future DRAMs beyond 256 M DRAM will utilize Ta₂O₅ (tantalum pentoxide, $\epsilon = 25$), BST (barium strontium titanate, $\epsilon = 500$), and then PZT (lead zirconate titanate, $\epsilon = 1000$). However, there are significant challenges in incorporating these materials into the DRAM process flow. The production-worthy deposition techniques for these films are not established yet, especially for BST and PZT. Furthermore, they may not be compatible with the existing DRAM process in terms of thermal stability and possible contamination. All of these new dielectric materials also require the use of metal electrodes which complicates the process requirement and poses a potential contamination problem.

In addition to the transition to higher-dielectric-constant materials and reduction of dielectric thickness, the DRAM has also gone through a significant change in its structure to gain additional capacitor area. The transition from a planar capacitor to either a stack capacitor or trench capacitor has been mentioned in Section 7.2.1. Even within the category of stack or trench capacitor, the structure has many

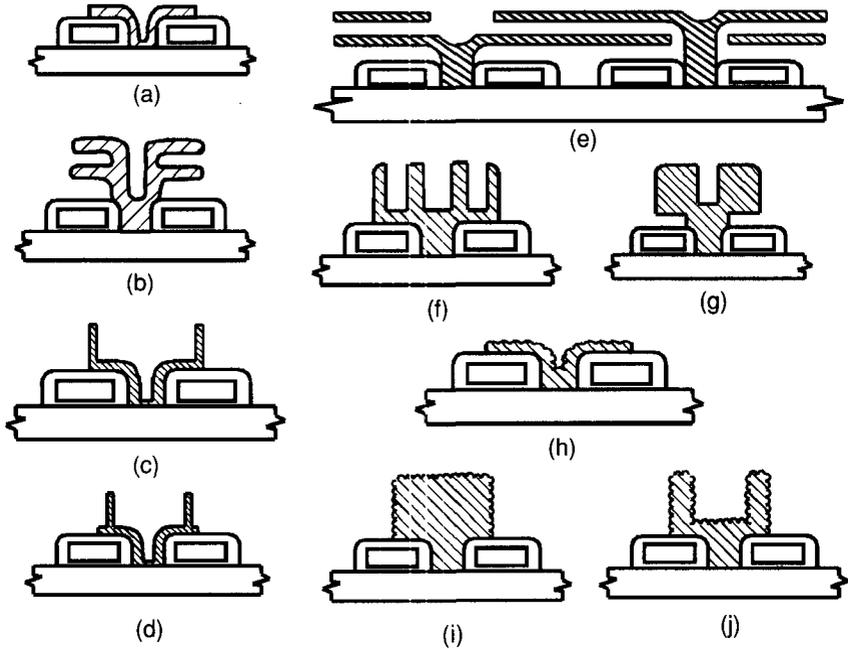


Figure 7.5 Various stack cell structures: (a) standard stack;¹ (b) fin stack;² (c) cylinder (crown) stack;³ (d) T-shaped cylinder stack;⁴ (e) spread stack;⁵ (f) double-crown stack;³ (g) dual-cell plate stack;⁶ (h) standard stack with a rugged poly;⁷ (i) simple stack with HSG;⁸ (j) cylinder stack with HSG.⁹

variations. For example, Figure 7.5 summarizes some of the variations of stack cells that have been reported. One noticeable technique is to roughen or texture the surface of the polysilicon electrode to increase the electrode surface area. This type of polysilicon is called rugged polysilicon⁷ (Fig. 7.5h) in the case of non selective deposition, and HSG (hemispherical grain)⁸ (Figs. 7.5i, j) in the case of selective deposition. The rugged polysilicon approach requires the etching back of polysilicon after the deposition to isolate the individual capacitors. Both techniques have been shown to increase the surface area by more than a factor of $1.8 \times$. This high area-enhancement factor (AEF) is obtained only after a careful optimization of the surface cleanup before the polysilicon deposition, the polysilicon deposition temperature, and the subsequent thermal processing. One advantage of using a high-dielectric constant material, especially BST or PZT, is that the capacitor structure can be simplified perhaps using a simple stack structure without going to these complicated three-dimensional structures.

Similar to the evolution of stack cell capacitors, the trench structure also have evolved over several generations of DRAMs from 1-Mb DRAM to 256-Mb DRAM. A simple trench structure as shown in Figure 7.3c was used for 1- and 4-Mb DRAMs by several manufacturers. Although various trench structures have since been

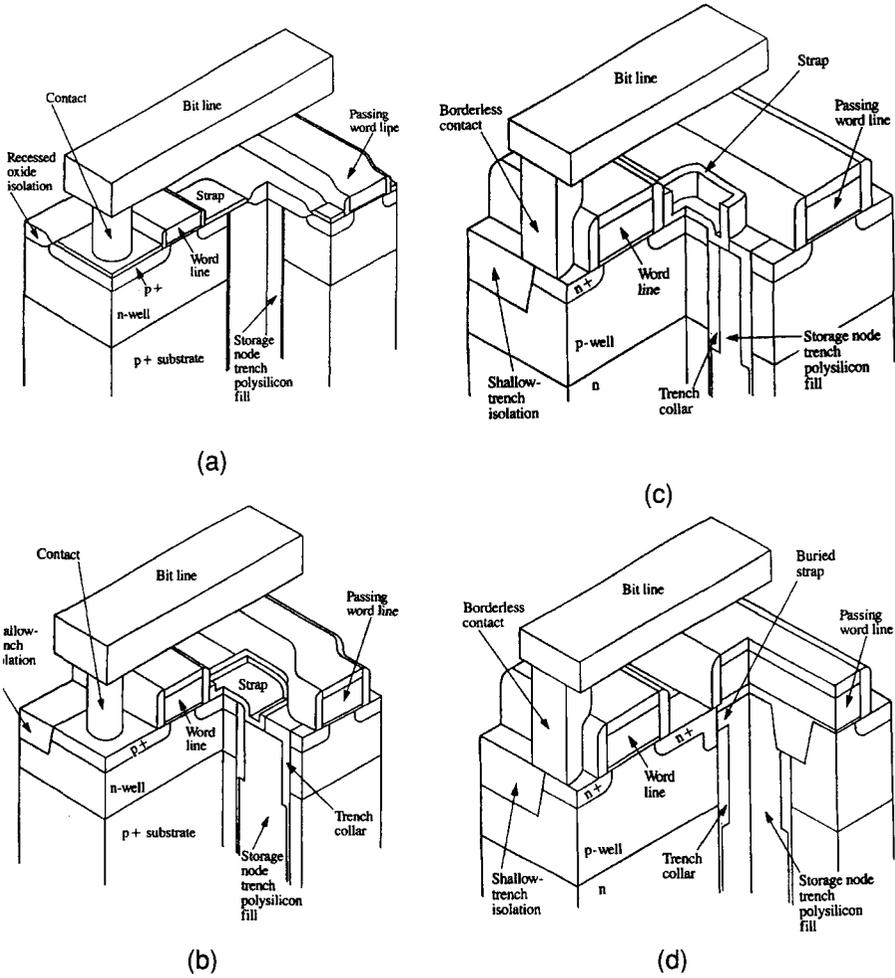


Figure 7.6 Evolution of IBM trench DRAM cell structure: (a) 4-Mb substrate plate trench (SPT); (b) 16-Mb merged isolation and node trench (MINT); (c) 64-Mb buried plate trench (BPT); (d) 256-Mb buried strap trench (BEST). (After Adler et al., Ref. 10, © 1995 International Business Machines Corporation. Reprinted with permission of the *IBM Journal of Research and Development*, Vol. 39, No.1/2.)

reported in the literature, the only trench structure that is still in manufacturing in significant volume is the one developed by IBM. Compared to the structure shown in Figure 7.3c, where the signal charge is stored in the bulk n^+ silicon region outside of the trench, the IBM structure stores a charge on the n^+ polysilicon region inside the trench. Figure 7.6 summarizes how the IBM's DRAM structure has evolved from 4- to 256-Mb DRAMs¹⁰.

7.3 DRAM OPERATION PRINCIPLE

7.3.1 Charge Share Sensing

The way to distinguish whether the memory cell is storing “1” or “0” information (read operation) is by a circuit called a *sense amplifier*. A sense amplifier, which is described in detail in Section 7.4.3, is basically a cross-coupled latch that detects a small voltage difference between the detected node and the reference node. Since the sense amplifier circuit takes up a large area compared to a memory cell, only one sense amplifier is provided for a large number of memory cells, usually for 128 or 256 memory bits. All the memory cells connected to a common sense amplifier share a common bit line as illustrated in Figure 7.7. The memory access is such that only one word line is activated among all the memory cells connected to a common bit line.

The sensing proceeds as follows. Before a word line is activated to access a memory cell, the bit line pairs (BL and \overline{BL}) are precharged to a same voltage, usually $V_{DD}/2$. At this point, the memory cell capacitor is isolated from the bit line since the access transistor is off. Assume that the memory cell stores “1” state, or voltage, V_{DD} . After the word line is activated and the access transistor turns on, the memory cell capacitor becomes connected to the bit line. Denoting the memory cell capacitance as C_S and the total bit line capacitance as C_B , the resulting bit line voltage rises above $V_{DD}/2$ by

$$\Delta V = \frac{V_{DD}}{2} \cdot \frac{C_S}{(C_B + C_S)} \quad (7.2)$$

Since none of the memory cells are connected on \overline{BL} (none of the word lines are activated on \overline{BL} side), the voltage on \overline{BL} side stays at $V_{DD}/2$. Therefore, ΔV is the signal that the sense amplifier can detect to distinguish the memory cell as storing

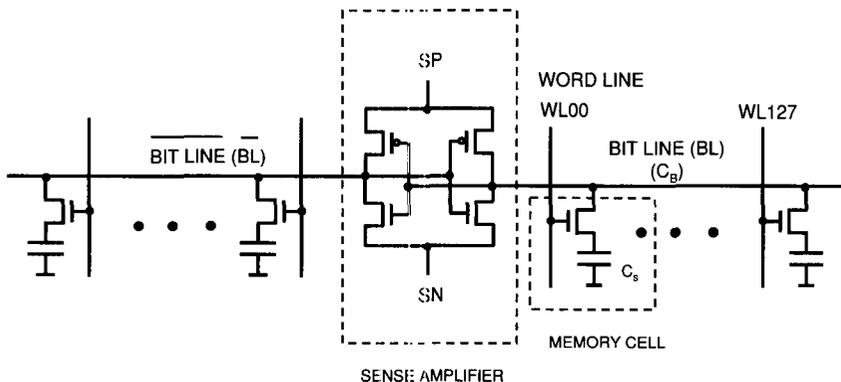


Figure 7.7 Schematic of a sense amplifier and a pair of bit lines. When one of the word lines on right side is selected, the bit line on the left side serves as a reference line.

“1.” Similarly, it can be shown that the voltage on the bit line decreases by the same amount, ΔV , if the memory cell stores “0.” Since C_B is much larger than C_S (usually at least 10 times), the voltage available to the sense amplifier for sensing is proportional to C_S/C_B and is usually a few tenths of a volt.

7.3.2 Refresh

The voltage stored on “1” memory cell slowly decays to “0” voltage (ground level) through various leakage mechanisms. There is no charge replenishing mechanism as in a static RAM (see Section 7.5). The only way that dynamic RAM memory cell does not lose the information is by periodically reading the data and rewriting the same data before the information is completely lost. This operation is called “refresh” and is an important feature of the DRAM. The cell refresh time requirement has increased from 2 ms for 64 K DRAM to as long as 64 ms for 16-Mb DRAM, and poses a tremendous challenge to reduce the cell-related leakage current for future DRAMs.

Several leakage mechanisms can affect the stored charge. This is summarized in Figure 7.8 and consists of the following:

1. Junction leakage
2. Pass transistor subthreshold leakage
3. Leakage through capacitor dielectric
4. Other parasitic leakage paths

Each mechanism is explained in details below.

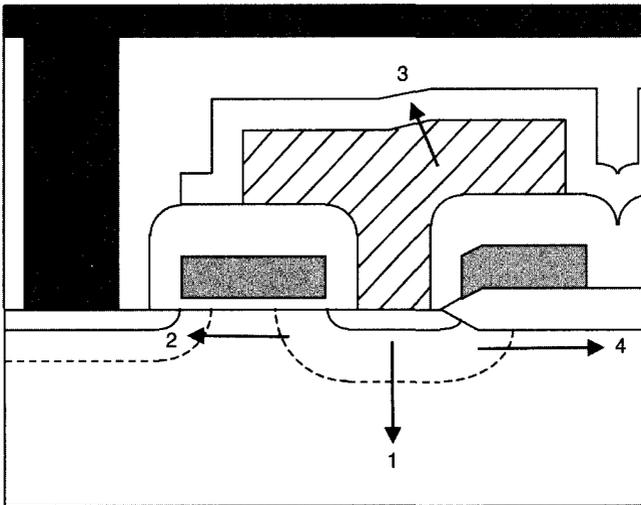


Figure 7.8 Various stored charge leakage mechanisms in a stack-type DRAM cell: (1) junction leakage; (2) pass transistor subthreshold leakage; (3) leakage through capacitor dielectric; (4) parasitic leakage path.

Junction Leakage

The pn junction in the DRAM memory cell is always reverse-biased. The leakage current through a reverse-biased pn junction can be expressed by the sum of the diffusion current and the generation-recombination current¹¹ as

$$J_R = \frac{qD_p p_{n0}}{L_p} + \frac{qD_n n_{p0}}{L_n} + \frac{qn_i W}{\tau_e} \quad (7.3)$$

where the first two terms represent the diffusion current, and the third term the generation-recombination current. Here, D_p and D_n are the diffusion coefficient of holes and electrons, respectively; L_p and L_n are the diffusion length of holes and electrons, respectively; p_{n0} and n_{p0} are the equilibrium hole density on the n side and the equilibrium electron density on the p side, respectively; n_i is the intrinsic carrier concentration; W is the depletion width; and τ_e is the effective lifetime in the depletion region. In most cases the generation-recombination current is dominant, but at zero or positive bias condition or at high temperatures, the diffusion current becomes more pronounced. In an ideal case, the diffusion current has an activation energy of E_g , the bandgap energy of silicon, whereas the generation-recombination current has an activation energy of $E_g/2$. Please also note that the depletion width, W , depends on the reverse-bias voltage, V , and the doping concentrations as

$$W = \sqrt{\frac{2\epsilon_s}{q} \cdot \left(\frac{N_A + N_D}{N_A N_D} \right) (V_{bi} + V)} \quad (7.4)$$

for a two-sides abrupt junction where N_A and N_D are the doping concentrations in p and n sides, respectively and V_{bi} is the built-in voltage.

Recently a new junction leakage mechanism that impacts the DRAM memory retention time has been reported.¹² This mechanism is based on the thermionic field emission of electron from the deep level in the depletion region. This leakage current increases with the doping concentration in the well because the electric field in the depletion region increases with the increasing well concentration and the thermal emission rate from the deep level also increases. Therefore, this current can be distinguished from the diffusion and generation-recombination current by its dependence on the well doping concentration since both diffusion and generation-recombination current decrease with the doping concentration as indicated in Eq. 7.4. This leakage has been shown to result in the tail of the distribution in the DRAM retention time.

Pass Transistor Subthreshold Leakage

Even when the word line voltage is low (at ground) and the pass transistor is in OFF mode, a small amount of current flows through the pass transistor below its threshold voltage. This is called a *subthreshold current*, and is given as

$$I_S = I_T \cdot \exp \frac{-qV_{th}}{mkT} \quad (7.5)$$

where I_T is the drain current at the threshold voltage, V_{th} is the threshold voltage, and m is a measure of the subthreshold slope and is related to the subthreshold slope parameter, S (mV/decade) by

$$m = \frac{0.001 \cdot qS}{kT \ln 10} \quad (7.6)$$

The transistor subthreshold current can be significant if the threshold voltage of the pass transistor is designed at too a low value.

Leakage through the Capacitor Dielectric

Although a capacitor dielectric is an insulating film, it conducts a small amount of current especially at high bias voltage. When the bias voltage exceeds the breakdown voltage of the dielectric film, the film becomes extremely conductive and the leakage current increases drastically. Below the breakdown voltage, the leakage mechanism through a nitride/oxide (NO) film is either Frenkel–Poole or Fowler–Nordheim. The Frenkel–Poole current is due to the field enhanced thermal excitation of trapped electrons into the conduction band and is given by¹¹

$$J = E \cdot \exp \left[\frac{-q(\phi_B - \sqrt{qE/\pi\epsilon_i})}{kT} \right] \quad (7.7)$$

where ϕ_B is the barrier height, E is the electric field, and ϵ_i is the high-frequency permittivity. On the other hand, the Fowler–Nordheim current is due to the electron tunneling enhanced under the electric field and is expressed as¹¹

$$J = E^2 \cdot \exp \left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar E} \right] \quad (7.8)$$

where m^* is the electron effective mass and \hbar is the reduced Planck's constant.

The dominant leakage mechanism in Ta_2O_5 and other high dielectric films such as $BaSrTiO_3$ (BST) and $PbZrTiO_3$ (PZT) has been found to be different from that of NO film and is due mostly to Schottky emission current at the interface between the dielectric and the electrode.¹³ The Schottky current is expressed as

$$J = AT^2 \cdot \exp \left[\frac{-q(\phi_B - \sqrt{qE/4\pi\epsilon_i})}{kT} \right] \quad (7.9)$$

where A is the effective Richardson constant, ϕ_B is the Schottky barrier height at the interface, and ϵ_i is the high-frequency permittivity. The Schottky current can be recognized by plotting J/T^2 versus $E^{1/2}$ on a logarithmic scale. On the other hand, the Frenkel–Poole current can be recognized by plotting J/E versus $E^{1/2}$ on a logarithmic scale.

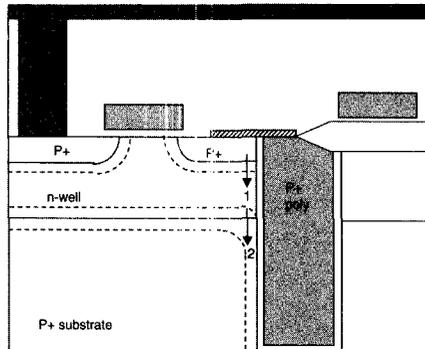


Figure 7.9 Two leakage mechanisms specific to a trench-type DRAM cell: (1) vertical parasitic FET; (2) trench-gated diode leakage.

Other Parasitic Leakage Path

In addition to these three mechanisms, there are some other parasitic leakage mechanisms some of which may be specific to the cell structure. For example, if the field oxide parasitic MOS threshold voltage is not sufficiently high, current can flow under the field oxide as indicated in Figure 7.8. In a trench capacitor structure, there are two specific leakage mechanisms as shown in Figure 7.9. One is a vertical parasitic MOS transistor that can leak the stored charge if the well doping concentration is not sufficiently high. Another is a trench-gated-diode leakage between the n-well and p⁺ substrate. Although this leakage does not cause signal loss, it can overload the n-well bias generator at high voltage and temperature during the burnin.¹⁰

7.3.3 Soft Errors

Soft errors are random, nonrecurring errors that are not due to the physical defects in a device (hard error). They are “temporary” in a sense that when new data are written, the memory operates normally, whereas a hard error causes the memory location to fail permanently. Soft errors can be caused by various mechanisms such as system noise, voltage marginality, pattern sensitivity, alpha particles, and cosmic rays. These errors can occur in both DRAMs and SRAMs. This section describes the soft errors in DRAMs caused by alpha particles and cosmic rays. The soft errors in SRAMs are described in Section 7.6.3.

Soft errors in DRAM memory cells are caused by the generation of unwanted (spurious) minority carriers that drift and are collected by the memory cell and destroy the stored information. There are two mechanisms by which minority carriers can be generated: alpha particles¹⁴ and cosmic rays.¹⁵ These mechanisms are compared in Figure 7.10.

Alpha Particles

Alpha particles are doubly charged helium nuclei that originate primarily from the decay of radioactive elements in the package or metallization materials. When the

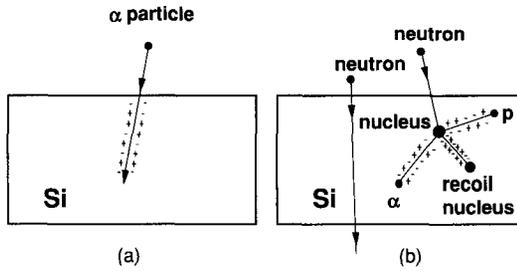


Figure 7.10 Comparison of electron–hole pair generation by (a) an alpha particle and (b) a neutron.

alpha particles penetrate the silicon substrate, they generate electron-hole pairs along the track as indicated in Figure 7.10. Typically, the track generates a charge density of 10–15 femtocoulombs (fC) per micrometer. Since a typical alpha particle with an energy of 3 MeV can penetrate the silicon substrate to a depth of 20 μm , a total charge of 200–300 fC is generated along the track.

Cosmic Rays

When primary cosmic ray particles (primarily composed of protons) enter the earth’s atmosphere, they collide with atmospheric atoms. These collision processes produce high-energy photons, electrons, protons, neutrons, muons, neutrinos, and so on. Most neutrons will pass through silicon with no interaction. However, a very small percentage of the high-energy neutrons interact by colliding with a silicon nucleus. This collision causes the silicon nucleus to recoil, which can generate several hundred femtocoulombs of charge in a distance of a few micrometers. The charge density can be as large as 100–150 fC/ μm , which is about 10 times larger than that for alpha particles.

Although the possibility of cosmic rays being a source of soft errors in DRAM was pointed out as early as 1979,¹⁵ it is only recently (at the time of writing) that some workers are beginning to conclude that cosmic ray neutrons may be a significant source of soft errors in DRAMs and SRAMs.

Figure 7.11 indicates the relative importance of neutron-induced and alpha-particle-induced soft errors described in a 1996 publication.¹⁶ The critical charge is

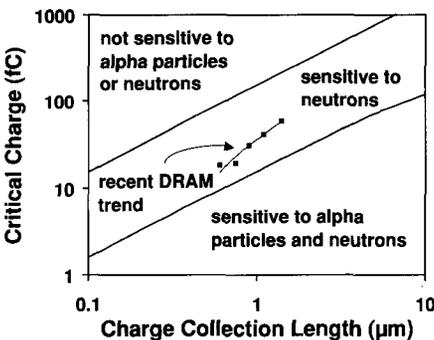


Figure 7.11 Regions of sensitivity to alpha particles and neutron induced recoils for DRAM storage nodes versus critical charge and charge collection length. (After McKee et al., Ref. 16, © IEEE, reprinted with permission.)

the amount of charge that must be collected by the memory storage node to cause a soft error, and the charge collection length is a depletion volume length of a storage node that contributes to the charge collection. This figure shows that for most modern DRAMs of 4 to 256 Mb density, the most dominant mechanism for soft errors is cosmic-ray neutrons rather than alpha particles.

7.4 DRAM CIRCUITS

7.4.1 Basic DRAM Structure and Operation

Figure 7.12 shows the structure of a typical DRAM chip. The solid arrows indicate the data flow and the dotted arrows indicate the flow of the control signals. A simplified read operation of a DRAM is described here. The signal timing is shown in Figure 7.13. Row and column addresses determine the X and Y coordinates of the memory cell that is being addressed, respectively, and are multiplexed in time to minimize the number of address pins. The row address is input first and is latched by the falling edge of the $\overline{\text{RAS}}$ (row address strobe) signal. Then the column address is latched in by the falling edge of the $\overline{\text{CAS}}$ (column address strobe) signal. The $\overline{\text{CAS}}$ signal is also used to enable the output buffer. The row address is decoded by the row decoder and selects a word line to be activated. The activated word line connects a memory cell to a bit line and establishes a small signal on the bit line. The sense amplifier detects this small signal. The column decoder connects one of many sense amplifiers to an input–output (I/O) line. This signal is amplified and sent to the data output buffer. As a result, the data output appears some time after $\overline{\text{CAS}}$ goes down. The time from $\overline{\text{CAS}}$ down to the data out is called “CAS access time,” while the time from $\overline{\text{RAS}}$ down to the data out is called “RAS access time.” During the write operation, the data flow from the I/O line to the outside pin is reversed. The row path is basically the same as in the read operation and selects a word line. In the meantime, the input data are sent to the I/O line through an input buffer. This signal on the I/O line is then used to drive the sense amplifier to write the data into the memory cell. The write operation is signaled by the $\overline{\text{WE}}$ (write enable) pin (not shown in Fig. 7.12) going low. For a read operation the $\overline{\text{WE}}$ pin is high.

One important thing to note is that when a word line is activated, all the sense amplifiers associated with the memory cells connected to that word line are activated regardless of which data are connected to the I/O line. Depending on the memory density and the exact way the memory array is divided into blocks, the number of sense amplifiers activated simultaneously is on the order of 1024. The data are available on these sense amplifiers if there is a way to extract these data. This will be important when we discuss the high-speed DRAM architecture in Section 7.4.5.

7.4.2 Memory Array

The simplest memory array architecture forms a memory cell wherever a word line and a bit line cross each other. Since a pair of bit lines (BL and $\overline{\text{BL}}$) must be

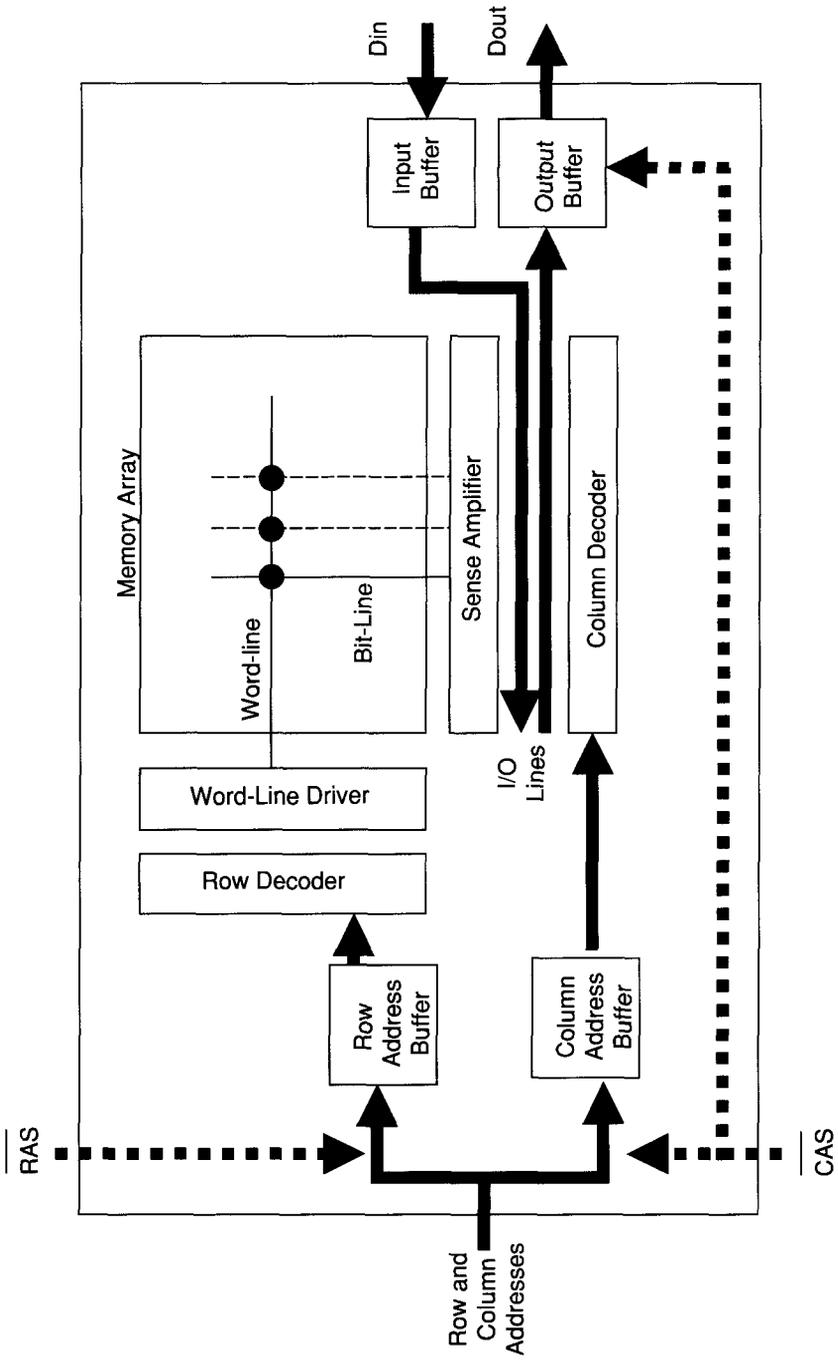


Figure 7.12 Basic structure of a typical DRAM chip.

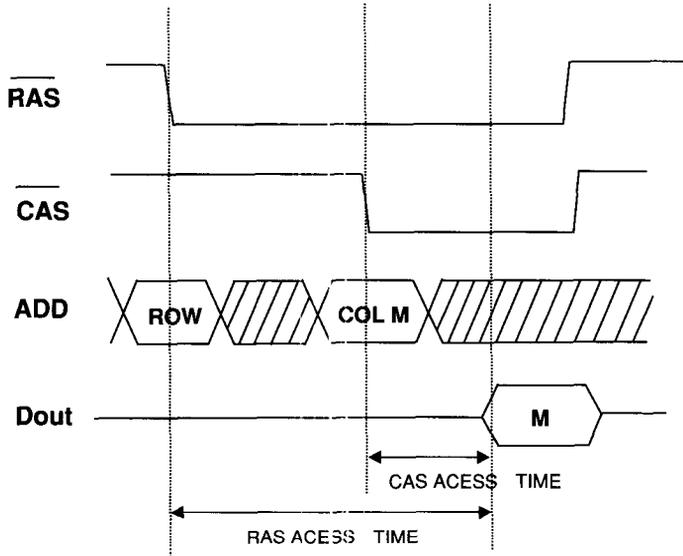


Figure 7.13 Signal timing diagram of the typical DRAM chip shown in Figure 7.13.

connected to a sense amplifier for data comparison, this simple architecture results in a configuration called an “open bit line” configuration shown in Figure 7.14a. When one of the word lines on the right side is selected, none of the word lines on the left side is selected (by decoding condition). The left-side bit line with no connected memory cells serves as a reference line for the right-side bit line. The open bit line configuration was used only in 16K and 64K DRAMs. There are two major problems with this configuration. One is that the sense-amplifier-circuit layout must fit into a very narrow pitch of a cell bit line (the sum of line width and spacing), which has proved extremely difficult because the memory cell size and pitch decrease with each successive generation of DRAMs. One solution to this problem is to use a “relaxed sense amplifier pitch open bit line” architecture¹⁷ as shown in Figure 7.14b. The second, and probably the most serious, problem with the open bit line architecture is the poor noise immunity of the open bit line configuration. The localized noise in the array couples into only one side of the bit line pairs and degrades the signal that the sense amplifier must detect.

The “folded bit line” configuration shown in Figure 7.14c was devised to avoid this noise problem. In this configuration, the BL and \overline{BL} extends side by side in the same direction from the sense amplifier. Any localized noise generated will couple into the both bit lines and be canceled by the differential operation of the sense amplifier. Therefore, the sense amplifier can detect a much smaller signal from the memory cell in the folded bit-line configuration. Since the development of 256K DRAMs, all DRAMs use this configuration.

Depending on how the memory cell is repeated in the array, the folded bit line architecture can result in different configurations as shown in Figure 7.15a,b. In this

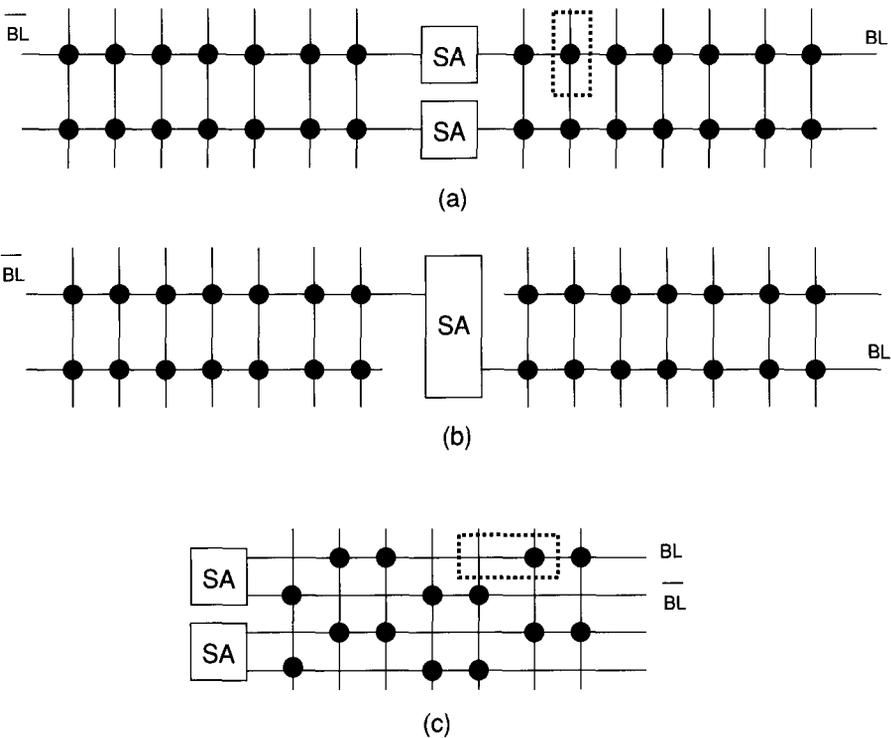


Figure 7.14 DRAM array configuration: (a) open bit line; (b) relaxed sense-amplifier pitch open bit line (after Inoue et al., Ref. 17); (c) folded bit line configuration. The dotted rectangle indicates a unit memory cell.

figure, the open circle indicates the bit line contact, the closed circle indicates the location of pass transistor, and the rectangle indicates the diffusion active region. The periodicity of the array is two bit line pitches in Figure 7.15a. In this arrangement, called “half-pitch array,” BL and \overline{BL} are next to each other. On the other hand, the periodicity of the array in Figure 7.15b is four bit line pitches. This arrangement is called a “quarter-pitch array.” BL and \overline{BL} are one bit line apart in this configuration.

The architecture that can be used is closely related to the actual memory cell structure. In the open bit line architecture, every time the word line crosses a bit line, the memory cell is formed. On the other hand, in the folded bit line architecture, one memory cell unit contains one passing word line (that does not form a memory cell) and a real word line as indicated by a dotted rectangle in Figure 7.15a. Therefore, the bit density in the folded bit line is seldom as high as in the open bit line architecture. The memory cell size is usually measured by the multiple of the minimum feature size, F . In the open bit line architecture, the minimum possible memory cell area is $2F \times 2F = 4F^2$ since the word line and bit line pitches are both $2F$ (F for a line and F for a space). In the folded bit line architecture, due to the

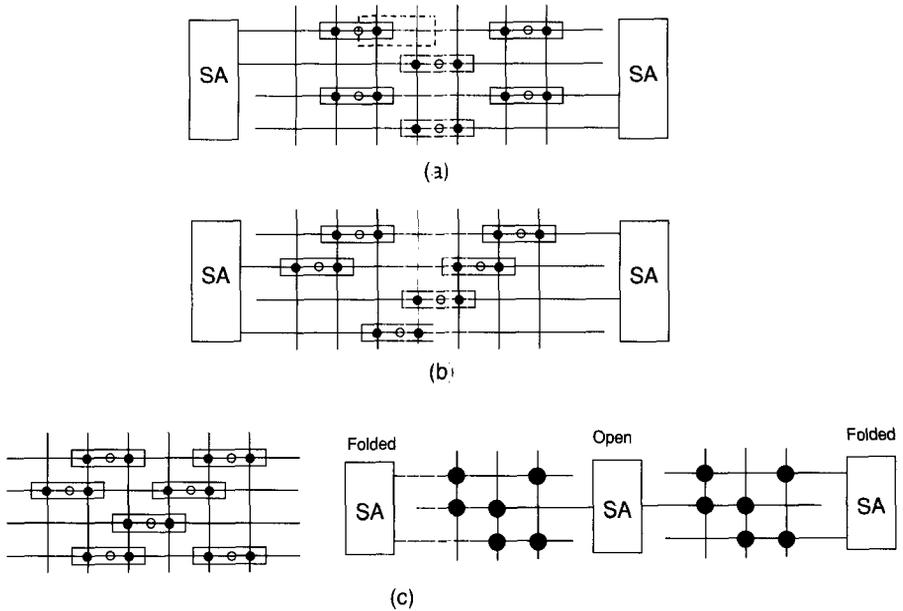


Figure 7.15 DRAM array architecture; (a) half-pitch array architecture; (b) quarter-pitch array architecture; (c) $6F^2$ cell array architecture and sensing scheme. (After Takashima et al., Ref. 18.)

existence of one passing word line, the memory cell dimension is $4F$ in the direction parallel to the bit-line. Therefore, the minimum possible memory cell size is $2F \times 4F = 8F^2$.

Since the minimum feature size, F , is usually determined by the technology (mainly by lithography), this $8F^2$ limit places a severe barrier to the scaling of the future DRAM memory cell size and chip size. Therefore, various architecture approaches are being studied to achieve smaller than $8F^2$ limit without sacrificing the benefit of folded bit line sensing. One such example is shown in Figure 7.15c.¹⁸ In this memory cell placement, there is only one passing word line between two active word lines as opposed to two passing word lines in a conventional folded bit line configuration. The memory cell size achieved is therefore, $6F^2$. The sensing of this array requires the mixing of folded and open bit line configurations as illustrated in the figure. In the actual implementation, a switch (transistor) is required between the sense amplifier and the bit lines to change the bit-line connection depending on which word line is selected.¹⁸ Another method to utilize $6F^2$ in a folded bit line configuration is to run the bit lines diagonally rather than horizontally.¹⁹ A circuit technique has also been proposed to utilize the $4F^2$ memory cell (cross-point cell), which conventionally requires an open bit line sensing, in a folded bit line sensing configuration by using a switch similar to the $6F^2$ folded/open sensing approach.²⁰ The use of switches, however, requires an additional timing and better control of the timing, which will be increasing difficult to get as the DRAM access time decreases

for future DRAMs. It is expected in the future that many other different approaches will be explored in order to break the $8F^2$ limit.

7.4.3 Sense Amplifier Operation

As explained in Section 7.3.1, a sense amplifier is used to detect a small-signal difference, typically several hundred millivolts to less than 100 mV, between the active bit line and a reference bit line. Figure 7.16 shows a schematic of a representative sense amplifier and operation waveforms in a $V_{DD}/2$ precharge scheme.²¹ Before the word line (WL) is activated, the bit line pairs are precharged to $V_{DD}/2$ through the precharge circuit. As soon as the WL is activated, the voltage of a bit line with a memory cell connected rises slightly according to Eq. 7.2 if the memory cell stores a “1” state, and drops by the same amount if it stores a “0”. After a short time, SN (sense NMOS) signal is brought low to initiate the sensing. As shown in the figure, both bit lines will fall together for a while as a result of the coupling effect until SN becomes V_{TN} below the high side of the bit lines, where V_{TN} is the threshold voltage of the pull-down transistor, Q_N . At this point, the transistor \overline{Q}_N turns on and pulls down the \overline{BL} side further. This turns off the transistor Q_N further and keeps the BL side from falling further. The \overline{BL} side continues to fall with SN. As soon as sufficient voltage difference is developed between BL and \overline{BL} , SP

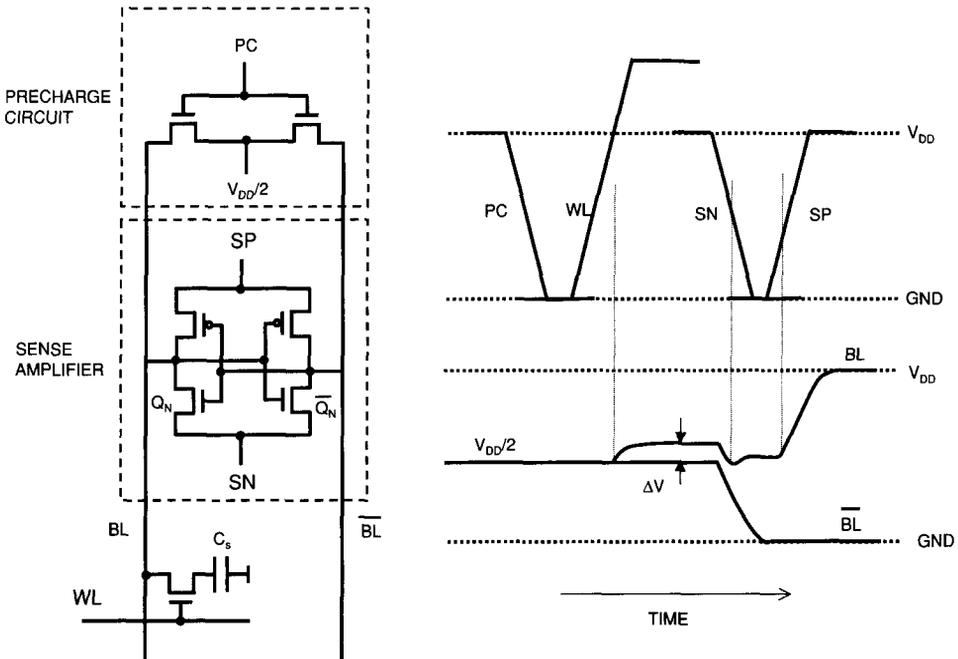


Figure 7.16 Sense amplifier schematic and operation waveforms in a $V_{DD}/2$ precharge scheme.

(sense PMOS) signal comes on to restore the high side to full V_{DD} . Since the WL is still high, this “1” state is written back into the cell. The voltage on the storage node at this point is actually $V_{DD} - V_T$, where V_T is the threshold voltage of the pass transistor, since the WL voltage is at V_{DD} . To restore the full voltage to the memory cell, the WL voltage must be well above V_{DD} to compensate for the V_T and the back-bias effect as explained in the next section. In some cases, both SN and SP are activated simultaneously at the start of the sensing to allow simultaneous sensing by NMOS and PMOS latches in the sense amplifier.

The sensitivity of the sense amplifier, or the minimum voltage difference that the sense amplifier can detect, is dependent on the rate at which the SN node is pulled down. The slower the SN is pulled down, the smaller the signal the sense amplifier can detect. The relationship between the time it takes for SN to go down, or the latch time, T , and the sensitivity of the sense amplifier, v , is given by²²

$$T = 2C_B \cdot \frac{(V_{TN} - fv)}{\beta f^2 V_{TN} v} \quad (7.10)$$

where C_B is the total bit line capacitance, V_{TN} is the threshold voltage of Q_N , $f = C_B/(C_B + C_G)$, where C_G is the gate oxide capacitance of Q_N , and β is the gain of Q_N . In most cases, $C_B \gg C_G$, $f \sim 1$, and $V_{TN} \gg v$. Therefore

$$T = \frac{2C_B}{\beta f^2 v} \quad (7.11)$$

In other words, the sensitivity, v , is inversely proportional to the latch time, T .

7.4.4 Word Line Boosting

In addition to the sensing of a small bit line signal, another important function of the sense amplifier is writing back the full signal into the memory cell. As indicated in Figure 7.16, the sense amplifier will latch completely after a sufficient time, and the BL and $\overline{\text{BL}}$ approach V_{DD} and 0 V, respectively, for the case of data “high.” This condition, however, is not sufficient to write back a full V_{DD} signal to the memory cell capacitor due to the threshold voltage drop through the pass transistor. The word line voltage must be “boosted” above the V_{DD} level in order to write a full V_{DD} signal to the memory cell. This situation is illustrated in Figure 7.17. Since the source of the pass transistor (bit line) is at V_{DD} , the pass transistor sees an effective back bias voltage of $-(|V_{BB}| + V_{DD})$, where V_{BB} is the normal back bias with the source at ground. For a DRAM chip with $V_{DD} = 2.5$ V, V_{BB} is usually between -1 and -1.5 V. Therefore, the effective back bias that the pass transistor sees can be as large as -4 V. The threshold voltage of the pass transistor with this back bias is much higher than the threshold voltage with normal $V_{BB} = -1.5$ V, and can be around 1.2 V or higher. The word line voltage V_{WL} must be above the source voltage, which is V_{DD} in this case, by at least this threshold voltage. Therefore, the required

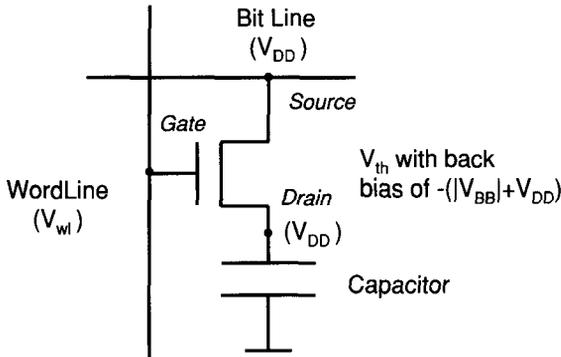


Figure 7.17 Voltage condition for word line boosting.

word line voltage level is $V_{WL} = V_{DD} + V_{th}$, which is at least 3.7 V (2.5 V + 1.2 V). A general rule of thumb is that the word line boost voltage must be at least 1.5 times the power supply voltage, V_{DD} .

7.4.5 High-Speed DRAM Architectures

As the operating speed of a microprocessor improves with technological advances (mostly by scaling), the speed of DRAM access must also improve in pace with the microprocessor since the DRAM as a main memory must supply the data to the microprocessor. However, as shown in Figure 7.18, the gap between the microprocessor clock speed and the DRAM access speed (as measured by the inverse of the RAS access cycle time) continues to widen. This is due mostly to

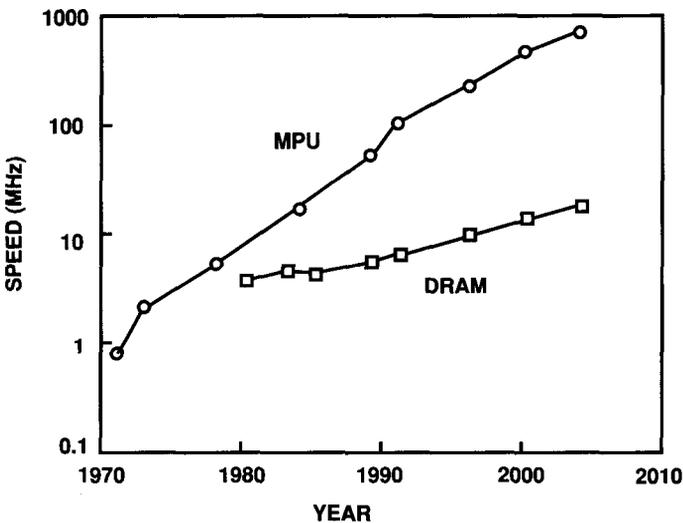


Figure 7.18 Speed trend of microprocessor (MPU) and DRAM.

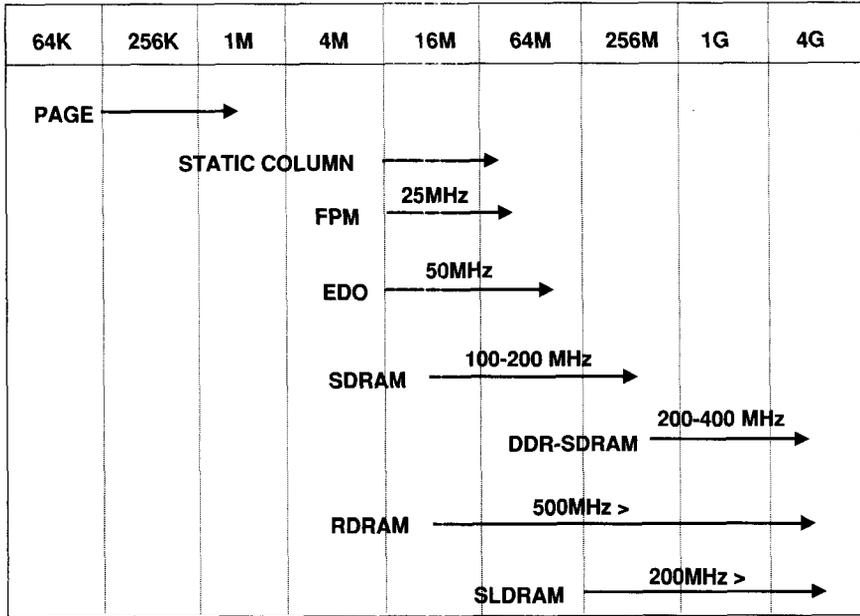


Figure 7.19 Trend of DRAM access mode change.

factors such as the address multiplexing time, the internal word line and sensing delay, and the limited data widths. Recently, there has been a tremendous interest and effort in improving the access time and bandwidth of DRAM access. Various new DRAM access modes and architectures have emerged as a result. Figure 7.19 summarizes the history of DRAM access mode changes. Some of these are explained in this section.

Fast Page Mode (FPM)

As explained in Section 7.4.1, once the row address is given and a word line activated, all the sense amplifiers associated with that word line have the data latched in. Therefore, only a column address is necessary to select the data out of these already activated sense amplifiers. This is the “fast page mode” (FPM), whose timing diagram is shown in Figure 7.20. After the first bit is accessed by a normal access using \overline{RAS} and \overline{CAS} , the subsequent access is done by supplying only the column address by toggling the \overline{CAS} signal.

Extended Data Out (EDO)

In the FPM mode, since the \overline{CAS} signal is also used as an output enable signal, once \overline{CAS} goes high, the output becomes invalid (undetermined state called tristate). This makes it difficult for the microprocessor to grab the data within a short period. The “extended data out” (EDO) DRAM makes the data available for a longer period. The output data stays valid even when \overline{CAS} goes up and stays valid until \overline{CAS} goes down the next time (see Fig. 7.20). Because of this added timing margin, the EDO

FAST PAGE/EDO MODE

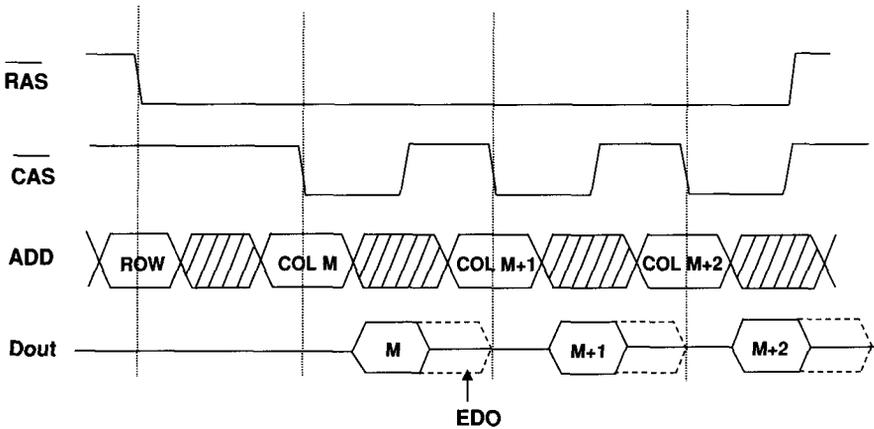


Figure 7.20 Timing diagram for fast page and extended data out (EDO) access modes.

DRAM is capable of tightening the entire timing and allowing a higher data rate than the FPM mode. In terms of circuit design, EDO design is only a small modification from the FPM design, but the data rate improvement can be as large as 50%.

Synchronous DRAM (SDRAM)

Instead of asynchronous operation of a conventional DRAM relying on the relative timings of \overline{RAS} and \overline{CAS} , a synchronous DRAM has a clock signal with which all operations are synchronized. Its timing diagram is shown in Figure 7.21. \overline{RAS} and \overline{CAS} signals behave the same way as in a conventional DRAM, but the row and column addresses are latched with the falling edges of the clock signal. The first data becomes available after a few clock cycles. Depending on the operating mode, a sequence of data more than 4 bits long (burst length) is available successively. The time from \overline{RAS} to “data out” is called a “row latency,” the time from \overline{CAS} to data out a “CAS latency,” and both are measured in the unit of clock cycles. This synchronization to the clock signal is advantageous to the microprocessor since it knows exactly when to expect the data output and can perform other operations without waiting for the data.

The diagram in Figure 7.21 shows only the falling edge of the clock to latch the addresses and data out. The latest development uses both edges (rising and falling) of the clock to essentially double the data rate. This is called DDR, or “double data rate,” and is expected to be a standard mode for future SDRAMs. Since it requires additional circuits to control and synchronize the tighter timing, the chip size will be larger compared to a standard-data-rate SDRAM.

Rambus DRAM (RDRAM)

In a conventional DRAM, one way to increase the bandwidth is to increase the number of bits on the data output. This has indeed been the trend with the

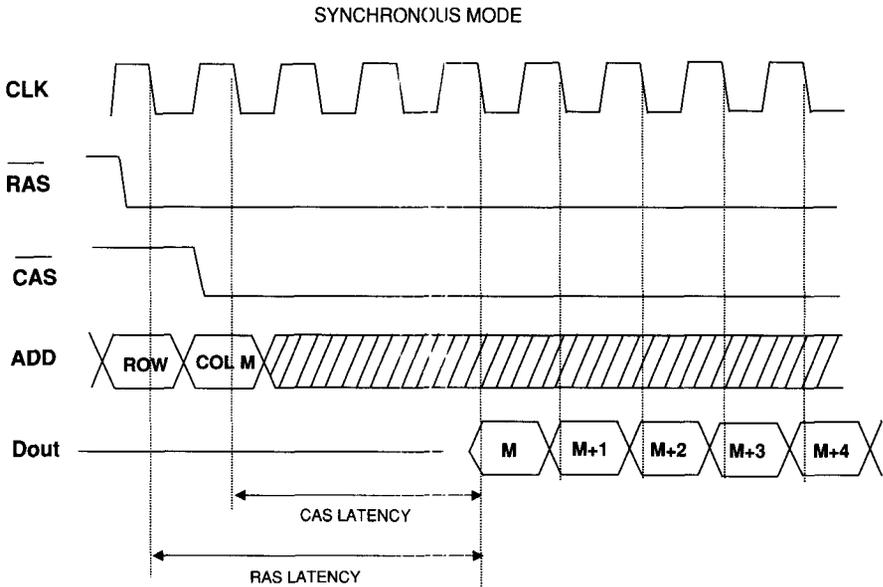


Figure 7.21 Timing diagram for synchronous access mode.

conventional DRAMs with the memory bus width increasing from $\times 1$ (by one, 1 bit wide), $\times 4$, $\times 8$, $\times 16$ to even $\times 32$ in most recent DRAMs. Rambus DRAM (RDRAM) and synchronous link DRAM (SLDRAM) relies on a different approach to increase the memory bandwidth. The approach is to keep the bus width to a small number (8 or 16 bits wide), but operates this bus at extremely high speed. It also relies on a “protocol” approach where the addresses and control signals for different modes are provided in a fixed sequential manner.

Figure 7.22 illustrates how several RDRAMs can be connected to a master processor with a common bus.²³ In this example, 32 units of RDRAMs are connected. The bus is 9 bits wide, and there are several other control lines and power supply lines. A read operation starts with the master processor placing an address on the bus. One of the 32 RDRAMs will have the matched address. This RDRAM will send an “okay” signal back to the processor indicating that the address has been matched. After several clock cycles, this RDRAM places the data from that address on the bus, which the processor can retrieve.

The protocol corresponding to this read operation is shown in Figure 7.23 as an example.²³ The reference clock speed is 250 MHz, but the data are latched at both falling and rising edges of the clock, yielding an effective data rate of 500 Mb/s per pin. The first packet, the read data request: “RDreq” consists of 6 bus cycles or 12 ns. The “Op” field indicates the type of operation, such as read memory or write memory. The “Adr” field indicates the device to access and the starting address. The “Count” field indicates the number of quadbytes to be transferred in the data packet. The “Rsrv” fields are reserved for the future. In this figure, Adr[9:2] means the 8 bits from the second bit to the 9th bit in the “Adr” field. Count[6,4,2] means three

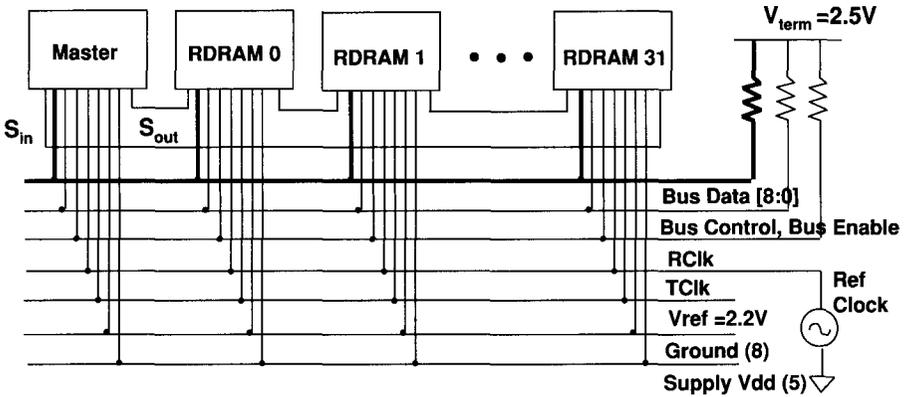


Figure 7.22 Simplified view of a Rambus system. (After Kushiyama et al., Ref. 23.) The number in parentheses indicates the number of pins. RClk and TClk are the *receive clock* (clock from master) and the *transmit clock* (clock to master), respectively.

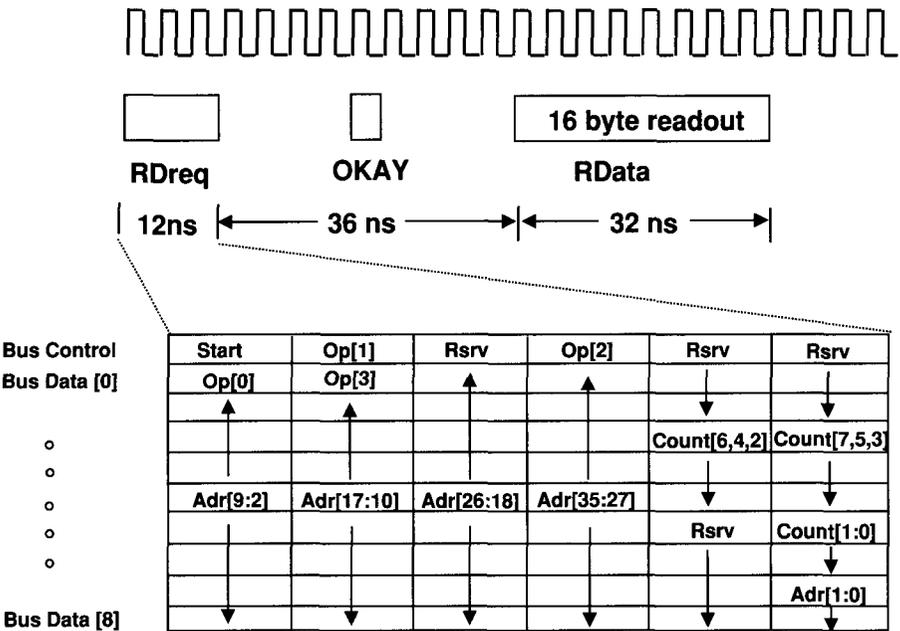


Figure 7.23 Example of a Rambus protocol packet format. (After Kushiyama et al., Ref. 23.)

bits, 2nd, 4th, and 6th bits in the “Count” field. Op[0] is the 0th bit in the “Op” field, and so on. This is an example of how the address and operational information is contained in the bus data packet.

The latest development further increases the bandwidth of the next generation of Rambus, called “direct Rambus” or “direct RDRAM.” There are significant

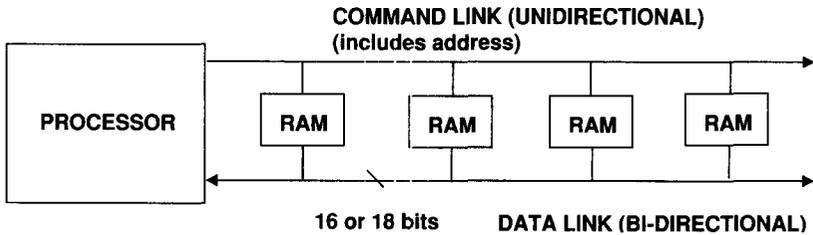


Figure 7.24 Simplified view of a SyncLink DRAM system.

differences in direct RDRAM from the first generation of RDRAM described above. One is that the data lines are separate from the address and control lines. Another difference is that the bus width has been increased from 8 to 16. This wider bus and higher frequency operation of the channel up to 400 MHz (800 MHz data rate) allow a peak bandwidth as high as 1.6 GB/s.

Synchronous Link DRAM (SLDRAM)

SLDRAM is an IEEE standard high-speed DRAM specification based on a protocol similar to the Rambus. It has command and address lines separate from the data line as shown in Figure 7.24 similar to the direct RDRAM but different from the first generation of RDRAM. The bus width is 16 as in the direct RDRAM. There are no commercial SLDRAMs available yet, but several DRAM manufacturers are working in a consortium to commercialize SLDRAMs.

Table 7.2 summarizes the features and speed of these new high speed DRAMs and a conventional synchronous DRAM. Figure 7.25 shows the peak bandwidth possible by these new devices as a function of year. It is expected that Direct RDRAM and SLDRAM will realize the peak bandwidth above 1 GB/s in the near future. To achieve the bandwidth well above 2 GB/s, however, a new scheme will be necessary. This may involve the integration of DRAM and logic (microprocessor) on the same chip. The DRAM/logic integration would allow much wider bus between them (256–1024, e.g.) which is not possible between separate chips because of pin capacitance, pin inductance, and power consumption.

7.5 SRAM MEMORY CELL STRUCTURE

7.5.1 Memory Cell Concept

An SRAM cell is a bistable transistor flip-flop, or two inverters connected back to back. Depending on its load type, there are four representative MOS SRAM cells:

1. Depletion load
2. Resistor load
3. Full CMOS load
4. TFT (thin-film transistor) load

TABLE 7.2 Comparison of High-Speed DRAMs

Items	SDRAM	DDR SDRAM	Rambus	Direct RDRAM	SLDRAM
Data addressing	Multiplexed address pins	Multiplexed address pins	Multiplexed address, control and data pins	Multiplexed address and control pins	Multiplexed address and control pins
Protocol method	No	No	Yes	Yes	Yes
Number of I/Os per channel	16	16	8	16	16
System clock frequency (MHz)	100	100	250	400	200
Data transfer frequency (MHz)	100	200	500	800	400
Total peak bandwidth (GB/s)	0.2	0.4	0.5	1.6	0.8

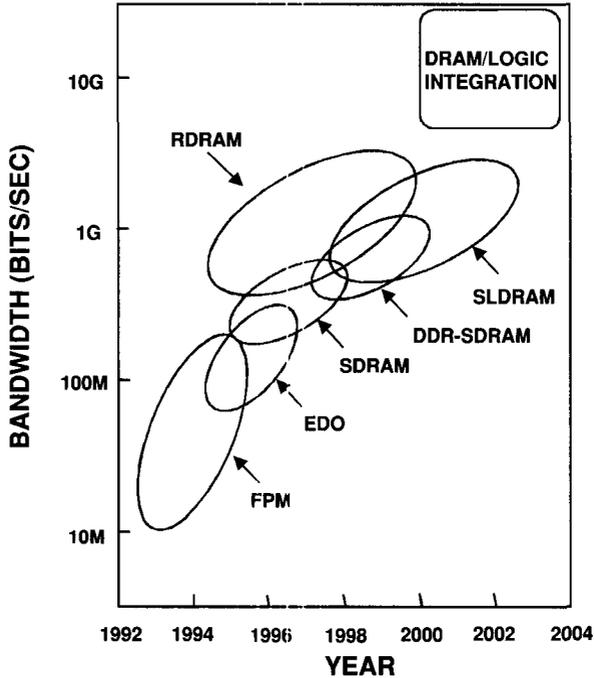


Figure 7.25 Roadmap of high-speed DRAM bandwidth and mode.

These four types are illustrated in Figure 7.26. Table 7.3 compares the advantages and disadvantages of each load type. The depletion-load SRAM cell is the oldest type and uses a depletion-mode (negative threshold voltage) n-MOS transistor as a load. This type was used before CMOS technology became predominant. The resistor-load (R-load) cell uses a high-resistance polysilicon resistor as a load. Since the polysilicon layer can be fabricated over the driver n-MOS transistors, it can achieve a very small cell size and high density. However, as the density of SRAMs increases, the resistance of the polysilicon load had to be increased significantly to proportionately decrease the standby current per cell. When the required polysilicon resistivity became prohibitively high, it was replaced by the TFT-load type. The full CMOS load uses a bulk p-MOS transistor as a load. It is compatible with the standard CMOS process with no additional polysilicon layer required. However, the memory cell size tends to be large compared to the R-load. The standby current can be very low since the pMOS load transistor can be turned off completely except for a very small subthreshold current. The TFT-load cell is an evolution of the R-load cell and uses the polysilicon layer as a channel of the p-MOS transistor. Since the mobility in the polysilicon layer is much lower than in the bulk silicon, the performance of the polysilicon p-MOS is much inferior to the bulk p-MOS, but is sufficient as a cell load. The TFT-load cell combines the low-power advantage of the full CMOS load cell and the high-density advantage of the R-load cell. As the power

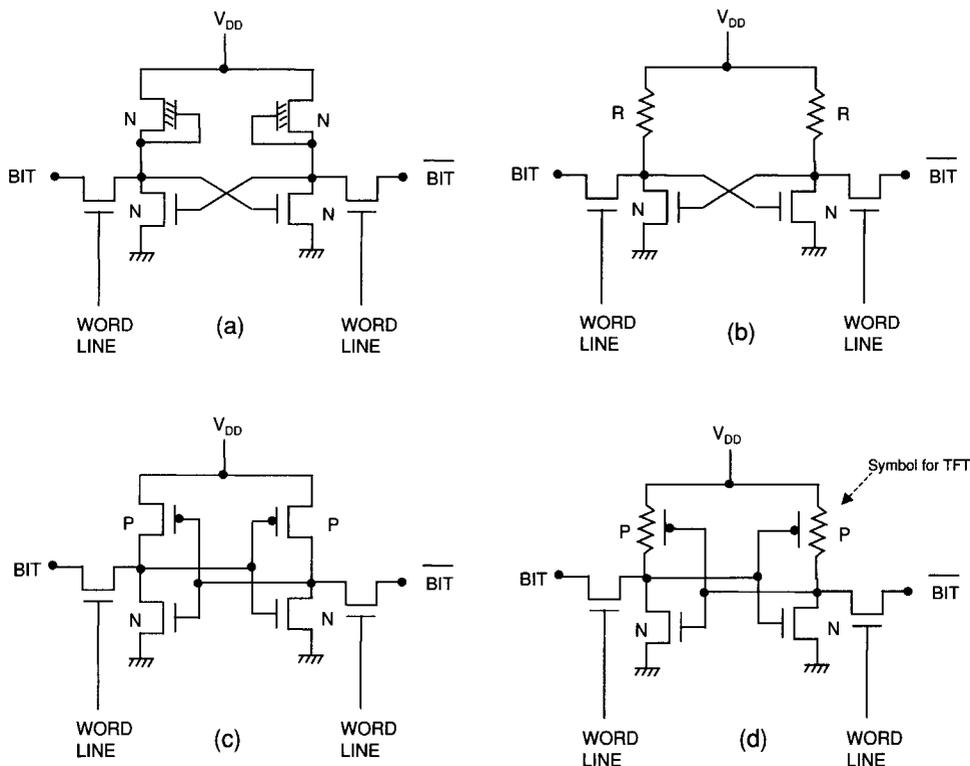


Figure 7.26 Four types of SRAM cell: (a) depletion-load SRAM cell; (b) resistor-load SRAM cell; (c) full CMOS SRAM cell; (d) TFT-load SRAM cell.

TABLE 7.3 Comparison of SRAM Memory Cells

	Depletion Load	Resistor Load	Full CMOS	TFT Load
Density	Medium	High	Low	High
Standby current	High	Medium	Low	Low
Cell stability	High	Low	High	Medium

supply voltage is reduced, however, the stability of the TFT-load cell still is not as good as that of a full CMOS cell and is gradually being replaced by a full CMOS cell.

Figure 7.27 shows how the SRAM memory cell sizes for full CMOS, R-load, and TFT-load types were reduced during 1988–1998. The reduction factor is about $0.4 \times$ for 3 years (one technology generation), which is similar to the reduction factor for the DRAM cell, mainly due to the scaling down of minimum feature size. As expected, the full CMOS type has the larger memory cell size compared to the R-load or TFT-load cell type.

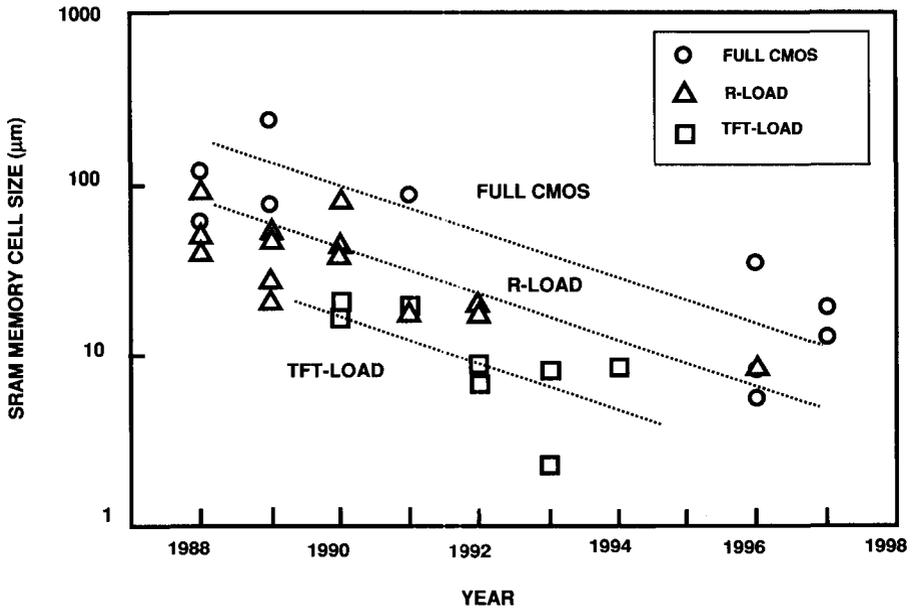


Figure 7.27 SRAM memory cell size trends. (Source *ISSCC Proceedings*.)

Figure 7.28 shows an example of an SRAM cell structure and its layout²⁴ with a TFT load. In this structure, a p-channel TFT MOSFET is fabricated on top of an n⁺ polysilicon gate that is shared between the top p-MOS TFT and the bottom bulk n-MOS. The drain side of the TFT p⁺ is offset from the gate to reduce the drain-induced leakage current. The p⁺ drain is connected to the common node through an n⁺ polysilicon layer. This connection results in a pn diode at the common node, but the SRAM operation is not seriously affected since the diode is always forward-biased under normal operating condition.

7.6 SRAM OPERATION PRINCIPLE

7.6.1 Sense Operation

The sense operation of an SRAM cell is illustrated in Figure 7.29. Before the sense operation, both sides of the bit lines are precharged high to $V_{DD} - V_T$, where V_T is the threshold voltage of the nMOS pullup transistors, Q_{L1} and Q_{L2} . Assume that the left-side node stores high (H) and the right-side node stores low (L). Then, Q_{P1} and Q_{D2} are ON and Q_{P2} and Q_{D1} are OFF. When the word line becomes high, the access transistor Q_{A2} turns on and a current path develops through Q_{L2} , Q_{A2} , and Q_{D2} , which draws a current, ΔI , as indicated by a dotted arrow. A sense amplifier connected to the end of bit line pairs can detect this differential current, and determine that the BIT side of the storage node is high and the $\overline{\text{BIT}}$ side is low.

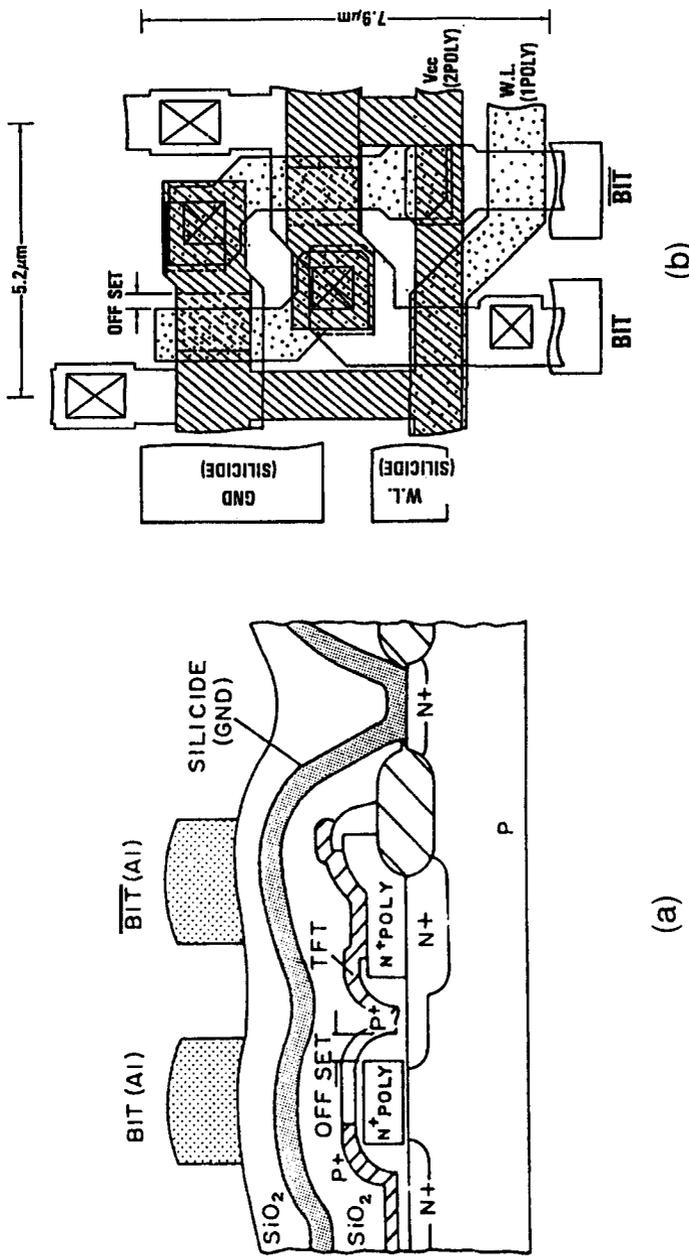


Figure 7.28 (a) Cross section and (b) layout of an SRAM memory cell structure with p-channel TFT load. (After Ando et al., Ref. 24, © IEEE, reprinted with permission.)

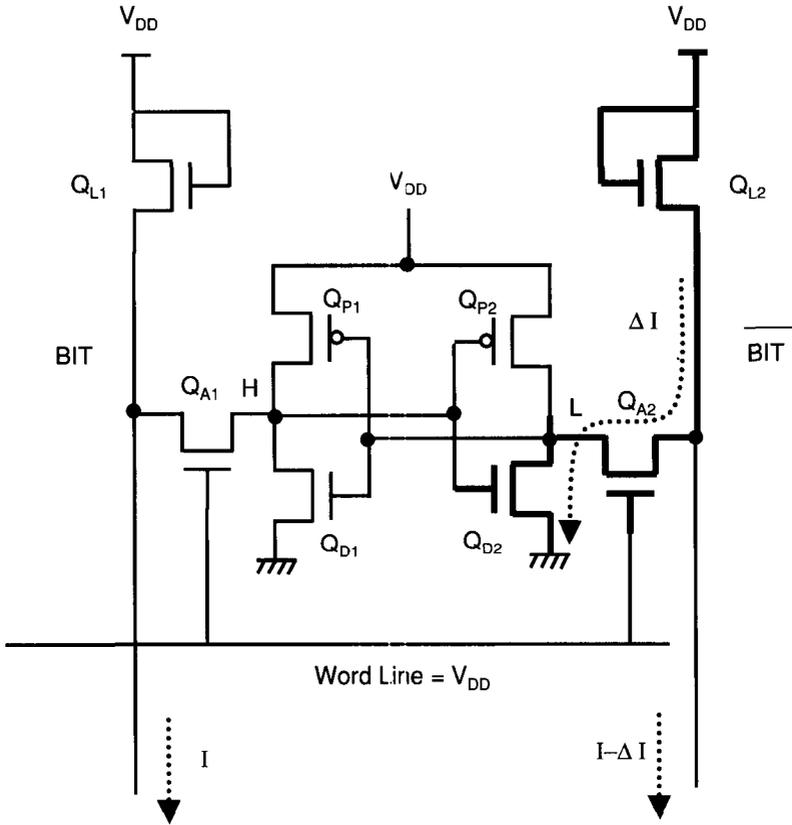


Figure 7.29 Circuit schematic for the sense operation of an SRAM cell.

7.6.2 Cell Stability Analysis

The SRAM cell stability determines the soft-error rate and the sensitivity of the memory cell to variations in process and operating conditions. One important parameter for the stability of the SRAM cell is the ratio between β (gain where $\beta = \mu C_{ox} W/L$) of the driver transistor and β of the access transistor. This ratio is usually called the “cell ratio” or “ β ratio” of the SRAM cell. During the read access the driver and access transistors are connected in series as shown in Figure 7.29. The voltage at the node between the read transistor and access transistor is determined by the ratio of β of the two transistors. If the β of the driver transistor is too small, it can bring the cross-coupled low side too high and can flip the cell. Therefore, higher β ratio results in better stability. However, it also results in a larger cell size. There is a tradeoff between the cell area and the stability of the cell.

The static-noise margin (SNM) of an SRAM cell can be analyzed graphically by drawing and mirroring the inverter characteristics and finding the maximum possible square between them. This is illustrated in Figure 7.30 for the full CMOS cell as well as the R-load cell.²⁵ In this simulation, the β ratio is equal to 2 for the full CMOS cell

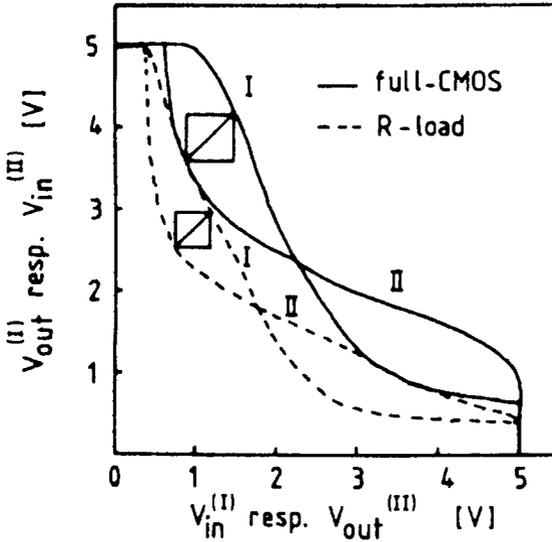


Figure 7.30 Comparison of static-noise margin (SNM) for resistor-load (R-load) and full CMOS SRAM memory cells. (After Seevinck et al., Ref. 25, © IEEE, reprinted with permission.)

and 3.5 for the R-load cell. As shown in the figure, the full CMOS cell has a much larger SNM compared to the R-load cell in spite of the larger value of cell ratio for the R-load cell. Another important difference between the full CMOS cell and R-load cell is the dependence of SNM on the power supply voltage. As seen in Figure 7.31, the SNM for the R-load cell (and also the TFT load cell) rapidly degrades at lower power supply voltage, whereas the SNM for the full CMOS cell stays fairly high even at a low power supply voltage. As the power supply voltage for SRAMs decreases below 1.8 V, especially for portable equipment applications, the full CMOS cell is becoming much more attractive.²⁶

7.6.3 Soft-Error Requirement

Alpha Particles

When an alpha particle hits the high side of the SRAM memory cell, the collected charge lowers the potential of the high node and can flip the memory cell, thereby causing a soft error. While the soft error in DRAM is determined only by the total amount of charge collected by the cell node, the situation is slightly different in SRAM because the cell load element (p-MOS, poly resistor, or TFT) can supply a current to the node that was hit in an attempt to restore the original cell state.

An alpha-particle-induced charge contains two components based on their time dependence.²⁷ One is a prompt charge, or funneling charge, which is collected by the depletion region in several hundred picoseconds. Another component is a diffusion

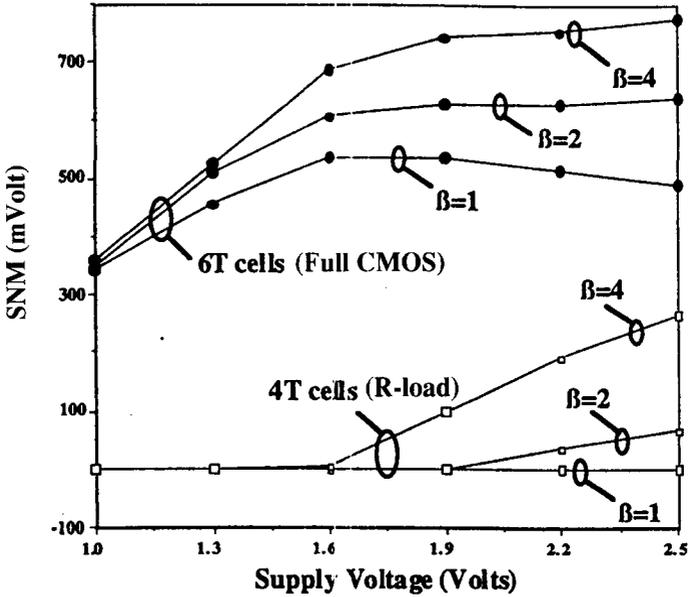


Figure 7.31 Power supply voltage dependence of static-noise margin for resistor-load (R-load) and full CMOS SRAM memory cells with three different β ratios. (After Lage et al., Ref. 26, © IEEE, reprinted with permission.)

component where the charge outside of the depletion region is collected slowly by diffusion. This collection occurs in a nano second or longer. Since the p-MOS load transistor in a full CMOS SRAM cell can supply around $30 \mu\text{A}$ when fully ON and the node capacitance is on the order of 5 fF , the charging time constant is $5 \text{ fF} \times 2.5 \text{ V} / 30 \mu\text{A} = 0.4 \text{ ns}$. This charging time is shorter than the diffusion charge collection time and can compete with the diffusion charge collection mechanism. On the other hand, the TFT or poly resistor load can supply only several orders of magnitude lower current and is too slow to compete with the diffusion collection mechanism. Therefore, the soft error of TFT or poly resistor memory cell is usually much worse than that of the full CMOS memory cell.

The alpha particle sensitivity of an SRAM memory cell can be simulated with the SPICE circuit simulator by modeling the alpha-particle-induced charge as a time-dependent current source connected to the node hit.²⁸ It is important to include the correct time dependence of the current source considering the prompt charge and the diffusion charge components.

Cosmic Rays

Similar to the DRAM, cosmic rays, mostly neutrons, can cause soft errors in SRAMs. It has been concluded recently that cosmic-ray-induced errors are more dominant in advanced SRAMs than are alpha-particle-induced errors.²⁹ This is the same conclusion drawn for DRAMs as mentioned earlier (Fig. 7.11).

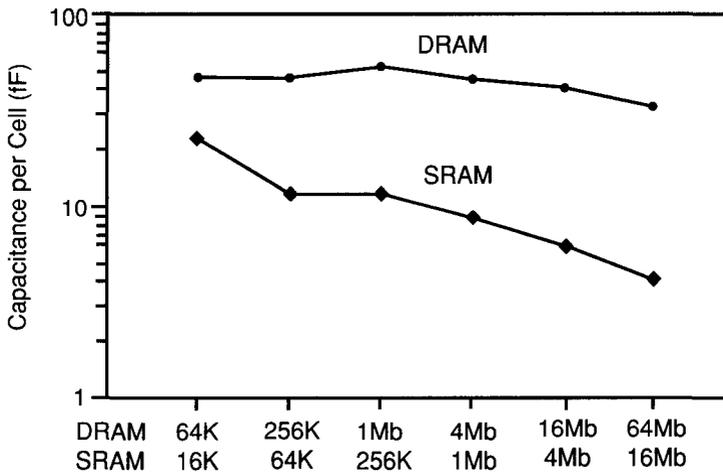


Figure 7.32 Trend of capacitance per memory cell for SRAMs and DRAMs. (After Lage et al., Ref. 29, © IEEE, reprinted with permission.)

Comparison of DRAM and SRAM Soft Errors

The capacitance of a sensitive node (cross-coupled node) in an SRAM memory cell is usually much smaller than the cell capacitance in a DRAM memory cell. Figure 7.32 compares the typical node capacitance for several generations of DRAMs and SRAMs.²⁸ While the DRAM cell capacitance has been maintained fairly constant because of the sensing requirement, the SRAM node capacitance has decreased rapidly. Although the active p-MOS load compensates for this smaller node capacitance somewhat, this still results in a smaller critical charge and higher number of soft errors for SRAMs as shown in Figure 7.33. It is expected that an additional capacitance must be added to the SRAM sensitive node for future SRAM generations to maintain an acceptable soft-error rate.

7.7 SRAM CIRCUITS

7.7.1 Basic SRAM Structure

The structure of an SRAM chip is basically similar to the DRAM chip structure shown in Figure 7.12 with a few differences. One difference is that in the SRAM chip the row and column addresses are not multiplexed in time, but are input simultaneously from separate pins. Another difference is that since the SRAM memory cell is static and the readout is not destructive, it is not necessary to have one sense amplifier for each bit line. Usually many bit lines (or data lines) are connected to a sense amplifier through the column select switch and a pair of I/O lines so that only one column is connected to the I/O line and the sense amplifier at a time.

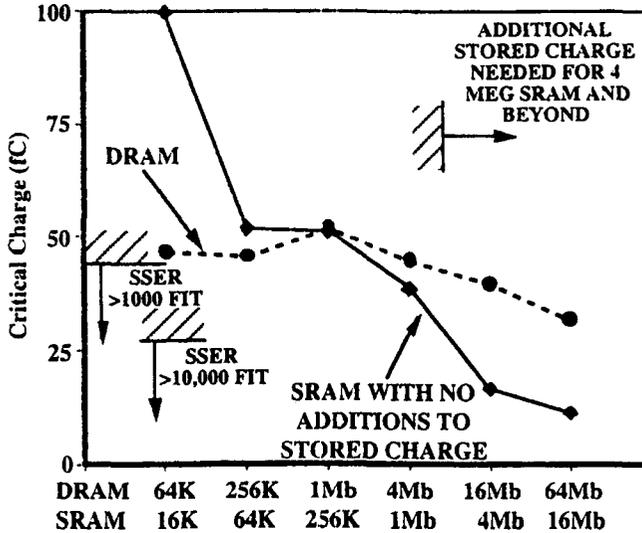


Figure 7.33 Critical charge trends for SRAMs and DRAMs. (After Lage et al., Ref. 29, © IEEE, reprinted with permission.) The region where SSER (system soft error rate) is larger than 1000 or 10,000 FITs (failures in time) is indicated (1 FIT = one failure in 10^9 h.)

7.7.2 Sense Amplifier

While the sense amplifier used in DRAM is almost always a CMOS cross-coupled latch such as shown in Figure 7.7, the SRAM sense amplifiers are available in many different forms. They can be classified into two types: a voltage-sensing type and a current-sensing type shown in Figure 7.34. The voltage-sense amplifier has been used in older generations of SRAMs but suffers from poor voltage gain at low-voltage operation.³⁰ It also requires a data line voltage swing of 100 to 300 mV, which is a disadvantage for low-power operations. On the other hand, the current-sense amplifier operates with a small data line swing of less than 30 mV and provides a large voltage gain even at low supply voltage.³⁰ It is therefore suitable for low-voltage, low-power SRAMs.

7.7.3 Divided Word Line Structure

The divided word line (DWL) scheme, first demonstrated in 64 K CMOS SRAM,³¹ divides the word line into several blocks as shown in Figure 7.35 to reduce the word line delay and the power consumption. The word line of each block (sub-word line) is activated by the AND gate with inputs of the row-select signal running horizontally and the block-select signal running vertically. Only memory cells connected to one sub-word line within the selected block are accessed in a cycle. This scheme is effective in reducing both the word line delay and active power dissipation. This technique can be further extended to a larger number of hierarchies as demonstrated in a three-stage hierarchical row decoder scheme.³²

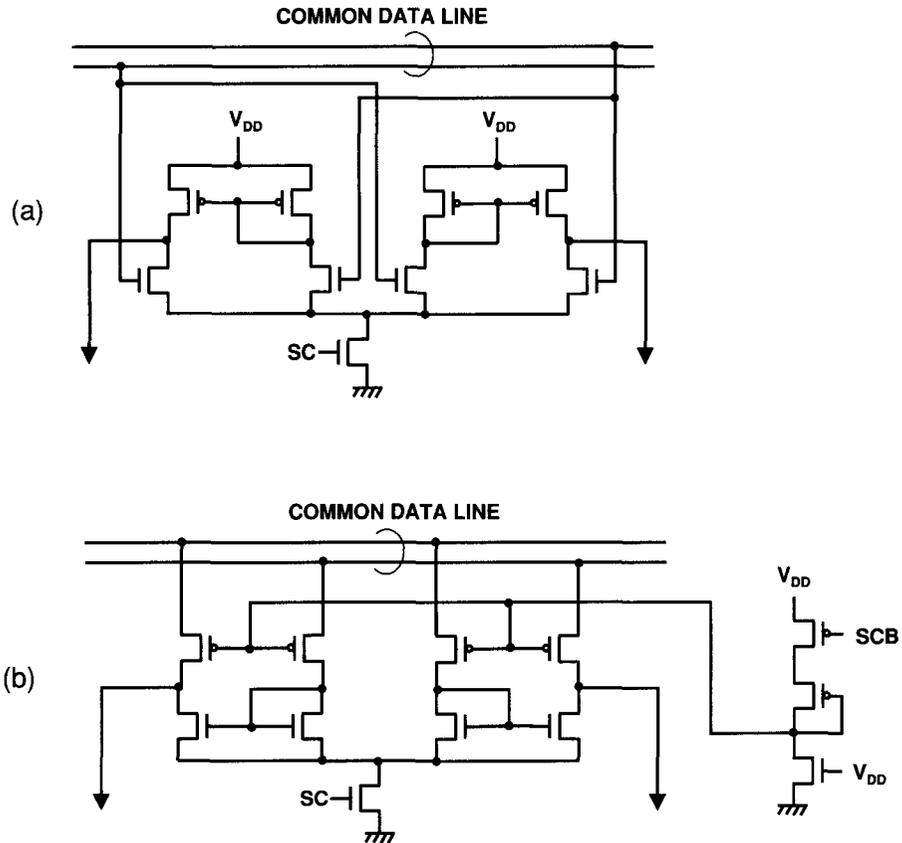


Figure 7.34 Schematics for (a) a voltage-sense amplifier and (b) a current-sense amplifier for SRAMs. (After Sasaki et al., Ref. 30.)

Although the DWL scheme has long been a standard technique for SRAMs, this technique is only beginning to be used in DRAMs starting from 64 Mb generation³³ due to the complexity of the boosting and refreshing requirements.

7.7.4 Address Transition Detection Circuit

The address transition detection (ATD) circuit is another widely used SRAM circuit technique. It detects the transition of any one of the address lines and is used to generate various required internal clock signals. One example of the ATD circuits is shown in Figure 7.36.³⁴ An ATD pulse AD_i is generated by detecting a low-to-high or high-to-low transition of any address signal. All the ATD pulses generated from all the address input transitions are summed up to one equalizing ATD pulse.

The ATD technique has been used to initiate various internal functions, such as bit line equalization before the read operation,³⁵ pulsed word line operation,³⁶ and

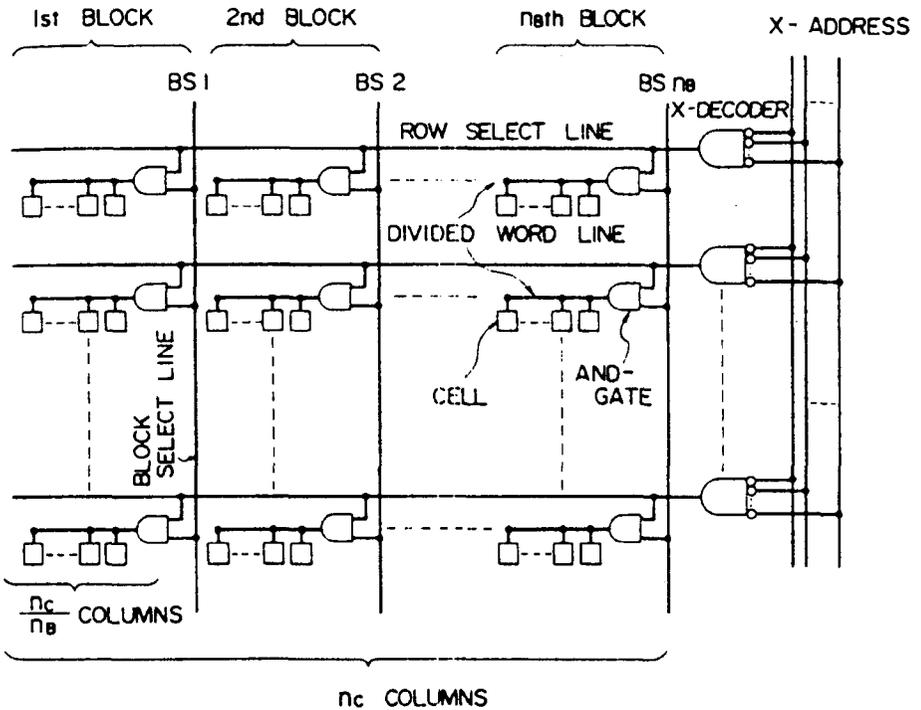


Figure 7.35 Divided word line scheme. (After Yoshimoto et al., Ref. 31 © IEEE, reprinted with permission.)

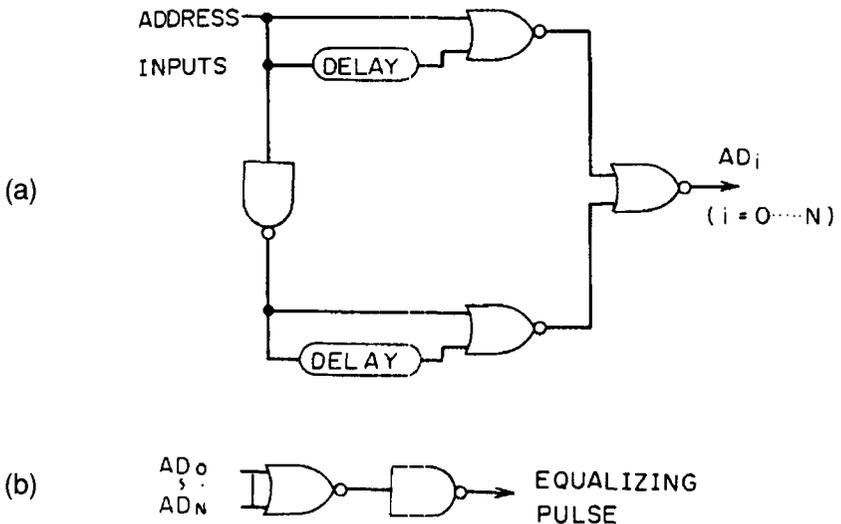


Figure 7.36 Address transition detection (ATD) circuit. (After Tsujide et al., Ref.34, © IEEE, reprinted with permission.) The ATD signal $AD_i (i = 0, \dots, N)$ from each address input is summed to generate one equalizing pulse.

others. These techniques are all effective in reducing the operating current and power dissipation.

7.8 SUMMARY AND FUTURE TRENDS

DRAM and SRAM have benefited tremendously from the general scaling trend since the late 1970s. Both DRAM and SRAM memory cell size have been decreasing $\times 0.4$ per generation or every 3 years. This trend is expected to continue for at least several more generations propelled by the advances in lithography, etching, and other microfabrication techniques. One of the most important future trends for both DRAMs and SRAMs is the low-voltage and low-power operation requirement driven by the proliferation of portable equipment. For example, Figure 7.37 shows how the power supply voltage of DRAM has decreased over several generations. Although the main reason for the voltage reduction has been to avoid the channel hot-carrier effect and to secure the transistor reliability, the power constraint has become the added reason in recent years. The circuit issues in low power DRAMs and SRAMs are well summarized by a 1995 review paper.³⁷

Some of the other future trends and problems specific to DRAM or SRAM mentioned in this chapter are

1. DRAM

- Introduction of new high-dielectric-constant materials for a cell capacitor
- Increasing requirement for low latency and high bandwidth

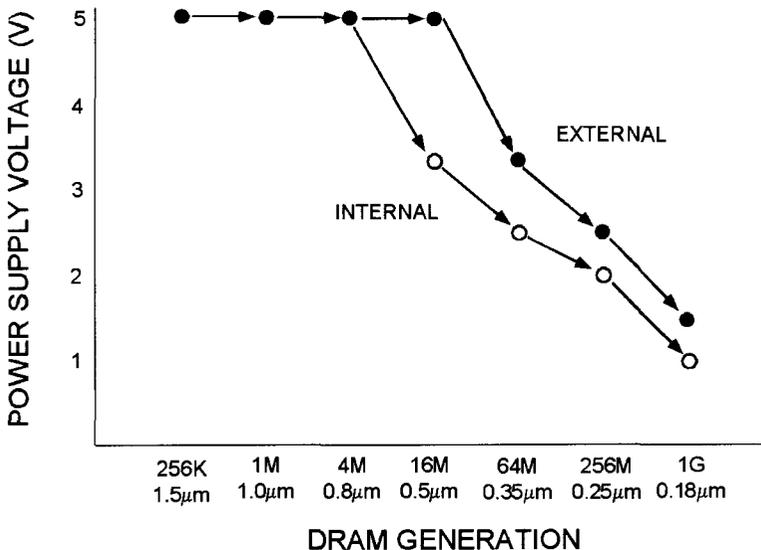


Figure 7.37 Trend of power supply voltage reduction for several generations of DRAMs.

- New memory architecture and circuit techniques to break $8F^2$ cell size limit
 - On-chip integration of DRAM and logic
 - Increasing difficulty in meeting soft-error requirement
2. SRAM
- Full CMOS memory cell becoming dominant due to low-voltage advantage
 - Increasing difficulty in meeting soft-error requirement

With the scaling of memory cell size and transistor sizes (gate length and gate oxide thickness) continuing, DRAM and SRAM chip access speed is also expected to improve, albeit with low-voltage operation. The circuit techniques to achieve the speed improvement at low voltage supply, however, are getting more difficult to come by. In order to avoid the soft error problem, it will be necessary to incorporate on-chip error correction circuits (ECC) for both SRAM and DRAM.

Another important issue for the future is the cost to manufacture high-density SRAMs and DRAMs, although a detailed analysis is beyond the scope of this chapter. The overall cost depends on many factors such as equipment cost, process complexity, wafer and chip sizes, and assembly cost. For both DRAMs and SRAMs, process simplification techniques are necessary to keep the cost per bit on the historical trend. One example for DRAM is the use of high-dielectric-constant materials for a cell capacitor to simplify the overall cell structure and the process steps. With the process simplification and wafer size increase from 200 to 300 mm, the cost per bit can continue to decrease.

REFERENCES

1. M. Koyanagi, H. Sunami, N. Hashimoto, and M. Ashikawa, "Novel High Density, Stacked Capacitor MOS RAM," *IEEE Int. Electron Devices Meeting*, Dec. 1978, p. 348.
2. T. Ema, S. Kawanago, T. Nishi, S. Yoshida, H. Nishibe, et al., "3-Dimensional Stacked Capacitor Cell for 16 M and 64 M DRAMs," *IEEE Int. Electron Devices Meeting*, Dec. 1988, p. 592.
3. T. Kaga, T. Kure, H. Shinriki, Y. Kawaroto, F. Murai, et al., "Crown-Shaped Stacked-Capacitor Cell for 1.5-V Operation 64-Mb DRAM's," *IEEE Trans. Electron Devices* **38**, 255 (1991).
4. W. Wakamiya, Y. Tanaka, H. Kimura, H. Miyatake, and S. Satoh, "Novel Stacked Capacitor Cell for 64 Mb DRAM," *Digest Symp. VLSI Technology*, Sendai, May 1989, p. 69.
5. S. Inoue, K. Hieda, A. Nitayama, F. Horiguchi, and F. Masuoka, "A Spread Stacked Capacitor (SSC) Cell for 64 Mbit DRAMs," *IEEE Int. Electron Devices Meeting*, Dec. 1989, p. 31.
6. H. Arima, A. Hachisuka, T. Ogawa, T. Okudaira, Y. Okumura, et al., "A Novel Stacked Capacitor Cell with Dual Cell Plate for 64 Mb DRAMs," *IEEE Int. Electron Devices Meeting*, Dec. 1990, p. 651.

7. Y. Hayashide, H. Miyatake, J. Mitsuhashi, M. Hirayama, T. Higaki, and H. Abe, "Fabrication of Storage Capacitance-Enhanced Capacitors with a Rough Electrode," *Extended Abstracts 22nd Conf. Solid State Devices and Materials*, Sendai, 1990, p. 869.
8. M. Sakao, N. Kasai, T. Ishijima, E. Ikawa, H. Watanabe, et al., "A Capacitor-Over-Bit-Line (COB) Cell with a Hemispherical-Grain Storage Node for 64 Mb DRAMs," *IEEE Int. Electron Devices Meeting*, Dec. 1990, p. 655.
9. H. Watanabe, T. Tatsumi, S. Ohnishi, T. Hamada, I. Honma, and T. Kikkawa, "A New Cylindrical Capacitor Using Hemispherical Grained Si (HSG-Si) for 256 Mb DRAMs," *IEEE Int. Electron Devices Meeting*, Dec. 1992, p. 259.
10. E. Adler, J. K. DeBrosse, S. F. Geissler, S. J. Holmes, M. D. Jaffe, et al., "The Evolution of IBM CMOS DRAM Technology," *IBM J. Res. Devel.* **39**, 167, Jan./March 1995.
11. S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1981.
12. T. Hamamoto, S. Sugiura, and S. Sawada, "Well Concentration: A Novel Scaling Limitation Factor Derived from DRAM Retention Time and Its Modeling," *IEEE Int. Electron Devices Meeting*, Dec. 1995, p. 915.
13. Y. Shimada, A. Inoue, T. Nasu, K. Arita, Y. Nagano, et al., "Temperature-Dependent Current-Voltage Characteristics of Fully Processed $\text{Ba}_{0.7}\text{Sr}_{0.3}\text{TiO}_3$ Capacitors Integrated in a Silicon Device," *Jpn. J. Appl. Phys.* **35 (Part 1)**, 140 (1996).
14. T. C. May and M. H. Woods, "Alpha-Particle-Induced Soft Errors in Dynamic Memories," *IEEE Trans. Electron Devices* **ED-26**, 2 (Jan. 1979).
15. Z. F. Ziegler and W. A. Lanford, "Effect of Cosmic Rays on Computer Memories," *Science* **206**, 776 (Nov. 1979).
16. W. R. McKee, H. P. McAdams, E. B. Smith, J. W. McPherson, J. W. Janzen, et al., "Cosmic Ray Neutron Induced Upsets as a Major Contribution to the Soft Error Rate of Current and Future Generation DRAMs," *1996 IEEE Int. Reliability Physics Proc.* April 1996, p. 1.
17. M. Inoue, T. Yamada, H. Kotani, H. Yamauchi, A. Fujiwara, et al., "A 16-Mbit DRAM with a Relaxed Sense-Amplifier-Pitch Open-Bit-Line Architecture," *IEEE J. Solid-State Circ.* **23**, 1104 (Oct. 1988).
18. D. Takashima, S. Watanabe, K. Sakui, H. Nakano, and K. Ohuchi, "Open/Folded Bit-Line Arrangement for Ultra High-Density DRAMs," *Symp. VLSI Circuits Digest Tech. Papers*, Honolulu, May 1993, p. 89.
19. K. Shibahara, H. Mori, S. Ohnishi, R. Oikawa, K. Nakajima, et al., "1GDRAM Cell with Diagonal Bit-Line (DBL) Configuration and Edge Operation MOS (EOS) FET," *IEEE Int. Electron Devices Meeting*, Dec. 1994, p. 639.
20. A. H. Shah, C.-P. Wang, R. H. Womack, J. D. Gallia, H. Shichijo, et al., "A 4 Mb DRAM with Cross-point Trench Transistor Cell," *ISSCC Dig. Tech. Papers*, Anaheim, Feb. 1986, p. 268.
21. N. C. Lu and H. H. Chao, "Half- V_{DD} Bit-Line Sensing Scheme in CMOS DRAM's," *IEEE J. Solid-State Circ.* **19**, 451 (Aug. 1984).
22. W. T. Lynch and H. J. Boll, "Optimization of the Latching Pulse for Dynamic Flip-Flop Sensors," *IEEE J. Solid-State Circ.* **9**, 49 (Apr. 1974).
23. N. Kushiyama, S. Ohshima, D. Stark, H. Noji, K. Sakurai, et al., "A 500-Megabyte/s Data-Rate 4.5 M DRAM," *IEEE J. Solid-State Circ.* **28**, 490 (Apr. 1993).

24. M. Ando, T. Okazawa, H. Furuta, M. Ohkawa, J. Monden, et al., "A 0.1 μ A Standby Current Bouncing-Noise-Immune 1 Mb SRAM," *Digest Symp. VLSI Circuits*, San Diego, May 1988, p. 49.
25. E. Seevinck, F. J. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE J. Solid-State Circ.* **22**, 748 (Oct. 1987).
26. C. Lage, J. D. Hayden, and C. Subramanian, "Advanced SRAM Technology—the Race Between 4 T and 6 T Cells," *IEEE Int. Electron Devices Meeting*, Dec. 1996, p. 271.
27. C. Hu, "Alpha-Particle-Induced Field and Enhanced Collection of Carriers," *IEEE Electron Device Lett.* **EDL-3**, 31 (Feb. 1982).
28. P. M. Carter and B. R. Wilkins, "Influence on Soft Error Rates in Static RAM's," *IEEE J. Solid-State Circ.* **22**, 430 (June 1987).
29. C. Lage, D. Burnett, T. McNelly, K. Baker, A. Bormann, et al., "Soft Error Rate and Stored Charge Requirements in Advanced High-Density SRAMs," *IEEE Int. Electron Devices Meeting*, Dec. 1993, p. 821.
30. K. Sasaki, K. Ishibashi, K. Ueda, K. Komiyaji, T. Yamanaka, et al., "A 7-ns 140-mW 1-Mb CMOS SRAM with Current Sense Amplifier," *IEEE J. Solid-State Circ.* **27**, 1511 (Nov. 1992).
31. M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, et al., "A 64 Kb Full CMOS RAM with Divided Word Line Structure," *ISSCC Digest Tech. Papers*, New York, Feb. 1983, p. 58.
32. T. Hirose, H. Kuriyama, S. Murakami, K. Yuzuriha, T. Mukai, et al., "A 20-ns 4-Mb CMOS SRAM with Hierarchical Word Decoding Architecture," *IEEE J. Solid-State Circ.*, **25**, 1068 (Oct. 1990).
33. M. Nakamura, T. Takahashi, T. Akiba, G. Kitsukawa, M. Morino, et al., "A 29ns 64 Mb DRAM with Hierarchical Array Architecture," *ISSCC Digest Tech. Papers*, San Francisco, Feb. 1995, p. 246.
34. T. Tsujide, N. Yasuoka, T. Hara, K. Tokushige, N. Hirakawa, et al., "A 25 ns 16×1 Static RAM," *ISSCC Digest Tech. Papers*, Feb. 1981, p. 20.
35. K. C. Hardee and R. Sud, "A Fault-Tolerant 30 ns/375 mW $16 K \times 1$ NMOS Static RAM," *IEEE J. Solid-State Circ.* **SC-16**, 435 (Oct. 1981).
36. O. Minato, T. Masuhara, T. Sasaki, Y. Sakai, and T. Hayashida, "A 20ns 64 k CMOS SRAM," *ISSCC Digest Tech. Papers*, Feb. 1984, p. 222.
37. K. Itoh, K. Sasaki, and Y. Nakagome, "Trends in Low-Power RAM Circuit Technologies," *Proc. IEEE*, **83**, 524 (April 1995).

PROBLEMS

- 7.1 As described in Section 7.4.2, the DRAM memory cell size can be expressed in the unit of a minimum feature size, F , as nF^2 . From the DRAM cell size and minimum feature size trend curves in Figure 7.4, calculate n , the number of F^2 for each generation of DRAMs.
- 7.2 Figure 7P.1 shows a cross section of a DRAM cell capacitor utilizing high-dielectric-constant Ta_2O_5 film. During the deposition of the Ta_2O_5 film, a thin layer of silicon dioxide is formed on the polysilicon bottom electrode as

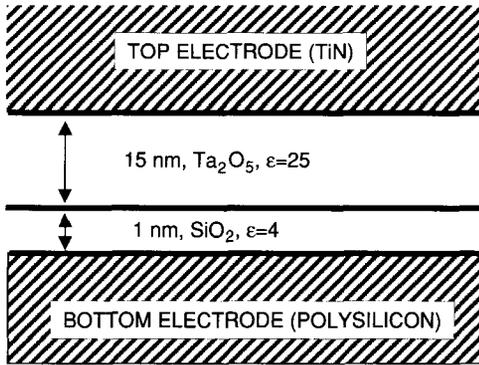


Figure 7P.1 Cross section of a DRAM capacitor with a Ta_2O_5 film.

shown. This oxide layer forms a series capacitor to the intrinsic Ta_2O_5 capacitor and reduces the effective dielectric constant of the combined structure. Using the thickness shown in the figure, calculate the effective dielectric constant of this composite structure.

- 7.3 The “equivalent oxide thickness, t_{eq} ” of a capacitor film is defined as the silicon dioxide thickness that gives the same value of capacitance as the film. What is the equivalent oxide thickness of the structure in Problem 7.2? What is the capacitance density (capacitance per unit area) of this structure?
- 7.4 Figure 7P.2 shows the dimensions of a DRAM stack cell capacitor. Assuming that it has a film structure such as that shown in Figure 7P.1, calculate the cell capacitance (there is no contribution from the bottom surface).
- 7.5 Derive Eq. 7.2.
- 7.6 Repeat the derivation from problem 7.5 for a memory cell storing a binary 0 instead of 1.
- 7.7 For a pn junction with the p side doped to $1 \times 10^{17} \text{ cm}^{-3}$, the n side doped to $1 \times 10^{19} \text{ cm}^{-3}$ and a reverse bias of -2.0 V , calculate the

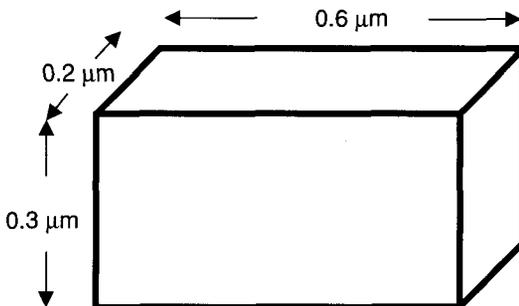


Figure 7P.2 Dimensions of a DRAM cell capacitor stack.

generation–recombination leakage current (third term of Eq. 7.3 assuming $\tau_e = 1 \times 10^{-5}$ s.

- 7.8** Calculate the subthreshold leakage current of a transistor with the threshold voltage $V_{th} = 0.5$ V and the subthreshold slope of 100 mV/decade assuming $I_T = 0.1$ μ A in Eq. 7.5.
- 7.9** What is the bandwidth of a 16-Mb synchronous DRAM with $8 \times$ output (8-bit output) operating with a clock frequency of 200 MHz? Devise several ways that the bandwidth can be increased to 400 Mb/s.
- 7.10** Assuming that the gain, β , of the transistors, Q_{L2} , Q_{A2} , and Q_{D2} in Figure 7.29 are β_L , β_A , and β_D , respectively, and the threshold voltage of all the transistors is V_T , estimate the voltage of the low-side node, L , during the read operation. What is the node voltage if the power supply voltage, V_{DD} , is 3.3 V, the threshold voltage is 0.5 V, and the β ratio is 4?

Nonvolatile Memory

JOHN CAYWOOD

Sub Micron Circuits, Inc.
San Jose, CA

GARY DERBENWICH

Celis Semiconductor Corp.
Colorado Springs, CO

8.1 INTRODUCTION

The focus of this chapter is on reprogrammable nonvolatile memory. It is, of course, possible to make nonvolatile memory by introducing permanent changes in the memory cell structures either during manufacture or in a “programming” step. The former approach is usually referred to as *read only memory* (ROM), and is the earliest form of commercially viable semiconductor memory. The latter approach is referred to as *programmable read-only memory* (PROM), operates by applying an electrical overstress to elements to either cause conductors to blow open (fuse technology) or to cause dielectrics to short-circuit (“short”) (antifuse technology). The ROM and PROM technologies are omitted from this chapter.

The fundamental challenge of reprogrammable nonvolatile technology is that it should be possible to store information in a short time ($\ll 1$ s) so that the time to fully write a memory with a million addresses is economically feasible in volume production while the information, once stored, is retained for more than 10 years ($\sim 3 \times 10^8$ s). This requires very nonlinear phenomena. The first two technologies that are discussed in this chapter, floating-gate memory and silicon nitride memory, employ processes that are very nonlinear with applied electric field. Applied electric fields cause charge to be injected through or into insulators, which modifies transistor thresholds. The thresholds are sensed to determine the data state. A third technology, ferroelectric memory, relies on a field across a dielectric film to switch the orientation of the remnant polarization.

The current from the switching of the remnant polarization is sensed to determine the polarization orientation. Because the remnant polarization in ferroelectric material is a collective phenomenon like ferromagnetism, the switching of the polarization is also a nonlinear phenomenon.

8.2 FLOATING-GATE MEMORY

As has been well documented, the potential on the gate of a MOSFET controls the current between the source and drain of the transistor. Silicon dioxide is an excellent insulator. Charge on a MOSFET gate composed of silicon surrounded by SiO_2 , called a "floating gate" because of the lack of electrical connection, will remain on the gate for a very long time in the absence of external stimuli to remove it. These two phenomena form the basis of floating gate nonvolatile memories.

An energy-band view through a typical, n-channel, stacked-gate nonvolatile memory transistor is shown in Figure 8.1. The band gap of the SiO_2 is ~ 9 eV, and the electron affinity of the Si is ~ 4.7 eV greater than that of the SiO_2 so that there is a barrier of ~ 3.2 eV between an electron in the conduction band in the Si and the conduction band in the SiO_2 and a larger barrier to hole transport. The consequences are that a carrier on the floating gate has a very long retention time but that charging the floating gate is difficult.

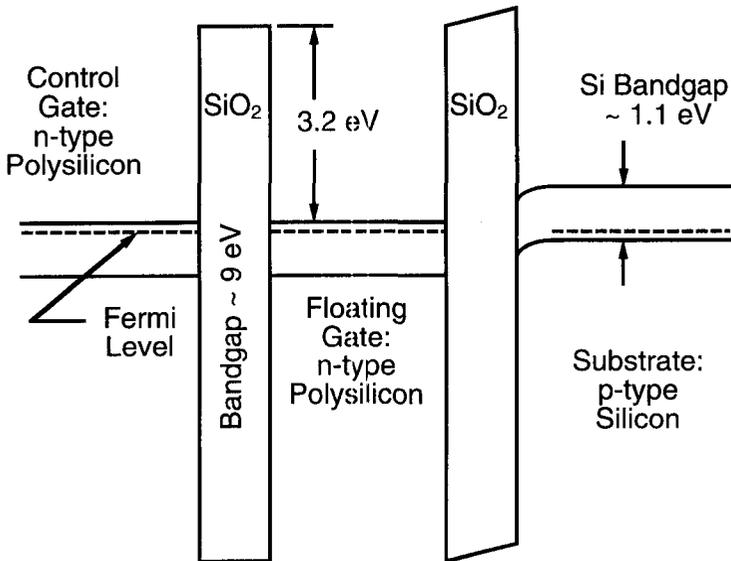


Figure 8.1 Band structure diagram of the floating-gate structure showing the n-type polysilicon control gate on the left, the n-type polysilicon floating gate in the center sandwiched between two SiO_2 layers, and the p-type substrate on the right. The high barrier between the conduction band in the silicon and the SiO_2 inhibits electrons from passing from the silicon into the SiO_2 in normal equilibrium conditions.

8.2.1 Physics of Floating-Gate Technology

Charge Transfer

Two mechanisms are commonly used to charge the floating gate: (1) injection of highly energetic carriers over the barrier or (2) tunneling of carriers through a barrier that has been modified by an electric field, a process called *Fowler–Nordheim tunneling*. These two mechanisms and their associated phenomena are discussed in the following sections.

Hot-Electron Injection For a carrier to be injected over the barrier between Si and SiO₂, a series of events must occur: (1) the carrier must be excited to a sufficiently large energy to surmount the oxide energy barrier; (2) it must travel from the region of energy gain to the Si/SiO₂ interface without losing significant energy; (3) it must cross from the silicon into the oxide (some of the carriers with sufficient energy will be reflected for reasons analogous with the partial reflection of light at a “transparent” glass pane); and (4) finally, the carrier must encounter a field in the oxide that is oriented to sweep the carrier across the oxide rather than oriented to reject the carrier back into the emitting silicon. An energetic electron can lose energy via phonon scattering, electron scattering, or pair production, but the electron density is usually low enough that this latter mechanism can be neglected.¹ A model describing this series of events called the “lucky electron model” has been developed.² Other models of hot electron injection using the hydrodynamic approach or the Monte Carlo method have been developed.^{3–6} These models are more accurate, but provide less physical insight into the process. The barrier height that appears in the calculations is not the low field optical barrier height, but rather an effective barrier height that includes the effects of image force lowering and direct tunneling that are important when the oxide is exposed to a high electric field.⁷

Several sources of hot carriers have been employed in fabricating floating gate nonvolatile memories. These include hot carriers generated in an avalanche region at the drain, channel hot electrons, and carriers generated in a depletion region in the substrate. These are discussed in the following paragraphs.

Avalanche Injection When a junction is reverse-biased, a depletion region is formed. In a MOS drain region, the drain is much more heavily doped than the substrate so that the depletion region lies primarily in the substrate. There is a field in the depletion region given by $E = (V_{\text{BIAS}} + \phi_{pn})/x_D$, where V_{BIAS} is the applied reverse bias, ϕ_{pn} is the difference in the Fermi levels in the p and n regions, and x_D is the depletion layer width as is discussed in Chapter 3. When the field is high enough, a carrier in the field may create an electron–hole pair by impact ionization before losing energy to lattice scattering. The pair created undergoes acceleration in the field and more carriers may be created.⁸

This effect was utilized to fabricate the first commercially available floating gate memory, called *FAMOS* (floating gate avalanche-injection metal oxide semiconductor).^{9,10} Figure 8.2 shows the cross-sectional and circuit schematic drawings of the FAMOS cell. The cell consists of a row-select transistor and a FAMOS

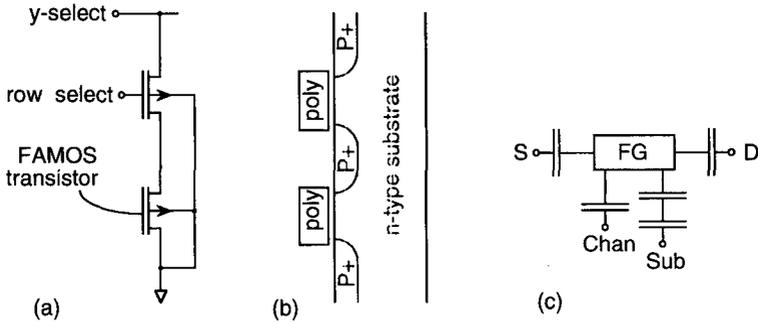


Figure 8.2 (a) Schematic and (b) cross-sectional drawings of a FAMOS memory cell and (c) a schematic drawing of the capacitive coupling of the floating gate.

memory transistor connected in series. There is no direct electrical connection to the gate electrode of the FAMOS transistor, which is fabricated from polycrystalline silicon. The row-select transistor is a standard enhancement p-channel device. The FAMOS transistor is a p-channel device with an enhancement threshold high enough that essentially no current flows through it when -5 V is applied to the drain of a discharged device. When the floating-gate electrode is charged negatively with electrons, the floating gate device is biased into conduction.

Because there are no direct connections to the floating-gate, its potential is strongly influenced by the biases on the neighboring nodes. There are capacitances to the drain, source, channel, and substrate. The substrate coupling is usually relatively weak because there are two capacitors in series, the oxide capacitance and the depletion capacitance. The channel capacitance depends on whether there is inversion charge in the channel. If the channel is in inversion, the inversion layer screens the substrate from the floating gate in the channel region. The potential of the channel varies from that of the drain to that of the source; thus, the channel capacitance can be replaced with two capacitors, one to the drain and one to the source. The division of the channel capacitance between the source and drain is bias dependent, but the drain usually receives 50 to 60% of the channel capacitance. In the case of the FAMOS structure, this means that about half of the bias applied to the drain is coupled to the floating gate.

The FAMOS transistor is programmed by applying a voltage of >-30 V to the drain of the transistor for the $\sim 10\text{-}\mu\text{m}$ scale technology with which these devices were manufactured. The combination of the row and column select circuitry allow selective programming.* The drain coupling applies a large bias to the floating gate so that the channel is in strong inversion. The remaining potential difference

*There is considerable confusion in the literature because the terms program and erase are used in inconsistent and contradictory manners. This chapter adheres to IEEE STD-1005, which states that "program" means to put electrons onto a floating gate and "erase" means to remove electrons from a floating gate.

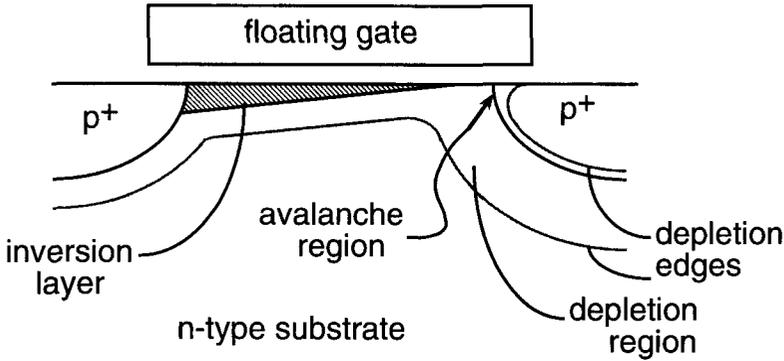


Figure 8.3 Cross-sectional view of the FAMOS transistor during programming showing the formation of the depletion region and the region near the drain where the avalanche occurs and where there is a field to direct the electrons toward the gate.

between gate and drain causes the depletion region in the drain region to bend as can be seen in Figure 8.3.

Hot electrons are created in the avalanche region near the drain. The difference in potential between gate and drain in this region provides fields that direct some of the hot electrons toward the gate oxide where a fraction is injected to the floating gate. A figure of merit for the injection efficiency of a floating-gate memory transistor is $\eta = I_{\text{gate}}/I_{PT}$, where I_{gate} is the floating gate current and I_{PT} is the current at the power terminal, in this case the drain. Avalanche is not a very efficient injection mechanism because there is a large drain current that doesn't contribute to injection and the conditions for collection of the hot electrons that are created in the avalanche are such that only a small fraction are collected which results in $\eta \sim 10^{-6} - 10^{-7}$.

Channel Hot-Electron Injection When an n-channel transistor is in saturation, hot carriers are generated in the high field region near the pinchoff region as is illustrated in Figure 8.4 and discussed in detail in Chapter 6. This phenomenon can be employed to make what is often called a drain-side programming, channel hot-electron (CHE) memory. The control gate is shown self-aligned to the floating gate in Figure 8.4. These gates need not be self-aligned, but usually are in high-density memory arrays to conserve area.

In this approach to a NV memory, the floating gate is supplied with a capacitively coupled control gate. This eliminates the row-select transistor required in the FAMOS approach because the cell can be designed so that the control gate capacitance is the predominant one. A discharged gate has an enhancement threshold selected such that with the control gate at ground, there is essentially no drain current for modest drain bias. With the control gate bias at a positive voltage, often V_{DD} , drain bias results in a drain current. The read biases are chosen such that a gate that is charged negatively by CHE programming does not conduct. Thus, the control gate performs the row-select function. The potential on the floating gate

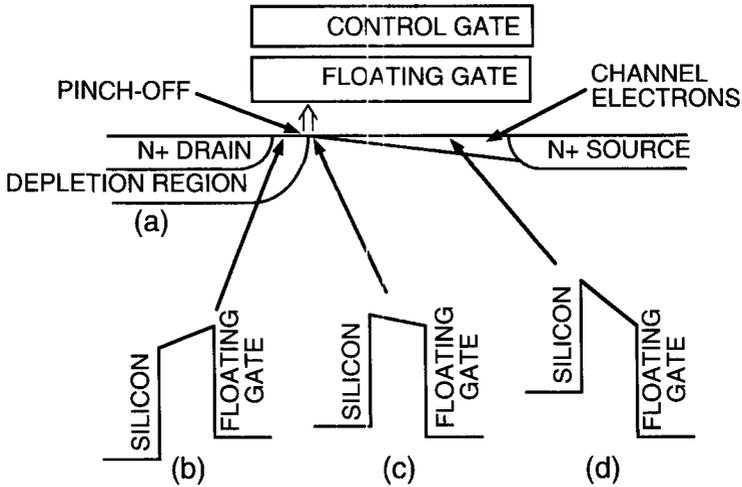


Figure 8.4 (a) Physical layout showing injection of electrons near the pinchoff point. (b)–(d) show the potential diagrams of the conduction bands at the points labeled b–d in cross-section a. Note that near the drain the field in the oxide repels electrons injected over the barrier from the substrate.

depends on the net charge on the gate and the biases on the neighboring nodes

$$\phi_{FG} = \frac{Q_{FG}}{C_{TOT}} + \frac{C_{CG}}{C_{TOT}} V_{CG} + \frac{C_D}{C_{TOT}} V_D + \frac{C_S}{C_{TOT}} V_S + \frac{C_{Chan}}{C_{TOT}} V_{Chan} + \frac{C_B}{C_{TOT}} V_B \quad (8.1)$$

where $C_{TOT} = C_{CG} + C_D + C_S + C_{Chan} + C_B$ and the C terms indicate the capacitance between the floating gate and the control gate, drain, source, channel, and body, respectively. The capacitance ratios appearing in Eq. 8.1 are often referred to as the *coupling ratios*.

To program with CHE, hot electrons must be generated in the pinchoff region and collected on the floating gate. For efficient collection, the floating gate must be at a more positive potential than the channel in the vicinity of the pinchoff point where most of the hot electrons are generated. These conditions can be created by applying a moderately high voltage to the control gate to provide a collecting potential for the hot carriers. A somewhat lower drain bias serves to produce the hot carriers.

The lucky electron model has been applied to drain-side CHE programming of EPROMs.¹¹ The model predicts that the gate current is given by

$$I_{gate} \approx 0.5 \cdot \frac{I_{DS} t_{OX}}{\lambda_R} \left(\frac{\lambda E_m}{\Phi_B} \right)^2 P(E_{OX}|_L) \exp\left(\frac{-\Phi_B}{\lambda E_m} \right) \quad (8.2)$$

where t_{OX} is the oxide thickness, λ_R is the electron elastic scattering length (i.e., the acoustic phonon scattering length), λ is the inelastic scattering length (the

combination of the scattering length for emission of optical phonons and impact ionization), E_m is maximum value of the channel lateral electric field, $P(E_{OX}|L)$ is the probability of a hot electron traveling from the generation site to the peak of the oxide potential barrier without inelastic scattering, E_{OX} is the field across the gate oxide at the drain pinchoff point, given approximately by $E_{OX} \approx (V_{DS} - V_{GS})/t_{OX}$, and Φ_B is the potential barrier between the silicon conduction band and the oxide conduction band, given by $\Phi_B = 3.2 - \beta\sqrt{E_{OX}} - \zeta(E_{OX})^{3/2}$, eV where β and ζ are constant parameters. The first term in the barrier height expression accounts for image-force lowering and the second for effective tunneling lowering.

Design of the drain-side CHE cell involves several tradeoffs. For efficient programming, it is desirable to maximize the lateral electric drain field to generate high-energy electrons. Factors that tend to advance this goal are to shorten the channel length, have an abrupt drain junction, and have a relatively high channel doping. However, there is a minimum allowed channel length. If the channel is too short, a device on an unselected row may punch through, which gives rise to excess bit line current and inadvertent programming of the cell. For cells without a series-select gate, coupling of the drain bias to the floating gates on the unselected rows can exacerbate punch through, an effect referred to as drain turnon. Increasing the channel doping concentration increases the threshold of erased cells, which reduces the read current.

Several factors limit the efficiency of this injection mechanism. One factor is that hot electrons are only a by-product of the channel current. Most of the channel electrons don't become hot enough to surmount the oxide energy-barrier because of electron-phonon scattering. Those hot electrons that are created have momentum parallel to the oxide-silicon interface. They must be scattered to be redirected toward the interface. Finally, as is shown in the potential diagrams in Figure 8.4, over most of the depletion region the potential between the floating gate and channel regions is repulsive for the injection of electrons. Only in the immediate vicinity of the pinch off region are there both hot carriers and an attractive field. Typical values for the injection efficiency of a drain-side channel hot-electron injection structure shown in Figure 8.4 are $\sim 10^{-6}$.

Source-Side Electron Injection An approach to channel hot-electron injection that has demonstrated significant improvement in injection efficiency is what is commonly referred to as *source-side injection*.¹² This technique adds an injection gate adjacent to the source side of the floating-gate as is illustrated in Figure 8.5. The injection gate is biased only slightly above threshold. This serves to reduce the channel current and provides a pinchoff region at the drain end of the injection gate, which is near the source of the floating gate. The floating gate is capacitively coupled by a control gate to a positive voltage. Electrons flowing through the channel of the injection gate are accelerated in the high electric field resulting from the potential difference between the injection and control gates so that they become hot. The hot electrons are injected into the depletion region that exists under the floating gate because the drain is biased sufficiently positive to keep the relatively small number of electrons injected by the injection transistor swept out of the region. The hot

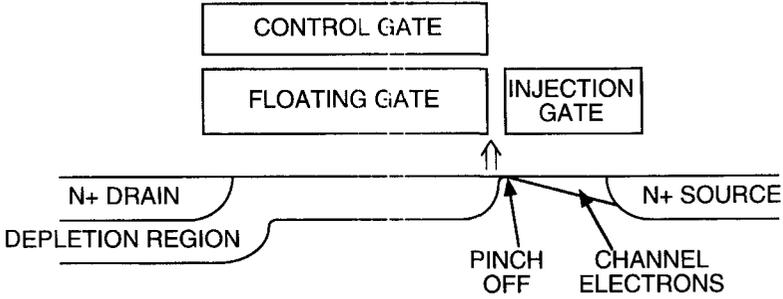


Figure 8.5 Schematic illustration of a generalized source-side injection cell showing the pinchoff of the channel at the end of the injection gate and the depletion region extending under the length of the floating gate to form a virtual drain at the drain end of the injection gate. The broad arrow indicates where the electrons are injected at the source end of the floating gate.

carriers are injected into a region in which the fields favor collection. Because of this, the injection efficiency can be as high as $\sim 10^{-3}$.¹³⁻¹⁵

Substrate Injection Another approach that has been demonstrated to yield relatively high injection efficiencies is the injection of electrons from the substrate. Several methods to accomplish this have been discussed.^{16,17} Typical structures utilize a buried n^+ emitter as is shown in Figure 8.6. During programming, the n^+ injector and the substrate are grounded. The control gate, source, and drain are biased at a high voltage. This causes a depletion layer to form under the floating gate and the junctions. The field from the depletion region punches through to the n^+ injector and causes electrons to be injected from the n^+ diffusion into the depletion region. Once the electrons enter the depletion region, they are accelerated by the depletion layer field in the direction of the gate. Because essentially all of the current

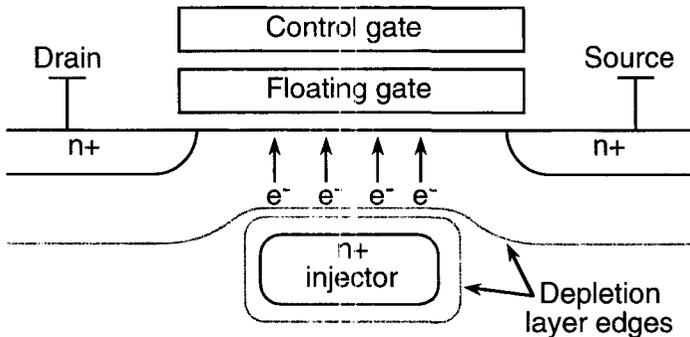


Figure 8.6 Cross section through a substrate injection EPROM cell during programming showing the electrons generated by punch through at the injector being accelerated through the field of the depletion region toward the floating gate.

is hot-carrier current and because the hot electrons inherently have the appropriate direction for injection, injection efficiencies in the range of 10^{-1} – 10^{-4} have been observed. EPROMs utilizing substrate injection have not been introduced as commercial products, probably because the injector diffusion does not scale well to the submicrometer regime.

Fowler–Nordheim Tunneling Tunneling is a quantum-mechanical process in which a particle can pass through a region in which it is classically forbidden. In the context of the current discussion, tunneling is a process that allows electrons to pass from the conduction band of one silicon region to that of another silicon region through an intervening region of silicon dioxide. Tunneling has been observed at low voltages through very thin oxide layers, ~ 20 – 30 Å thick. This is, however, not very useful for nonvolatile memories because charge that is placed on a gate by this process finds it equally possible to leave, which adversely affects the retention.

However, the probability that an electron will tunnel through an oxide barrier under low bias conditions decreases exponentially with thickness of the oxide, so that for an oxide with thickness ~ 10 nm, the probability becomes practically zero. The tunneling barrier for an electron in the conduction band of the silicon is the shaded area seen in Figure 8.7a. If a very high electric field is applied across the

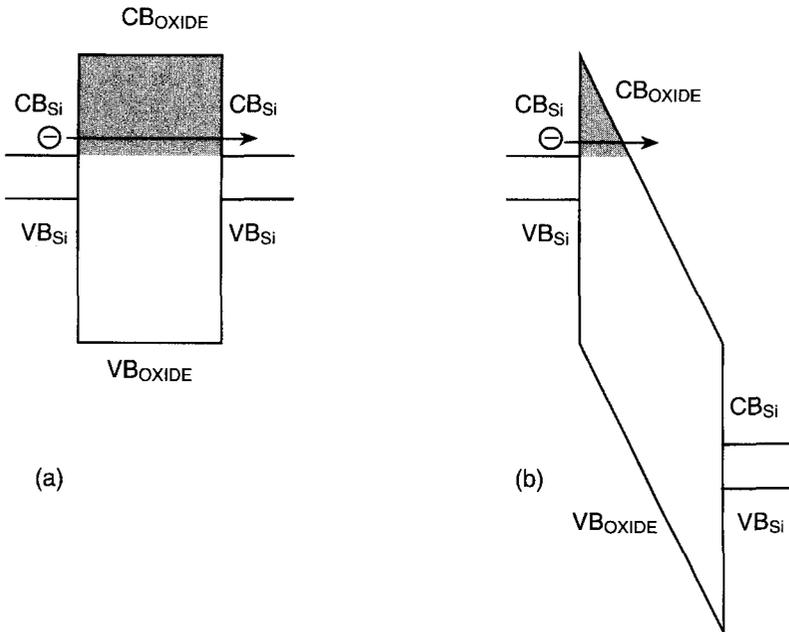


Figure 8.7 Potential diagram showing the band structure of a Si/SiO₂/Si sandwich under low-field (a) and high-field (b) conditions. The arrows illustrate the paths of any tunneling electrons. Figure 7b is scaled approximately correctly for a 10-nm-thick oxide with an applied bias of 10 V.

oxide, the potential barrier is distorted and the barrier to an electron tunneling into the conduction band in the *oxide* is reduced to the shaded area seen in Figure 8.7*b*.

The problem of tunneling through a trapezoidal barrier was first solved with the use of the WKB approximation for tunneling from a metal into vacuum by Fowler and Nordheim in 1928.¹⁸ The resultant equation is

$$J = \left(\frac{q^3 E^2}{8\pi h \Phi_B} \right) \exp \left[\frac{-4\sqrt{2m}\Phi_B^{3/2}}{3\hbar q E} \right] \quad (8.3)$$

where E is the applied electric field, Φ_B is the potential barrier between the metal Fermi level and the vacuum ground state, m is the electron mass, and the other symbols have their usual meanings.

Lenzlinger and Snow slightly modified Eq. 8.3 to describe tunneling through an oxide layer on silicon by replacing the electron mass with the effective mass of the electron in the SiO₂ conduction band.¹⁹ This effective mass is usually taken to be about 0.5 of the free-electron mass. There are corrections which account for the effect of temperature on the electron distribution, but these are small effects. For example, the tunneling current changes by 30% over the temperature range 90 to 473 K.²⁰ Other effects have been shown to modify the tunneling probability slightly. Maserjian detected oscillations in the current-voltage characteristic that he attributed to interference effects of the wavefunctions in the conduction band of the oxide.²¹ From careful analysis of the dependence of the oscillations on thickness and voltage, it was concluded that the scattering length of electrons in the conduction band of the oxide, that is, the mean coherence distance, is 13 Å.²²

For tunneling between two silicon layers, the formation of a thin, essentially two-dimensional, accumulation layer in the silicon on one side of the oxide and a depletion layer in the silicon on the other side of the oxide, provide voltage drops in series with the oxide.^{23,24} These effects can decrease the voltage across the oxide by over 0.5 V, which is large enough to cause the current to change by approximately an order of magnitude. Modeling based on this work has been successful in matching experimental results without adjustable parameters.²⁵ Tunneling into the oxide conduction band shown in Figure 8.7 is called *Fowler-Nordheim tunneling* to distinguish it from tunneling from one conductor to another conductor which is called direct tunneling.

All the work referred to so far is for tunneling between two planar surfaces that form equipotential planes. There are cases of technological interest that do not fit this simple description. One case is tunneling between two nonplanar polysilicon surfaces. Another case is tunneling from a floating polysilicon gate to a diffused junction that extends under the gate with the result that the fields across the oxide separating the gate from the junction are laterally non-uniform.

Work on electron conduction through oxides grown on polysilicon established that the electron current was a result of Fowler-Nordheim injection aided by the geometrically enhanced fields that are the natural result of convex features on the surface of the polysilicon. The important geometric features for tunneling

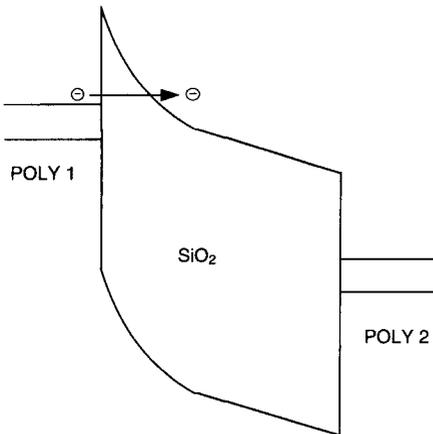


Figure 8.8 Band diagram for tunneling from a textured poly emitter to a poly collector. Note that the electric field is concentrated at the emitting poly 1 surface.

enhancement have radii of curvature that are small with respect to the thickness of the oxide film.^{26–28} The enhancement effect can be large; the mean electric fields across the oxides are often only 10–20% as high as for those needed to obtain similar emission from planar electrodes.

The band diagram in Figure 8.8 may help make this result a bit more intuitive. Electrostatics assures that the electric fields will be highest in the regions of greatest positive curvature. The high electric field near the emitting surface means that the potential energies decrease rapidly near this surface and decreases more gradually away from the emitting surface. The tunneling occurs through the thin potential barrier near the emitting surface. The reduced field away from the emitting surface serves to sweep the electrons injected into the oxide conduction band across the oxide to the collecting electrode. The electron emission into the oxide depends essentially only on the electric field at the emitting electrode as long as the dielectric is thick with respect to the radius of curvature of the emitting surface.

Development of commercial memories operating on this principle spurred work on calculating the fields resulting from surface curvature.^{29,30} Roy and co-workers correctly pointed out that the tunneling occurs not through a triangular barrier, as Fowler and Nordheim assumed, but through a barrier with a concave-shaped top surface.³¹ A consequence of this is that the current varies more rapidly with electric field than would be the case for a triangular barrier potential, especially at lower fields.

Although early work focused on emission resulting from enhancement coming from texture on the surface of polysilicon, Marcus and Sheng and Kao et al. showed that oxidation of polysilicon traces can result in the formation of edges on the poly with small radii of curvature.^{32–34} The resultant shapes are more uniform than the features observed on the top surface of polysilicon, in spite the dependence of the shape of the edges on reduced oxidation rates coming from compression in the oxide and plastic flow of the oxide and polysilicon, which, in turn, depend on oxidation ambient, silicon crystal structure and doping in complex ways.³⁵ For this reason,

commercial memories that utilize interpoly tunneling all seem to depend on tunneling from edges.

The I-V characteristics for geometrically enhanced tunneling are very asymmetrical with respect to bias polarity because of the asymmetry of the shape of the surfaces on the two sides of the interpoly oxide. The interpoly oxides are conformal so that if the surface of a polysilicon layer has structure with positive radii of curvature (i.e., is convex), the opposing surface will have negative radii of curvature (i.e., be concave). At a given bias, the tunnel current for the two bias polarities can differ by over 10 orders of magnitude. This is very different from the behavior of tunneling from planar surfaces for which the I-V characteristics of the tunnel oxide are essentially symmetrical.

There are memories that depend on tunneling from the floating gate to a positively biased source or drain junction lying under the floating gate to remove electrons from the gate. The fields in the vicinity of the drain can be rather nonuniform for reasons illustrated in Figure 8.9. The high field between the n^+ diffusion region and the gate that is necessary to induce tunneling will deplete a portion of the n^+ diffusion region and cause the depletion region to curve back under the gate, as is illustrated in Figure 8.9. The electric fields across the gate oxide are reduced in the neighborhood of the depletion region. There are very high fields in the silicon in this depletion region that will lead to *band-to-band tunneling* (BBT), which occurs when the reverse field across the junction is high enough that electrons in the valence band in the n region can tunnel to states in the conduction band in the p region as shown in Figure 8.9. The hole remaining behind in the n region can be accelerated by the electric field with effects on cell performance and reliability that are discussed in later sections.

Additionally, oxidations that are conducted after the floating gate edge is defined will result in growth of the oxide near the gate edge with the consequent formation of

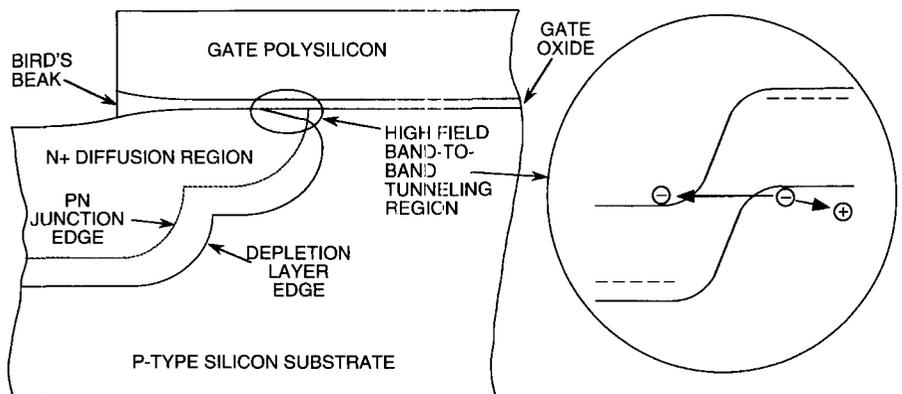


Figure 8.9 Cross section showing representative features of a floating gate during erasure from a diffusion under an edge of the gate. The curvature of the depletion region resulting from the depletion of a portion of the n^+ drain region and the bird's beak extension under the gate can cause tunneling as illustrated in the band diagram on the right.

a bird's beak extending under the polysilicon. The electric field across this bird's beak region is, of course, less than that across the thinner oxide further from the poly edge.

Because of the rapid variation of the oxide electric field along the channel, accurate numerical simulation of the FN tunneling in this structure must be done using a very fine mesh in the junction region.

Ultraviolet Light Erase Light (photons) can be absorbed by matter and excite electrons to sufficiently high energy that the electrons can surmount the barrier that binds them to the matter and be emitted. Einstein won the Nobel Prize for explaining that the energy that the electrons acquire depends on the energy of the individual photon ($E = h\nu = \hbar/\lambda$) and not on the intensity of the light (i.e., the number of photons). Thus, if the wavelength is longer than the minimum necessary for emission, no emission will occur. If the light is of sufficient energy to cause emission, the number of electrons emitted will be proportional to the number of photons (the light intensity).

Although Einstein's work related to emission from metals into a vacuum, the same principles apply to the emission of electrons from silicon over the barrier of surrounding SiO_2 . As is illustrated in Figure 8.10, electrons can be excited from either the valence band or from the conduction band. Excitation from the conduction band requires a lower energy, $h\nu_1$, than excitation from the valence band, $h\nu_2$. However, excitation can only occur from the conduction band if there are electrons

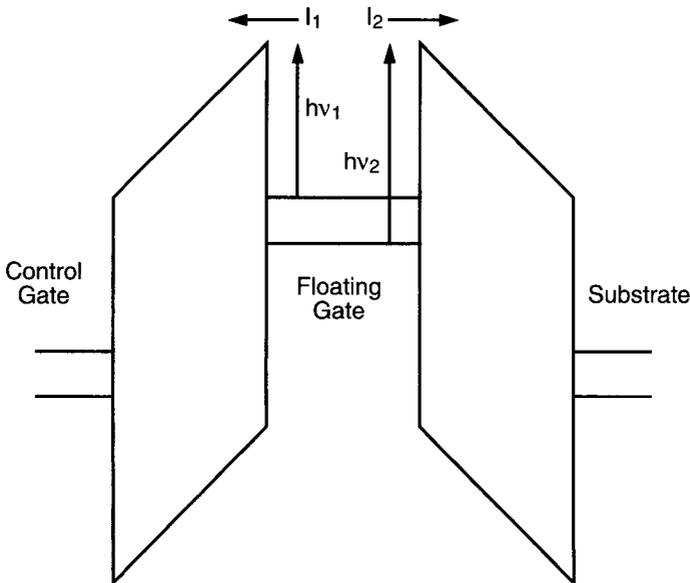


Figure 8.10 Schematic band diagram showing the excitation of electrons from the conduction band or from the valence band to energies at which they can surmount the oxide barrier and escape to the control gate, I_1 , or the substrate, I_2 .

present to excite (i.e., the material must be fairly heavily n type doped). Moreover, longer-wavelength light that will excite electrons from the conduction band to above the oxide barrier can penetrate farther into the silicon before being absorbed than can shorter wavelength light, so there is a good chance that the excited electrons will lose energy while traveling from the site of excitation to the Si/SiO₂ interface. The consequence of this is that the chance is much less for an electron excited from the conduction band to surmount the barrier than that for an electron from the valence band. These features can all be seen in the data shown in Figure 8.11. Commercial erasers utilize the 2540-Å Hg line for efficient operation.

Several factors complicate this picture. Usually the UV erasure takes place with the device unpowered so that photocurrents can flow in both directions, that is, to both the control gate and the substrate, as shown in Figure 8.10. Modern EPROMs are commonly CMOS ICs with the floating and control gates doped n type. However,

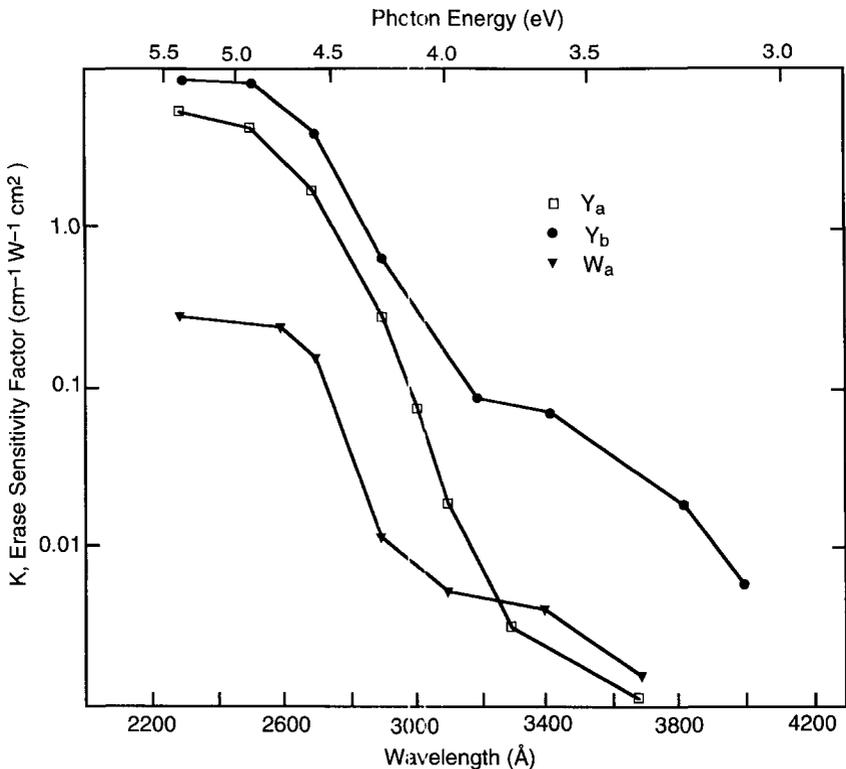


Figure 8.11 Spectral erase sensitivities of various UV-EPROM devices. Note that while all three show a large increase in sensitivity at photon energy of ~ 4.2 eV, which corresponds to the energy required to excite electrons from the top of the valence band to above the oxide barrier, the two cells, Y_b and W_a , with n-type floating gates also have weak shoulders extending to about 3.2 eV, the energy required to excite a conduction band electron over the barrier.³⁶

the substrate is n type in the source–drain overlap regions and p type along the channel so the Fermi level diagram is location dependent. More importantly, the control gates are connected to drive transistors whose pn junctions will act as photodiodes with the result that the control gate will not be at ground potential.

Another potential complication is that the diagram in Figure 8.10 assumes that there is SiO₂ between both the floating gate and control gate and between the floating and channel as is indicated by the insulator bandgaps. In practice, the dielectric between the two gates is often an oxide/nitride/oxide sandwich that acts to trap the photocurrent emitted in the direction of the control gate. This trapped charge inhibits further emission in the control gate direction. It can also later cause threshold shift of a programmed cell as the trapped charge redistributes.

A final complication is the obstacles the light may face in reaching the interfaces where charge transport can occur. Because the silicon is a good absorber of light in the wavelengths of interest, the gates shadow the effective emission regions. There is evidence that the light depends on both reflection and diffraction effects to slip in under the gates and that this occurs primarily in the field oxide regions. Some modern high-density technologies with multilevel-metal interconnection layers may leave only a small portion of the cell exposed to light. This can cause rather slow erasure of these devices.

Sensing of Floating-Gate Cells

Binary Sensing The most common sensing condition for a floating-gate memory is to couple voltage to the floating gate such that the channel under the gate will be inverted for one charge state of the floating gate and not be inverted for the other state. If the surface under the gate is inverted, current can flow and be sensed as one logical state; if the surface is not inverted, no current flows which is sensed as the opposite state. For the typical *NV* array, Eq. 8.1 simplifies to

$$\phi_{FG} = \frac{Q_{FG}}{C_{TOT}} + \frac{C_{CG}}{C_{TOT}} V_{CG} + \frac{C'_D}{C_{TOT}} V_D \quad (8.4)$$

in the sense condition where the prime on the drain capacitance indicates that the portion of the channel capacitance that can be associated with the drain has been included in this term. The apparent threshold, V_{ta} , as measured from the control gate is for this case given by

$$V_{ta} = \frac{C_{TOT}}{C_{CG}} V_m - \frac{C'_D}{C_{CG}} V_D - \frac{Q_{FG}}{C_{CG}} \quad (8.5)$$

where V_m is the “natural” threshold of the floating gate poly transistor when a contact is made to this layer.

If the cell has a select transistor, as illustrated schematically in Figure 8.12*b*, it is enough that the floating gate threshold be above the high level of bias for the control gate for the programmed state and be below this value for the erased state, as shown in Figure 8.13*a*. If the control gate is biased at the positive supply, V_{DD} , the

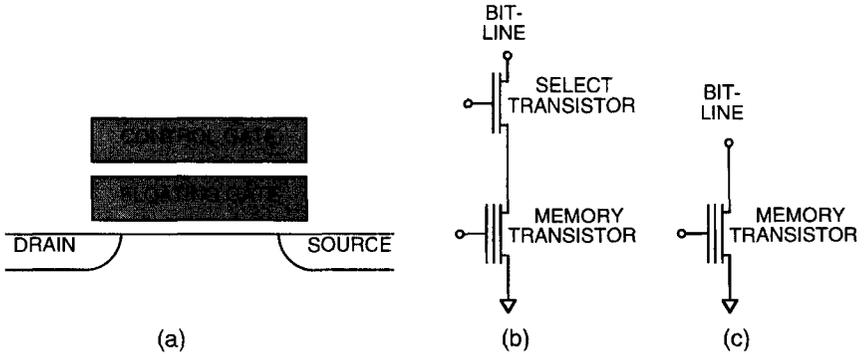


Figure 8.12 (a) Cross section through a stacked self-aligned floating-gate memory transistor; (b) schematic of the memory transistor in a cell with a select transistor; (c) schematic of the memory transistor in a cell without a select transistor.

programmed threshold voltage should be above V_{DD} and the erased threshold voltage less than V_{SENS} . If current is sensed, the cell is erased; if no current is sensed, the cell is programmed. It is also possible to take advantage of the select transistor and arrange it so that the erased state is conducting and the programmed state nonconducting with the control gate grounded. This choice sometimes provides opportunities for clever circuit design.

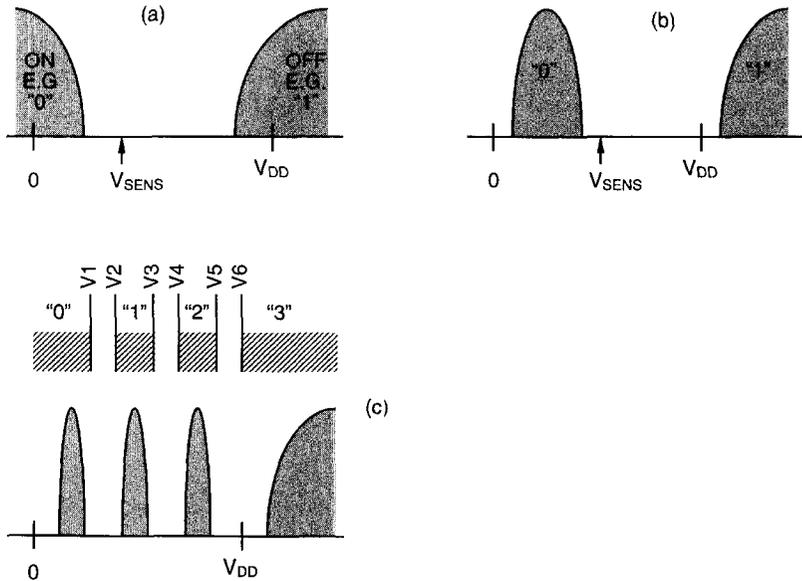


Figure 8.13 Typical threshold voltage ranges for (a) a 1-bit memory cell with select transistor, (b) 1-bit memory cell without select transistor, (c) 2-bit memory without select transistor. These are the apparent thresholds as measured from the control gate.

If the cell lacks a select transistor, as shown in Figure 8.12c, the bias on the control gate must serve to select the cell. This requires that there be a positive bias on the control gate in the sense mode. As in the previous case, the programmed cell threshold must be above the high level of the control gate. The erased cell threshold must be somewhat above 0 V, but still below V_{SENS} , as shown in Figure 8.13b. When these conditions are met, the floating gate will not conduct with the control gate at ground for either the programmed or erased state, which acts to deselect the cell. When the control gate is biased at its sensing voltage, the erased cell will conduct, but the programmed cell will not, allowing the charge state of the floating gate to be sensed.

Multilevel Storage If the threshold of the floating gate could be controlled so as to fall into a series of four bands with gaps in between, as illustrated in Figure 8.13c, it would be possible to determine into which of the four bands the threshold fell by applying multiple biases on the control gate. This would allow two bits to be stored in a single cell. (Alternatively, the cell current could be sensed and divided into four groups.) As is discussed in later sections, certain phenomena will tend to affect the charge on the floating gate and, hence, the cell threshold. The smaller the gap between the threshold values that constitute a state, the more difficult it becomes both to sense the state correctly and to maintain the threshold in the desired range. However, the potential gains in bit storage density make this an attractive approach.

8.3 FLOATING-GATE MEMORY ARRAYS

It is not very meaningful to discuss floating gate memory cells without also discussing the arrays into which the cells are embedded because the key to a cell's usefulness is whether the cell can be written to and read from without affecting the surrounding cells. It is also true that the function of a cell may be determined by the surrounding circuitry. Floating-gate nonvolatile memories are conventionally classified as EPROMs, E²PROM, or flash memories. EPROMs, or UV EPROMs as they are sometimes called, are fairly easily distinguished from the other two categories. E²PROMs are distinguished from flash memories, sometime called *flash EEPROMs*, in that they may be written to an arbitrary pattern on a byte basis. Flash memories, on the other hand, can be cleared to one state on a block basis. Bytes within the block can be subsequently written to an arbitrary pattern by setting some of the bits to the state that is opposite to the clear state. Some cells are used for both E²PROM or flash arrays; others are used for only one type of array. In the discussion that follows, cells that can be used for both E²PROM and flash arrays are classified with the more functional memory, the E²PROM.

8.3.1 UV EPROMs

All commercially available EPROMs utilize drain-side CHE programming and UV erase. The most common approach is that usually referred to as the "T" cell. The

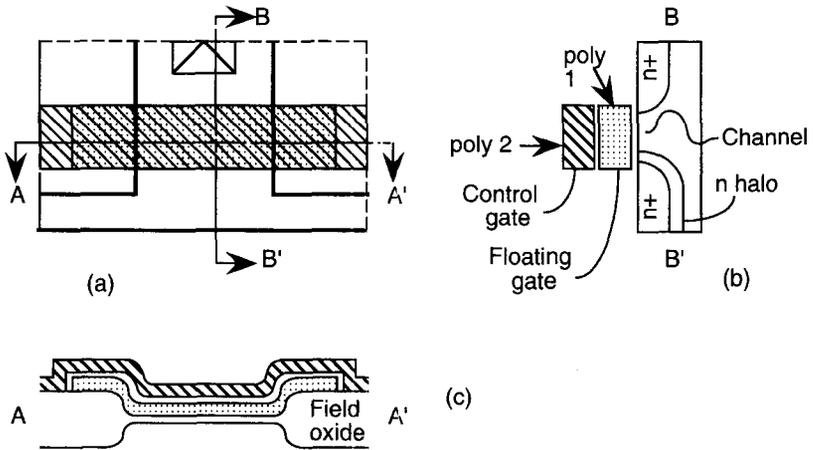


Figure 8.14 (a) Top view and (b,c) cross sections through a T-cell UV EPROM.

name is derived from the shape of the active region for a single cell, which can be seen as the region defined by the heavy line in Figure 8.14. The view through the cell along the channel is exactly that shown in Figure 8.14a. In the lateral direction the floating gate extends over the field oxide to provide “wings” that add to the floating gate to control gate capacitance without increasing the other capacitances significantly. This improves the capacitive coupling ratio C_{CG}/C_{TOT} , sometimes referred to as the *gate coupling ratio*, which provides better control of the floating gate potential by the control gate bias. This is an important factor in efficient programming.

The T EPROM cells are organized into arrays as shown schematically in Figure 8.15. The bit line contacts are shared between two cells, as can be seen in Figure 8.14, and all of sources are common. Consequently, during programming, the drain programming bias is applied to every cell on a bit line that is being programmed. Only one row will be biased to V_{PP} for programming. Typically, for submicron channel lengths, 10–14 V is applied to the control gate for programming and about half of that bias is applied to the drain. These bias conditions cause peak currents in the range of 0.3–1 mA to flow through each cell that is being programmed.

The inhibit function depends on all the unselected word lines being grounded so that the cells on these word lines do not conduct. As can be seen from Eq. 8.1, the drain capacitance will couple a fraction of the drain bias applied to program the selected cell onto the floating gate of the unselected cells on the same bit lines. This coupling will tend to turn the unselected gates on and promote punchthrough. The cell must be designed with a choice of natural threshold, drain coupling, and minimum channel length to avoid the punchthrough effect.

Attempts have been made to improve on the bit density of EPROM arrays by minimizing or eliminating the area occupied by the bit line contact by using what is referred to as a virtual ground array. Two approaches that have found commercial

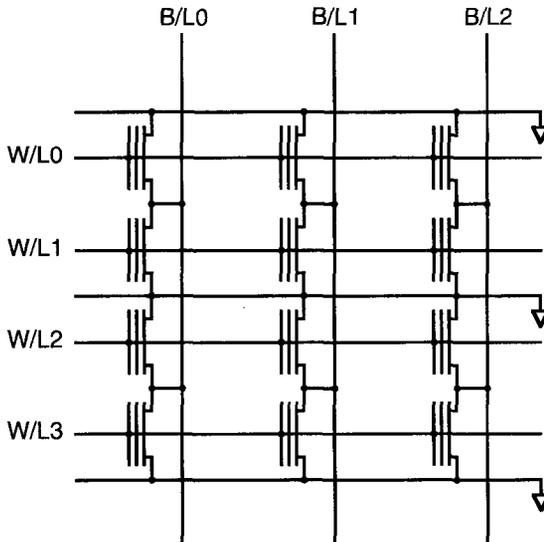


Figure 8.15 Schematic diagram of a portion of a T-cell EPROM array.

application have the same appearance in cross section along the channel as the channel as the T-cell EPROM. These have been named the X cell and alternate metal ground (AMG*) approaches.^{37,38}

The X-cell approach is so named because the active region on a 2×2 array of cells forms the letter “X” when viewed in layout, as seen in Figure 8.16. The cross section parallel to the current flow, indicated by the line B–B’, is the same as that shown for the T cell in Figure 8.14; similarly, the view perpendicular to the current flow and indicated by A–A’ matches this similarly labeled section in Figure 8.14.

The advantage of this approach over the T cell is that the bit line contact is shared over four cells rather than two cells. The cost for this approach is more complicated peripheral circuitry. Figure 8.17 shows a schematic diagram of the cells in Figure 8.16 and the operation conditions. As can be seen from the table, the price paid is that the source line must be decoded and switched rather than being simply tied to ground as is the T-cell source.

Although the X cell increases the number of cells that can share a bit line contact to four, area could be saved by eliminating all drain and source contacts from the interior of the array. This is the approach adopted by the AMG cell and array. A layout of this approach together with cross section views is shown in Figure 8.18. Note that there are no contacts in the array core and that field oxide is not used for isolation in this region.

The n^+ diffusions that are seen in cross sections A1–A1’ and A2–A2’ are formed after the first layer of poly is patterned and before the second layer of poly is deposited. The portions of the poly 1 stripes laying between the poly 2 stripes are

*AMG is a registered trademark of WSI.

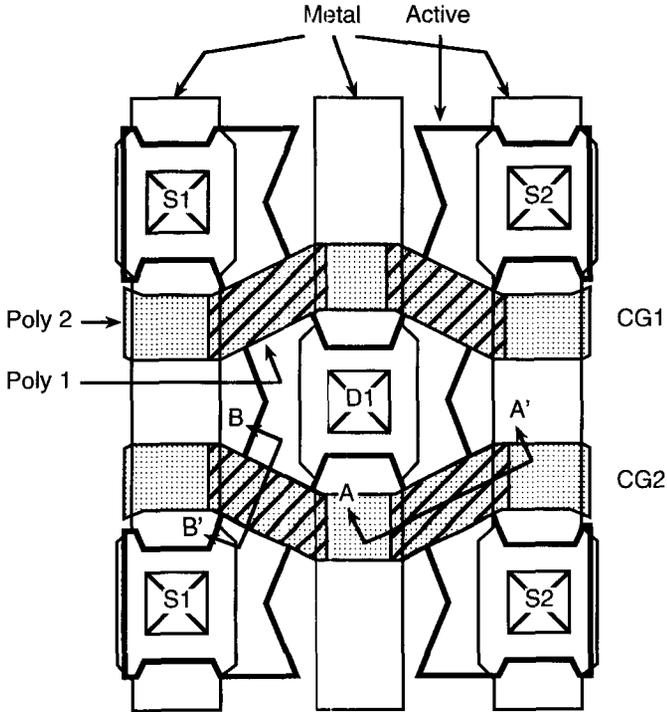


Figure 8.16 Layout view of a 2 × 2 array of the X-cell EPROM. The active region is outlined in the heavy line and forms an “X.”

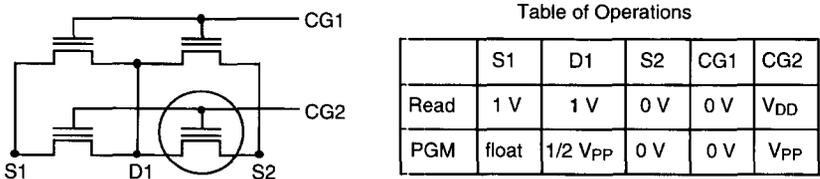


Figure 8.17 Schematic representation and a table of operation for the X-cell matrix shown in Figure 8.16.

removed using the poly 2 pattern as a mask leaving the separated diffusions seen in cross section A1–A1'. A boron implant is used to raise the threshold of these newly exposed regions so that they do not conduct.

The array connections are illustrated in the array schematic in Figure 8.19. Metal lines, which are identified as (M–1), (M), (M+1) in Figure 8.19, strap every other diffusion bit-line, which are shown as the n⁺ regions in the Figure 8.18 cross sections. Strapped diffusion lines are continuous and run along the entire array. These are the lines seen with the symbol \square on them. Nonstrapped diffusion lines are

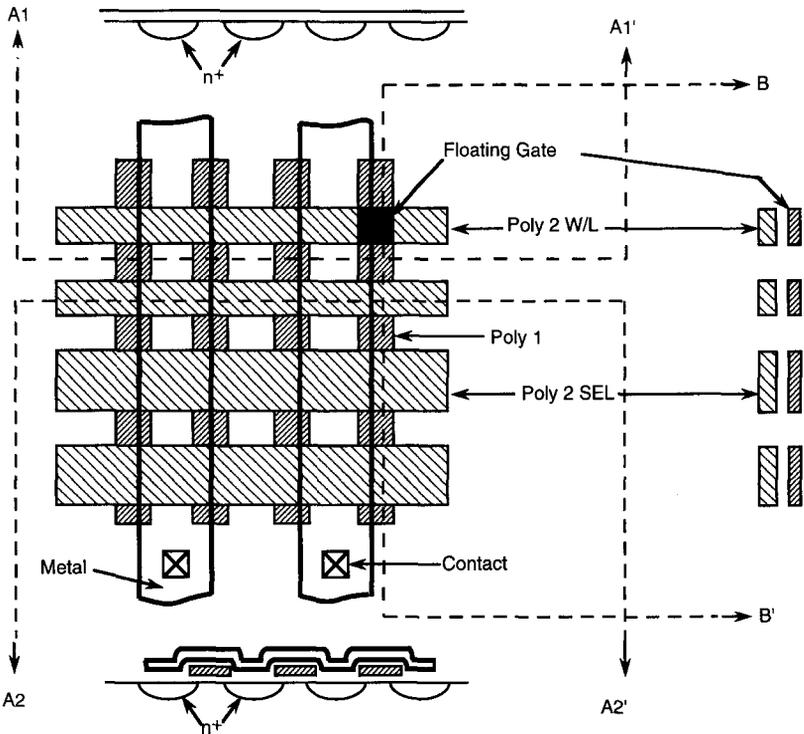


Figure 8.18 Layout view and representative cross section views through a portion of an AMG array. The poly 1 is shown as it would be after the poly 1 etch. After the poly 2 is etched to form the stripe patterns shown, the poly 1 is etched again with the poly 2 mask pattern to leave only small areas shown as the darkened region in the upper right of the array. The field oxide areas are not shown in the interests of clarity. There are no field oxide regions in the array region in which the poly is used to provide isolation, but field oxide is used to separate the contacts outside of the array core.

broken into segments. Each segment contains N cells (typically $N=64$). The segmented bit lines, identified as $(S-1)$, (S) , and $(S+1)$ in Figure 8.19, are electrically connected to metal strapped lines through select transistors. The select transistors are located at both ends of the segmented bit line to reduce the series resistance associated with a selected cell and driven with the lines identified as $SEL(n-2)$, $SEL(n)$, and $SEL(n+2)$. The select transistors can be stacked-gate transistors for process simplicity.

Programming of the floating gate is accomplished under conditions of CHE injection; the programming conditions are presented in the functionality Table 8.1. The current flow is between adjacent diffused bit lines under the selected word line. To access any cell, one word line and one select line are selected together with two metal bit lines.

Because there is no field oxide region over which the floating gate-control gate capacitance can be increased without increasing the floating gate to channel

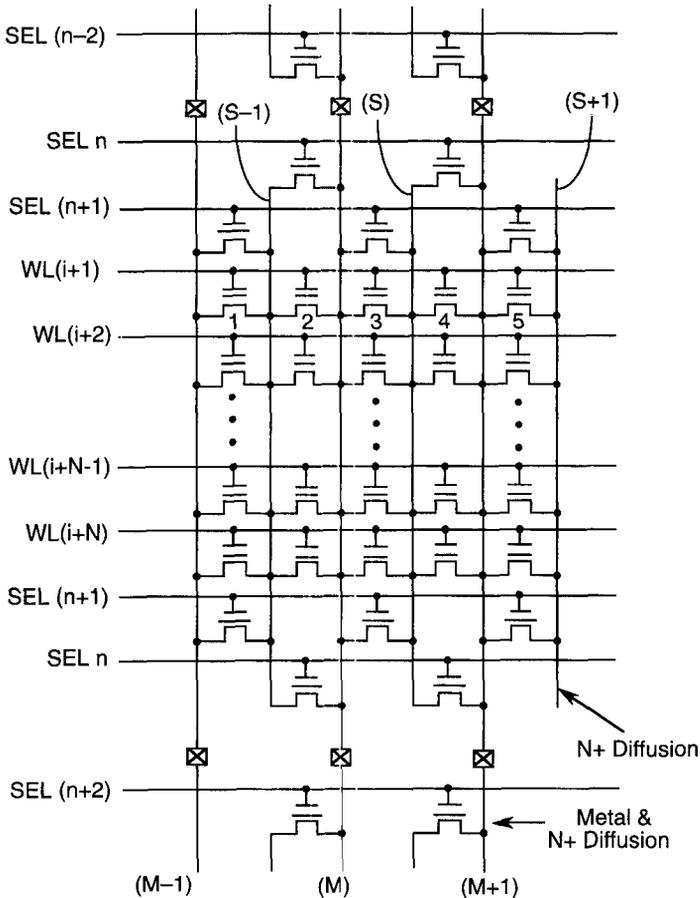


Figure 8.19 Schematic diagram of an AMG array subarray showing the select transistors.

TABLE 8.1 AMG Table of Operation for Programming

OPERATION	SEL (n)	SEL (n + 1)	BL M - 1	BL (M)	BL (M + 1)	WL (i + 1)	OTHER WL
PROGRAM CELL 1	V _{CC}	0 V	7 V	0 V	float	V _{PP}	0 V
PROGRAM CELL 2	0 V	V _{CC}	0 V	7 V	float	V _{PP}	0 V

capacitance, the control-gate coupling ratio is relatively low and the drain coupling ratio is relatively high compared with a T cell. To ensure good charge transfer characteristics, the AMG array is optimized by scaling the effective channel length of the EPROM cells. During programming, a high voltage (~ 7 V) must be sustained on the bit line of the selected cell. The BL-BL (metal-metal) leakage in the array is minimized by having two cells in series (instead of one cell as in a T-cell EPROM

array), for instance, cell 1 and cell 2 between BL (M-1) and BL (M) in Figure 8.19, thus allowing a shorter L_{eff} than in a T-cell array. The L_{eff} scaling helps offset the effects of the drain–source resistance in the array and the lower coupling relative to a T cell. About 0.2- μm effective channel length in 0.8- μm technology has been obtained as the shortest operational L_{eff} .

A major architectural difficulty to integrating a stacked-gate cell in a virtual ground environment is to ensure an inhibit function for the neighboring cells on the word line being selected during programming. In Figure 8.19, if cell 2 is being programmed, then cell 3 experiences a high voltage on the drain and thus, by design, the voltage drops across cell 3, cell 4, and so on, remain small enough to avoid any spurious programming throughout the programming cycle. This is achieved by keeping the unselected bit lines floating during the programming cycle. The floating bit line capacitor is charged by the current flowing through cell 3 and cell 4 on the selected word line in about 100 ns. Note that cell 3 has its source connected to a segment of diffused line which has a very small capacitance (<0.5 pF) and is charged up very quickly. By design, both the bit line and word line voltages are ramped up slowly (by a few hundred nanoseconds).

The functional selection scheme for read is the same as program, but with lower voltage levels as shown in Table 8.2. During the read cycle, all bit lines are precharged to about 2 V. The bit line that is acting as source is discharged and the presence of word line and select line voltages help develop the read current of an erased cell, thus discharging the selected bit line. The programmed cell does not discharge the bit line and this potential difference is detected with the sense amplifier.

Not all EPROM cells have the same profiles along the channel as the T cell. In particular, some memories employ what is known as a *split-gate cell*. The split gate is combined with a virtual ground in the approach that has been named the staggered virtual ground (SVG) (SVG is a registered trademark of WSI) array.³⁹ This approach is shown in layout and cross section in Figure 8.20. As the cross section B–B' shows, this approach has isolation areas in the core of the array. The cross section A–A' illustrates two aspects of this split-gate cell. One thing to note is that the control gate forms a series transistor on the source side of the memory gate; another is that there is a fairly large overlap of the floating gate on the bit line diffusion region that acts as a drain to this transistor. This diffused bit line is formed by an ion-implantation step so that it is self-aligned with the floating gate. The gate–drain overlap capacitance can be used to capacitively couple the floating gate to higher potentials for

TABLE 8.2 AMG Table of Operation for Reading

OPERATION	SEL (n)	SEL (n + 1)	BL (M - 1)	BL (M)	BL (M + 1)	WL (i + 1)	OTHER WL
READ CELL 1	V _{CC}	0V	2V	0V	2V	V _{CC}	0V
READ CELL 2	0V	V _{CC}	0V	2V	2V	V _{CC}	0V

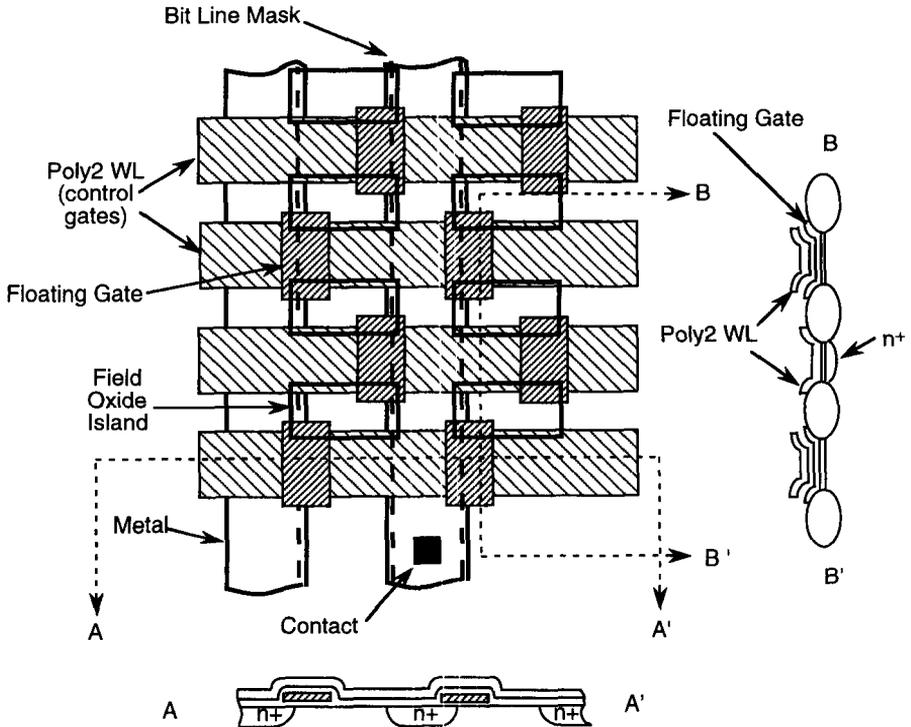


Figure 8.20 Layout and cross section views of a SVG array.

programming because the series transistor prevents punchthrough on the unselected rows.

This is a virtual ground array without contacts in the array core to save area. The asymmetric structure of the split gate cell is key to achieving good program inhibit characteristics, which leads to simpler decoding of the array, thus preserving the advantage of a smaller array size. The split-gate structure also allows scaling of the effective channel lengths, L_{eff} , of the floating gate.

To program a cell, one word line and two bit lines are selected (one as drain and one as source). Since the orientation of the cells on adjacent word lines are opposite of each other, every bit line could be selected as a drain or as a source. The conditions required in the array are summarized in Table 8.3 (see also Fig. 8.21). The split-gate cell has inherent immunity to reverse programming (“disturb” of cell 3 when cell 2 is being programmed). This is because the drain of the floating gate is exposed to a gradual lateral potential extending along the channel of the adjacent select device.

The functional selection scheme for a read cycle is same as programming, but with lower voltage levels as shown in Table 8.4 (again, see Fig. 8.21). During the read cycle, all bit lines are precharged to about 2 V. The bit line that is acting as the source is discharged and the bias on the selected word line couples enough potential onto an

TABLE 8.4 Operating Conditions of the SVG Array during a Read Cycle

OPERATION	BL (M-2)	BL (M-1)	BL (M)	BL (M+1)	WL (i+1)	WL (i+2)	OTHER WL
READ CELL 1	0 V	2 V	2 V	2 V	V_{CC}	0 V	0 V
READ CELL 2	2 V	0 V	2 V	2 V	V_{CC}	0 V	0 V
READ CELL 6	2 V	0 V	2 V	2 V	0 V	V_{CC}	0 V
READ CELL 7	2 V	2 V	0 V	2 V	0 V	V_{CC}	0 V

erased cell to cause read current to flow, thus discharging the selected bit line. A programmed cell does not discharge the bit line. This difference in voltages is detected with the sense amplifier.

8.3.2 Byte-Alterable E²PROMs

In read mode, electrically erasable programmable read only memories (E²PROMs) are very similar to EPROMs. A heavily doped polysilicon field plate that is completely insulated from all other electrodes and silicon by high-quality insulators forms a floating gate for a MOSFET. Typically, one or more electrodes are coupled capacitively to the floating gate. Electron conduction of the transistor channel can be controlled by the charge stored on the floating gate. This charge modulates the channel current in the floating gate transistor and can be sensed as discussed previously.

The major difference between an EPROM and an E²PROM is that there are electrical elements in the E²PROM structure that allow the transport of electrons both to and from the floating gate under high bias voltages. The E²PROM transport process is Fowler–Nordheim (FN) tunneling from planar structures or enhanced FN tunneling from nonplanar structures (e.g., textured poly). These processes were discussed in Section 8.2.1. Memory cells are designed to couple the necessary high voltages for charge transport across the tunneling element over intervals of milliseconds. The transport of electrons results in a change of potential in the floating gate, which reduces the electric field for tunneling. Therefore, the electron transport process is a self-limiting process. Because the floating gate is conductive, it is possible to have charge transported in a small area away from the channel region.

It is relatively easy to design a single memory cell that has two memory states. It is a very different problem to put the cell in an array and control the cell program and erase conditions so that no disturb of other cells occurs. Applying voltage in one way to program a cell in the array would usually mean that another cell would be erased in the array. Special inhibit conditions must be designed into the cell. These considerations are discussed in the following sections.

Tunnel Oxide Defined Lithographically in Drain Extension The generic name for this memory cell is (*floating-gate tunnel oxide*) (FLOTOX), but variations of the

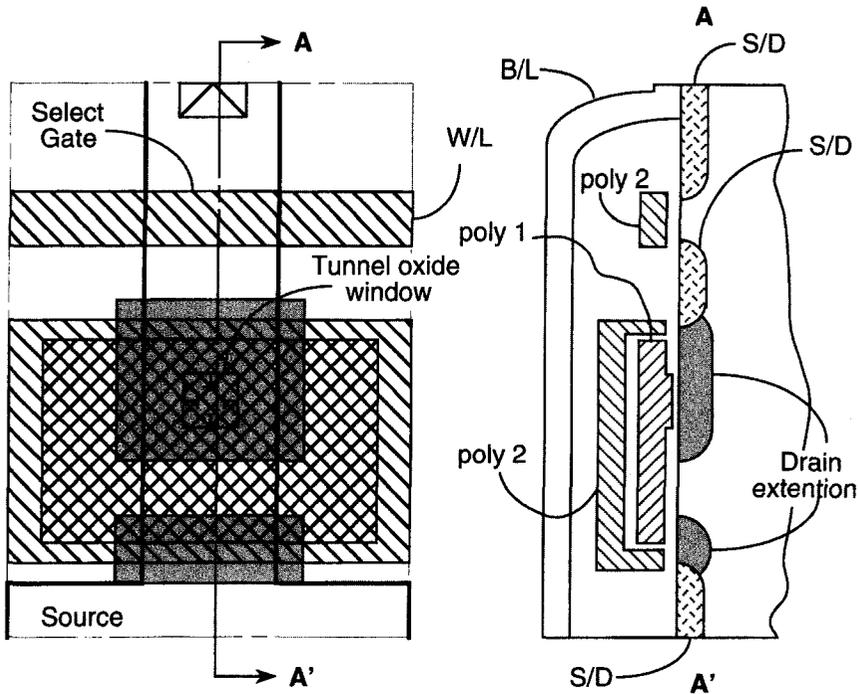


Figure 8.22 Layout view and cross section through a FLOTOX memory cell.

basic cell have been reported and used for E^2 PROM products.^{40–42} The basic structure of the FLOTOX memory cell is shown in Figure 8.22. The tunnel dielectric (which is typically less than 12 nm thick) is grown over an area defined lithographically in an extension of the drain region, which is usually more lightly doped than the drain region. When the drain extension is in-line as is shown in Figure 8.22, it is usual to include a similar diffusion on the source end of the memory transistor so that the channel length is alignment-independent. A second layer of polysilicon forms the control-gate capacitor and gate of the row-select transistor.

The operation of the cell in an array is most easily understood by reference to Figure 8.23. An array operates on words, typically of one byte. The cells in the byte may be programmed by applying high voltages to the byte select line and the word line. If a voltage V_{PP} is applied to the W/L, a voltage ($V_{PP} - V_t$) can pass through the byte select transistor to bias the control gates. This voltage is usually chosen to be in the range 14–18 V, depending on the tunnel oxide thickness and the control gate coupling ratio, $R_{CG} = C_{CG}/C_{TOT}$. The cell is usually designed such that $R_{CG} \geq 0.6$. The B/L bias is maintained at 0 V. The resultant field across the tunnel oxide is sufficiently high that electrons will tunnel through the thin oxide via FN tunneling to charge the floating gate negative.

The floating gates within a byte can be selectively erased by applying high voltages to the W/L and the selected bit lines with the byte line biased at ground.

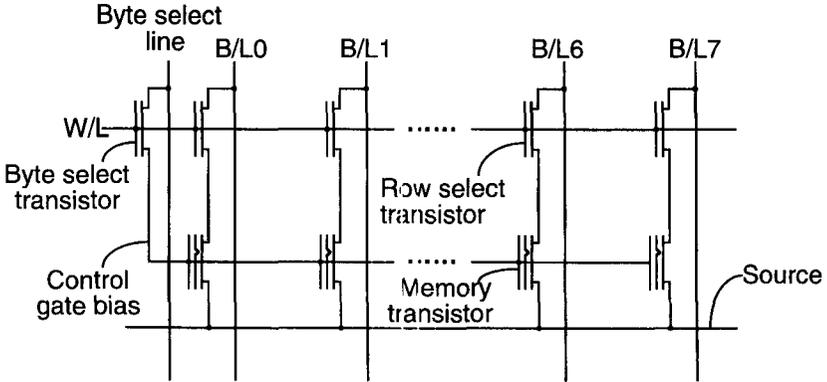


Figure 8.23 Schematic diagram of one byte of a FLOTOX E²PROM array.

Again, a bias of V_{PP} on the W/L will allow $(V_{PP} - V_i)$ to appear on the drain extension of the selected bits. Usually, $\geq 70\%$ of the bias applied to the drain extension is coupled across the tunnel oxide. This bias causes the electrons to tunnel from the selected floating gates to the drain extensions, selectively erasing these gates. The source is allowed to float during the erase operation. Because all the gates in the selected word were previously programmed, the transistors on the unselected B/Ls do not conduct with the control-gate at ground. This prevents any current flow from the selected bit lines that are at a high voltage to the unselected bit lines that are at ground or are floating.

Writing an arbitrary pattern to a byte is a two-step process. All bits of the byte are first programmed; then selected bits within the byte are erased to set the byte to the desired pattern. Note that because of the operation sequence, there are no static current paths in the write operation. The current that must be supplied is very low; it is only the peak tunnel current, typically ~ 0.1 nA/cell, and the leakage currents of the various junctions. This allows the necessary voltages to be generated on chip from the V_{DD} supply. Usually all of the timing and control functions are also supplied, so that this memory appears as simple to write to as an SRAM.

The inhibit functions are provided by the addition of transistors to the memory transistor. The unselected rows are isolated in the read, program, and erase modes by the byte- and row-select transistors. The bytes within the selected row can be further isolated depending on which byte select lines and bit lines are biased.

Textured Poly E²PROM Cell Another E²PROM cell used commercially is the textured poly cell. This cell is shown in layout and cross section in Figure 8.24. The cell utilizes three layers of polysilicon. The oxide that separates these layers of polysilicon is much thicker than the tunnel oxide of the FLOTOX cell. The interpoly oxides on the textured poly arrays are typically >500 Å thick compared to the ~ 100 Å thickness of the FLOTOX tunnel oxide. The voltages applied to write in the two approaches are comparable because the electric fields on the top surfaces of the polysilicon is geometrically enhanced as was discussed with respect to Figure 8.8.

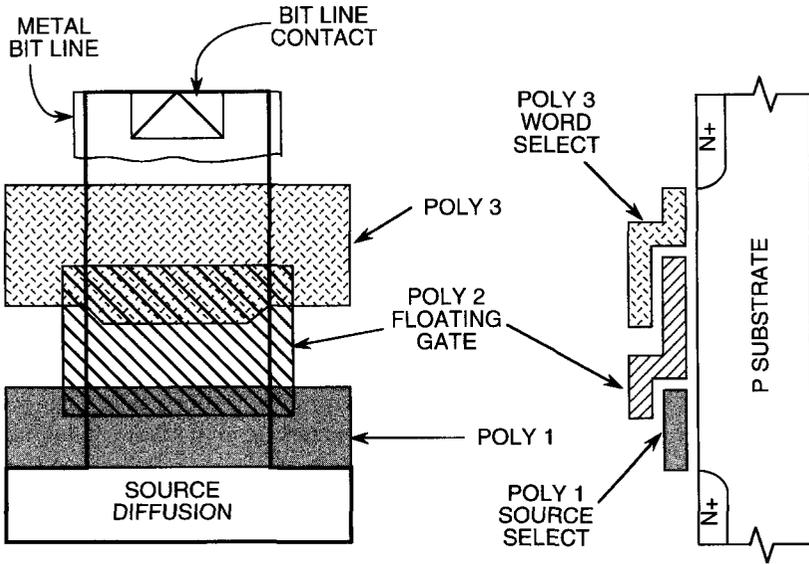


Figure 8.24 Layout and cross-sectional views of a textured poly E²PROM cell.

As discussed in Section 8.2.1, the *I-V* characteristics for geometrically enhanced tunneling are very asymmetrical with respect to bias polarity. Because of the asymmetrical tunneling behavior of the textured poly, the cell depends on electrons tunneling from the poly 1 electrode to the poly 2 floating gate for programming and tunneling from the poly 2 floating gate to the poly 3 electrode for erase.

Figure 8.25 shows a schematic diagram of the textured poly cell of Figure 8.24. As will become apparent in the discussion of the operation of this cell, the transistors in the cell provide sufficient isolation that the bias on the surrounding cells is not

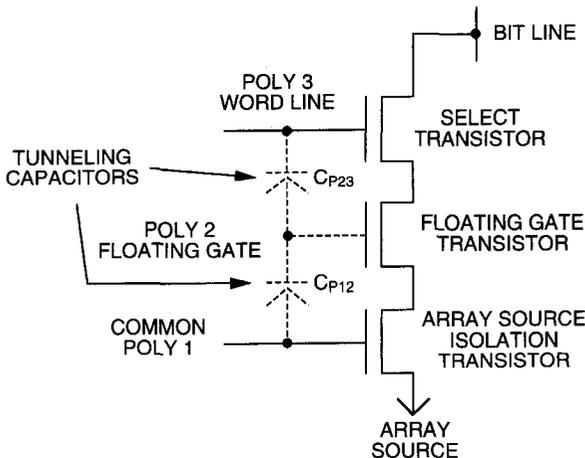


Figure 8.25 Schematic diagram of the textured poly E²PROM cell.

important. The symbols labeled C_{P12} and C_{P23} in Figure 8.25 are tunneling capacitors. The symbols are intended to indicate that the poly–poly overlap creates a capacitor through which electrons may flow more easily in one direction under high electric fields.

Rather than having separate program and erase operations like the FLOTOX cell, the textured poly cell has only a write operation. During the write operation, the ground-select line is biased at ground to isolate the floating gate transistor channel from the array source. The poly 3 word-select line is biased at V_{PP} . The potential on the floating gate depends on the bias on the bit line, which acts to steer the cell to either program or erase.

Assuming that the bit line is biased at ground, the capacitance between the channel and the floating gate couples the floating gate to a low potential and most of the applied voltage appears across the oxide between the floating gate and the poly 3 word-select gate. This bias causes electrons to tunnel from the floating gate to the word-select gate. If the floating gate was programmed prior to this write operation, the negative charge of the electrons on the floating gate increases the field and tunneling occurs, which erases the gate. If the floating gate were erased prior to the write operation, the positive charge on the floating gate would reduce the field and little tunneling would occur.

On the other hand, assuming that the bit line is biased at V_{PP} , the voltage on the bit line acts via the channel–floating gate capacitance to couple the floating gate to a high potential so that most of the voltage drop occurs between the poly 1 source-select transistor and the floating gate. Again, the charge on the gate prior to the write event acts to increase the field to program an erased gate. A floating gate that is already programmed is merely refreshed.

To read the textured poly cell, the source-select transistor is biased at V_{DD} to provide a current path between the floating gate channels and the source. The row-select line is biased at V_{DD} to select the desired row. The bit lines on the selected columns are biased to a modest voltage, typically in the range 1–2 V. If the floating gate is programmed, the channel of this transistor will be cut off and there will be no current flow. If the floating gate is erased, the channel will conduct and current will flow from the bit line to the source which will tend to pull the bit line low.

The reader can see how the channel potential in the textured poly cell acts to steer the floating gate to either program or erase. Inhibition for unselected rows are provided by the word- and source-select transistors. During read mode, the selection of the desired columns is provided by column select circuitry. Disturb is prevented in the write mode by initiating a write by loading the data present in the selected row into latches. The latches are updated as desired before the data in the latches are written back into the row. Because the write operation only reinforces the cells that contain the data being written, the unaltered cells experience little stress.

8.3.3 Flash EEPROM Memories

Flash EEPROMs were born out of demand for less expensive electrically rewritable memory than could be provided with byte-rewritable E²PROMs. The approach was

to abandon the feature of byte alterability in favor of simpler devices. By allowing one of the write operations to be a block operation, it is not necessary to provide inhibition for one of the write operations. This extra freedom has been exploited in a wide variety of cells and arrays. Obviously, the byte-writable cells described previously can be simplified slightly, which has been done, but the real area reductions have been achieved with novel approaches. Some of the approaches utilize hot electron injection for programming and FN tunneling for erase; others use FN tunneling for both programming and erase. Some of the approaches that have met with the most commercial success will be described in the following sections.

T-Cell Flash EEPROM

The cell structure, array architecture, operation and reliability of T-cell flash memory has been discussed extensively.⁴³⁻⁴⁶ The one-transistor, T-cell, flash memory array has evolved from the conventional T-cell EPROM technology. The top view and cross section through the transistor parallel to the word line are the same as that shown in Figure 8.14. The gate oxide is thinned to ~ 10 nm to allow the cells to be erased by FN tunneling from the floating gate to the substrate or the source. For some erase schemes, the source diffusion is modified as discussed in the following paragraphs.

The cells are programmed the same way as the T-cell EPROM is programmed. Because the flash memories are usually intended for applications involving in-system programming, it is common for the memories to have circuitry to implement algorithms for efficient programming and threshold margin measurement designed into the memory chips rather than relying on intelligence in a programming system as is the usual case for EPROMs. The read operation and design considerations for T-cell flash memory arrays also mirror those for T-cell EPROM memory arrays.

Erase can be accomplished by causing electrons to tunnel either to the source or to a p well.⁴⁷ Tunneling to the source is very nonuniform, while tunneling to a p well provides uniform erase. There are advantages and disadvantages to both approaches.

Tunneling to the source involves biasing the source to a high positive potential with respect to the floating gate so that a high field is established across the gate oxide separating the junction from the floating gate. This causes a high field in the silicon in the vicinity of the junction as illustrated in Figure 8.9. If the junction is abrupt, there is a real possibility that the high field in the silicon will lead to gate-aided breakdown. The resulting currents would be too large for the intended mass operation. To avoid gate-aided breakdown, common practice is to provide a graded source junction, as illustrated in Figure 8.26a.

Nonuniform erase can be accomplished by either applying a bias of ~ 12 V to the source with the control gate grounded or by applying ~ 5 V to the source with the control gate biased at about -10 V.⁴⁸ The drain is allowed to float in either case. The advantage of the negative gate erase is that the source current is supplied from a low voltage supply, often directly from V_{DD} . The significance of this is that the source current during erase can be as large as 1 nA/cell. Although 1 nA may seem like a small current, when cumulated over 1 M cells, the current is ~ 1 mA. Generating 1 mA from an on-chip charge pump is difficult; consequently, negative gate erase is

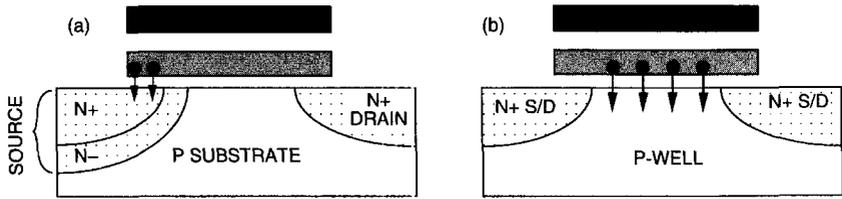


Figure 8.26 Cross sections parallel to the channel current flow illustrating the T flash cell operating with (a) nonuniform erase and (b) uniform erase.

essentially mandatory where in-system erase is desired from a single logic-level supply. The disadvantage of negative gate erase is the extra chip area required for the row decoders, which must support both negative and positive voltage signals.

Note that the sum of the biases for the negative gate erase is larger than the bias applied for the grounded gate erase. This is because the channel capacitance is reduced in the grounded gate erase case by the depletion region that forms under the floating gate.

A T-cell flash array can also be erased by applying about -9 V to the control gate and about $+7$ V to the p well with the source and drain both floating. Under these bias conditions, the channel under the floating gate is in accumulation; a large vertical field is established between the holes accumulated at the oxide silicon interface and the floating gate. Because both source and drain junctions float, there is no significant field established between the S/D junctions and the p well; graded junctions are not required. However, to bias the p-type body of the transistors at a negative bias, a triple-well CMOS process, in which the memory transistors are formed in p wells created in an n-type substrate, must be employed, which results in additional process complexity.

Erase algorithms are often implemented in silicon for the same reasons that the program algorithms are. The erase and program algorithms essentially consist of short write pulses followed by read cycles to verify if the write function has been accomplished. The program pulses are typically ~ 1 μ s and the erase pulses typically ~ 10 ms. The time to program a byte is <10 μ s for memories from most manufacturers. The time to erase a byte or block is typically ~ 1 s.

The individual program pulses can occupy a greater fraction of the program time than can the erase pulses of the erase time because program threshold is a less critical parameter. As is shown in Figure 8.13, the program threshold is unbounded on the high side. The only requirement is that the threshold be greater than the bias applied to the control gate during read. The erase threshold is bounded on two sides. If the erase threshold is too high, the read current will be low and the read access time will be slow. If the read current is too low, the sense amplifier may detect the incorrect information. On the other hand, if the erase threshold is too low, the cell may conduct during read even when deselected, which would lead to erroneous reading of the data from other programmed cells on the same column. In order to tighten the distribution of erase thresholds sufficiently for a practicable memory, all the bits must be programmed prior to erase.

A factor that can complicate the erase operation of a cell without a select-gate per row is an effect called “erratic erase.”⁴⁹ This effect results in a cell suddenly erasing much faster than normal as is illustrated in Figure 8.27.⁵⁰ The enhanced erase tunneling illustrated in Figure 8.27a is not, in general, a permanent effect; rather, after some indeterminate number of further program/erase cycles, the erase tunneling will revert to the “normal” tunneling characteristic that it had initially; hence the name “erratic.”

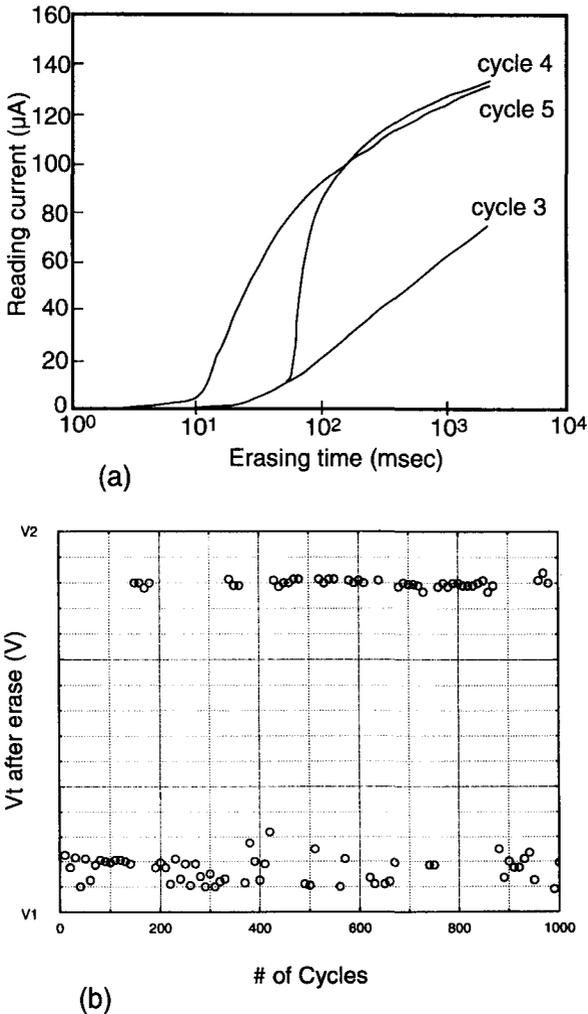


Figure 8.27 (a) The current for reading is shown as a function of erasing time on three successive erase cycles. The dramatic increase in reading current after about 5 ms on the fourth read cycle marks the onset of the enhanced erase tunneling. (b) The measured erase threshold as a function of cycles for an erratic bit.

Erratic erase is believed to be a result of trap-assisted tunneling through traps that are created as a result of the holes injected into the oxide during the erase process.^{51,52} These traps may be subsequently filled and thereby deactivated only to later be emptied and reactivated. Erratic bits have been observed to switch back and forth between normal tunneling and enhanced tunneling multiple times, as is shown in Figure 8.27b.

When the flash T-cell is integrated into an array, the selection for reading and programming is exactly the same as for a T-cell EPROM. Erase selection depends on segmentation of the erase blocks, such as the segmentation of the source connections for grounded gate erase.

Field-Enhanced Tunnel Injector Flash EEPROM Cell

The field-enhanced tunnel-injector flash (FETIF) EEPROM cell has a T-shaped active area and relies on hot-electron injection programming and FN tunneling erase,⁵³ but is quite different in operation from the T-cell flash EEPROM. The cell is shown in cross section and layout view in Figure 8.28. Note that the poly 2 lies partially over the poly 1 floating gate and partially over the gate oxide, where it forms a transistor in series with the floating gate transistor. The poly 2 forms a portion of the

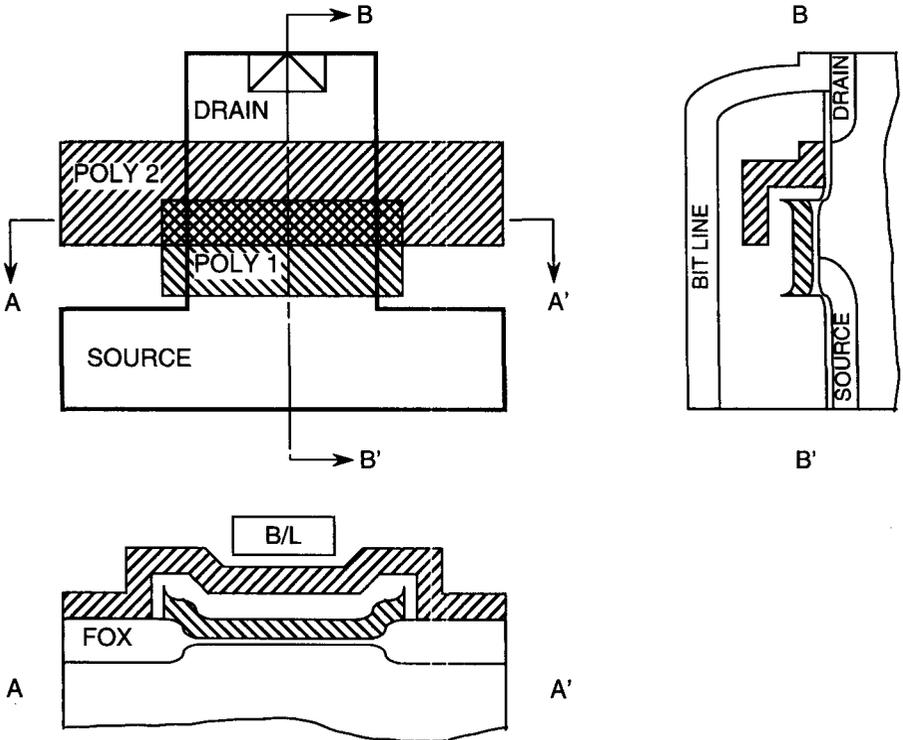


Figure 8.28 Layout and cross-sectional views of the field enhanced tunnel injector flash EEPROM.

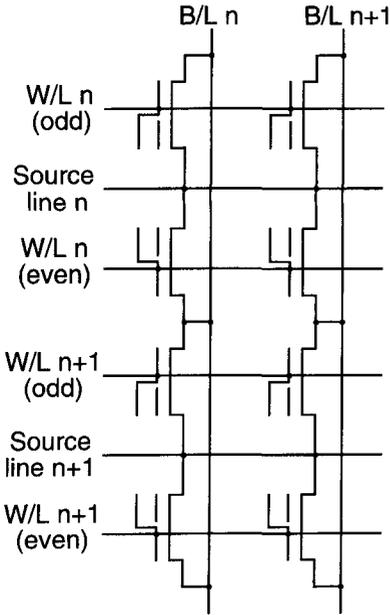


Figure 8.29 Schematic diagram of an array of FETIF cells.

word line; the poly 2 series transistor acts as a word line select transistor. Nominal values of oxide thickness for a 1- μm -scale process are floating-gate channel oxide = 15 nm; select-gate channel oxide = 40 nm; floating-gate-to-control-gate sidewall oxide = 40 nm; floating-gate-to-control-gate top oxide = 200 nm. The metal bit line is shown in the cross-sectional drawings but omitted from the layout drawing for clarity.

A schematic diagram and table of operating biases of an array of FETIF cells are shown in Figure 8.29 and Table 8.5, respectively. A major difference between the FETIF cell and most other nonvolatile memory cells is that the oxide between the poly 2 word line and the poly 1 floating gate is relatively large so that the capacitive coupling between poly 1 and poly 2 is relatively small. The largest capacitive coupling to the floating gate is that of the source because of the large overlap of the floating gate on the source diffusion.

The FETIF array operates by erasing a block of cells to precondition them for subsequent selective programming. The minimum erase block consists of a pair of

TABLE 8.5 Operating Biases of a Selected FETIF Cell in an Array

	ERASE	PROGRAM	READ
WORD LINE	$\approx 15\text{ V}$	V_{t+}	V_{REF}
BITLINE	V_{SS}	“1” $\approx V_{\text{DD}}$; “0” $\approx V_{\text{SS}}$	$\approx 2\text{ V}$
SOURCE LINE	V_{SS}	$\approx 12\text{ V}$	V_{SS}

word lines that bracket a common source diffusion. In Figure 8.29, these are referred to as the even and odd word lines associated with a source line. Erasing takes place by applying a bias of ~ 15 V to the word lines in the erase block while all of the bit lines and the source lines associated with the selected word lines are grounded. With the B/Ls and source lines grounded, an inversion layer will be formed in the floating gate channel as soon as the floating gate is coupled above its native threshold. The capacitances of the floating gate to the inversion layer and the source overlap region dominate the total floating gate capacitance so that most of the applied W/L bias appears across the interpoly oxide. The poly 1 floating gate is oxidized in a way that encourages the formation of sharp edges along the edge of the poly 1 as shown schematically in Figure 8.28. The field enhancement along these edges allows tunneling through relatively thick interpoly oxide as discussed with respect to Figure 8.8. Because the tunneling is interpoly, high voltage is not applied to junctions in the array; hence, there is no band-to-band tunneling. The 15-V bias can be generated with a low-power, on-chip charge pump.

Cells are selectively programmed by applying about 12-V of bias to one of the source lines. One of the word lines neighboring the selected source line is biased at ground, which prevents current through cells on that row. The other word line is biased slightly above threshold for the drain-select transistors. The unselected bit lines are biased at V_{DD} ; the bit lines containing the cells to be programmed are grounded, which allows current flow through the cells on the selected row and bit lines. The large source underlap diffusion couples the floating gate upward in potential. Since the channel of the floating gate transistor is essentially depleted and the interpoly oxide is relatively thick, the source to floating gate coupling ratio ($R_{S_FG} = C_{S_FG}/C_{TOT}$) is high, typically $\sim 80\%$. This means that the 12-V source bias increases the potential on the floating-gate by ~ 9.6 V. These are the conditions described with reference to Figure 8.5 for efficient source side injection. The channel current can be limited by the bias on the poly 2 transistor to < 1 μA , which allows the 12-V bias to be generated with an on-chip charge pump.

In the read mode, the source lines are all grounded. The selected row line is biased at a voltage that turns on the select transistor, for example, V_{DD} , and the selected bit lines are biased at ~ 2 V. Those cells that are programmed don't conduct while those that are erased conduct. Note that because of the existence of the drain select gate, the cells may be erased into depletion as indicated in Figure 8.13a, which is advantageous for reading under low supply voltage conditions.

Triple-Poly, Virtual Ground (TPVG) Flash Cell

The TPVG flash also utilizes hot-electron injection for programming and FN tunneling for erase as do the two flash cells discussed previously, but differs from each of these cells in two important ways. This approach (1) employs a virtual ground and (2) uses CHE injection for programming and geometrically enhanced tunneling for erase, so that it differs from each of the two previous approaches in one of the write modes.

A layout view and cross-sectional drawings through the cell along the channel A–A' and across the channel of the TPVG cell are shown in Figure 8.30.^{54,55} This

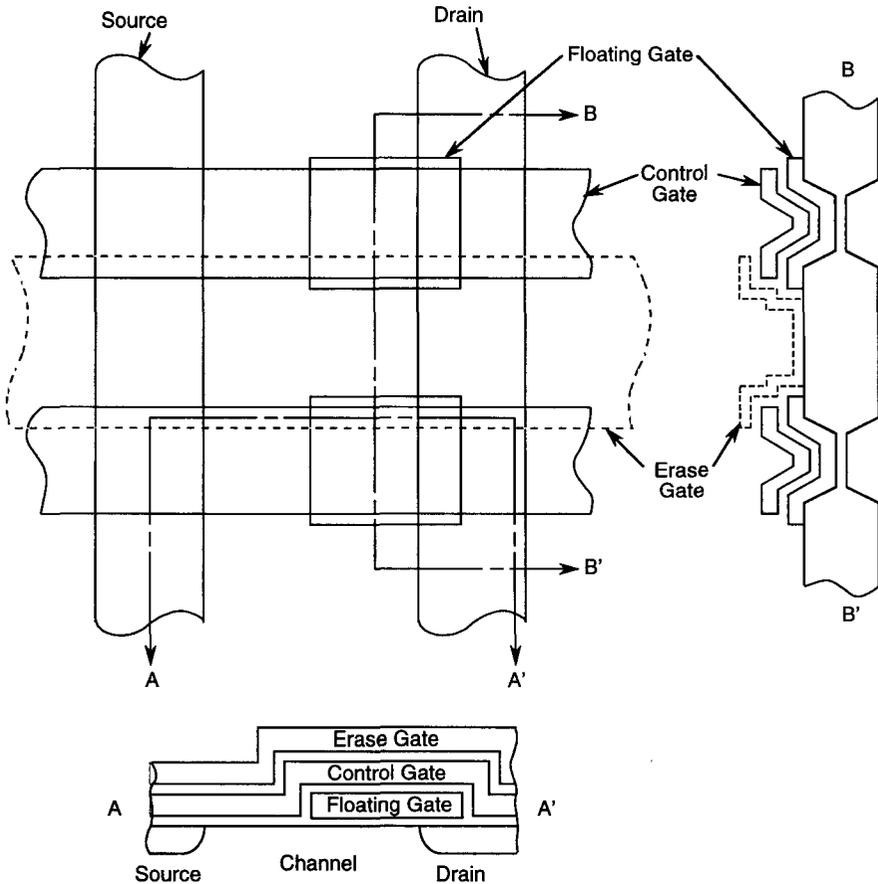


Figure 8.30 Layout and cross-sectional views through a triple-poly, virtual ground flash cell.

physical view can be seen in the context of the circuit schematic view shown in Figure 8.31. Comparison of these figures with Figures 8.20 and 8.21 for the SVG cell shows that the schematic diagrams and the views along the channel are the same for the two arrays if the erase gate is ignored for the TPVG approach. Indeed, these two approaches operate in the same manner in the read and program modes. The difference is in the erase mode in which the erase gate of the TPVG approach is biased to a high voltage with all other nodes grounded to cause electrons to tunnel from the edge of the floating gate to the erase gate in the same way as the FETIF approach. Because the TPVG approach is intended for mass storage applications, there is no attempt to provide small block erase as the FETIF approach does. Although it is not important from a device physics point of view, the TPVG approach has adopted some circuit design features to further optimize it for mass storage applications. These are described in the references.

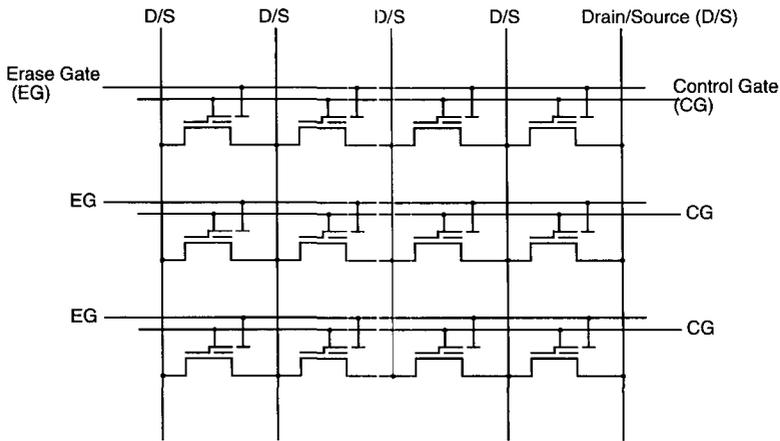


Figure 8.31 Schematic diagram of a portion of a triple-poly, virtual ground flash array.

NAND Cell

In mass storage applications, cell area is very important. A major limitation to reduction of the cell area of the T-cell flash EEPROM is the area required for the drain contact and the required spacing of this contact from the gate electrodes. One way of reducing the cell area is to eliminate this spacing. The NAND architecture minimizes the area needed for the drain contact by sharing the contact over a number of cells as is illustrated in Figure 8.31.⁵⁶⁻⁵⁸ As can be seen in Figure 8.32, the NAND architecture sandwiches a number of memory transistors, in this case 8, in series between a pair of select gates. These transistors all share a common source and a common drain and can each store a bit of information. Because the drain and source are shared by a number of memory transistors, the effective area for drain and source *per memory transistor* is significantly reduced. The memory transistors are composed of a floating gate formed in a first layer of polysilicon that is self-aligned to the overlying polysilicon trace in the direction of current flow. The dielectric between polysilicon layers is a sandwich of silicon dioxide/silicon nitride/silicon dioxide, often referred to as “ONO,” as is usually the case for scaled floating gate transistors.

A cross section of the structure of a NAND cell and the surrounding peripheral circuitry is shown in Figure 8.33. The memory arrays are placed in a p well formed in an n-type substrate. The NAND memory transistors program and erase via uniform Fowler–Nordheim tunneling from the substrate. This write mechanism, combined with the small read current allows the use of lightly doped source–drain regions in the stack of memory transistors. The NAND array and the peripheral circuitry are placed in different p wells that are electrically isolated from one another. The two isolated wells allow the memory wells to be biased at a high positive voltage in order to erase the floating gates while the p well for the periphery is biased at ground to provide for proper circuit operation.

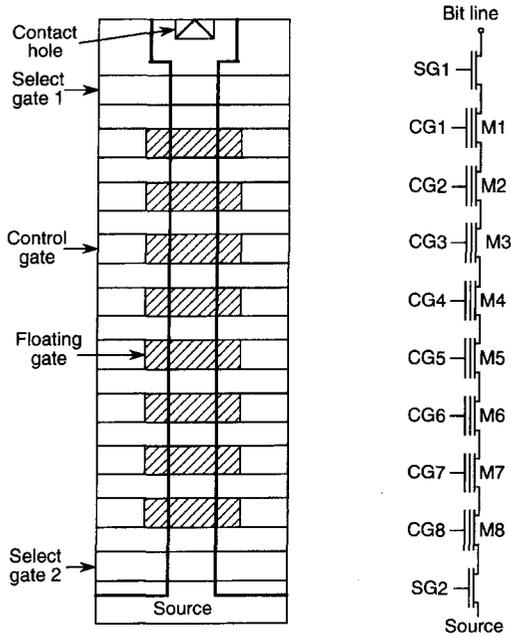


Figure 8.32 Layout view and schematic diagram of the NAND architecture side by side to illustrate the correspondence between the elements.

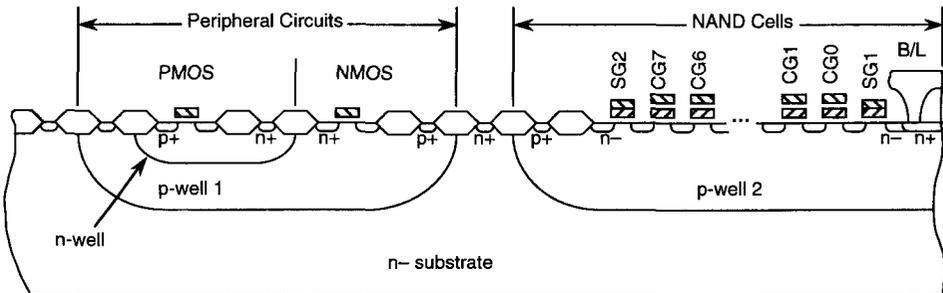


Figure 8.33 Cross section of NAND cell and accompanying peripheral circuitry.

The conditions for proper operation of the array are given in the table in Figure 8.34. The erase operation is implemented by biasing the p well 2 at 20 V with the control gates of the memory blocks that are to be erased biased at ground. The source and drain junctions float so that there is no significant potential difference between the well and the source and drain junctions. Consequently, gate-aided breakdown and band-to-band tunneling are of no concern.

The control gate lines in the cell stacks that are not to be erased are biased at 20 V so that there are only small potential differences between the floating gates and p

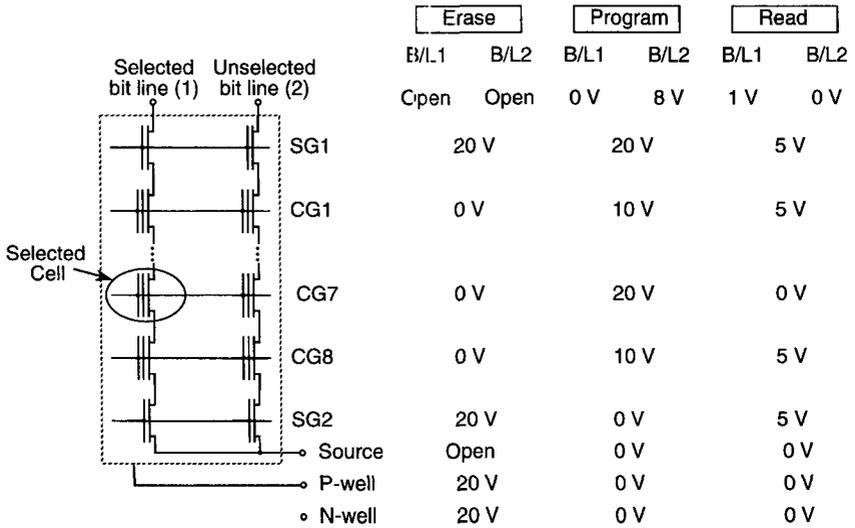


Figure 8.34 Schematic drawing of a portion of a NAND array with its table of operation. The table has been arranged so that the entries are aligned with the many traces in the array for simplicity.

well 2 for these stacks. The minimum erase block size is equal to the number of memory transistors between the two select gates times the number of columns in a row. In a typical 16-Mbit memory device the minimum erase block size is typically 4 K. The typical erase cycle requires 10 ms. It is not necessary to program all of the bits prior to erase.

The memory can be selectively programmed by tunneling electrons from an inversion layer in the channel to the selected floating gates. In the program mode, the drain-select gate is biased at 20 V and the gate of the second select transistor is biased at ground. The drain is biased at ground on the selected columns and at 8 V on the unselected columns.

Under these bias conditions, the selected memory transistor experiences a bias of 20 V on its control gate with a channel potential of 0 V. This bias is large enough to cause electrons to tunnel from the inversion layer in the channel to the floating gate. Tunneling is inhibited on the unselected bit lines on the selected row because the inversion-layer potentials in the channels of these devices is 8 V, which reduces the field across the gate oxide. The unselected control gate lines are biased at 10 V. This value is high enough to ensure that the unselected transistors on the unselected bit lines are biased to a state that allows the potential to propagate along the stack of transistors and low enough that the potential between the unselected control gates and the selected channel is too small to cause significant tunneling to the floating gates of these transistors. Because the oxide tunneling is uniform with no band-to-band tunneling, the current required for programming is low enough to allow many bytes to be programmed in parallel, such as the byte-alterable E²PROMs discussed previously.

To read a selected cell, all the transistors in series with the selected cell must be biased to a state in which conduction is possible. To accomplish this, the two select gates and all the unselected control gates are biased at V_{CC} . The selected cell is interrogated by biasing the control gate of this cell at 0 V while the drain of the stack is biased at ~ 1 V. To function properly, the erased memory transistors must have depletion thresholds such that they conduct with the control gate grounded. Meanwhile, the programmed cells must have enhancement thresholds that are high enough to prevent significant current leakage through a programmed transistor with the control gate grounded, but low enough that the memory transistors with the unselected control gates at V_{CC} are biased into conduction. Since the current through an erased memory transistor must flow through the series combination of the grounded-gate depletion transistor and at least seven enhancement transistors with their control gates biased at V_{CC} , the read current is low, $\sim 1 \mu\text{A}$. The worst case is that for reading an erased transistor nearest the drain contact with all the other memory transistors in the stack programmed because the voltage drop across the programmed transistors provides source bias for the erased transistor. The low read current allows random access times of $\sim 10 \mu\text{s}$. NAND memories are typically organized to read multiple bytes in parallel into a shift register so that serial-read data rates of 10 Mb/s are achieved.

DiNOR

Another architecture that attempts to improve on the area efficiency of the T-cell flash memory by reducing the area allocated to the drain contact is the divided bit line NOR flash architecture, DiNOR.⁵⁹ This approach uses a polycide minor bit line, seen in Figure 8.35, which requires less space than the metal bit line contact of the T-cell flash device. The reduction in area is large enough that, even with the addition of the select transistors for the minor bit lines, the array is smaller than a T-cell array of the same bit density designed with comparable design rules. Like the NAND approach, this architecture uses a triple-well CMOS process to support negative voltages for writing.

Both the word lines and the source lines are strapped in first-level metal to reduce the word line delay and to provide low resistance for the common source lines. The major bit lines are bussed through the array in a second layer of metal that connects to the n^+ drain of the minor bit line selection transistors via tungsten plugs as is seen in Figure 8.35. The interconnection of the minor bit lines with the major bit lines in the array is shown in Figure 8.36. The use of the minor/major bit line architecture reduces the bit line capacitance for improved bit line response time.

The block memory map has been used to optimize the program/erase performance. A typical 64 K block is composed of 2048 main bit lines; each main bit line has four select transistors, four sub-bit lines, and 256 stacked-gate memory transistors in one block.

The DiNOR architecture depends on FN tunneling for program and erase such as with the NAND approach. Like the NAND and FLOTOX cells, the DiNOR approach uses a block program followed by a selective erase to write an arbitrary pattern into a

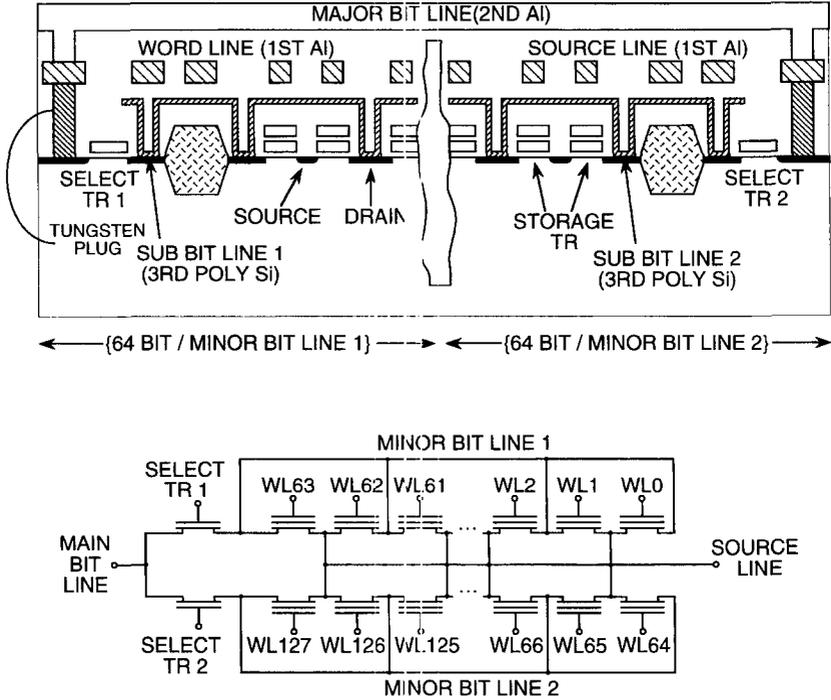


Figure 8.35 Cross-sectional drawing and schematic diagram of the equivalent circuit of one subblock of the DiNOR array.

group of memory transistors. The operation of this approach to flash memory will be described with the assistance of Figure 3.37 and Table 8.6.

The program operation depends on uniform tunneling from the substrate to the floating gate. In the DiNOR array, this is accomplished by applying a bias of 10 V to the gate while the source line and the p well are biased at -9 V. The select gate is also biased at -9 V to prevent any current flow through the memory transistors. A triple-well process is used to split the potential difference and avoid applying more than 10 V bias across the junction in any well.

In the selective erase mode, the word line of the selected row is biased at -9 V. The select gate is bias at 10 V to allow access to the memory transistors. Figure 8.38 shows the erase conditions within a partial array. Those transistors that are to be erased have 6 V applied to their bit lines. These transistors will be erased to a conducting state, "0", by tunneling of electrons from the floating gate to the drain. The cell on WL0 that is connected to MBL1 has 0 V on the drain. This results in a lower field across the oxide between floating-gate and drain so that the tunneling current is reduced to such a small value that this cell remains programmed to a nonconducting, "1," state.

Because the low threshold voltage erased state is the state that is selectively written, the DiNOR operation is able to tighten the threshold distribution by the

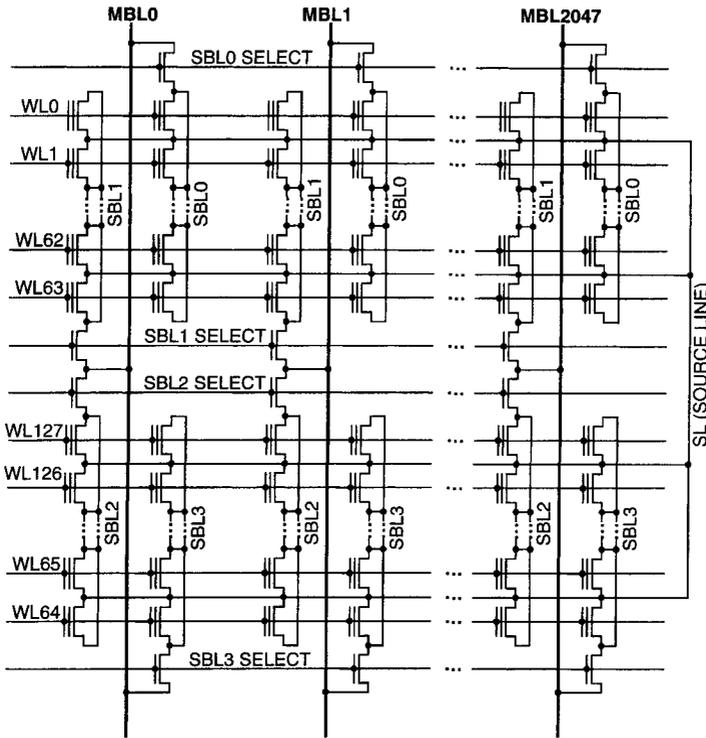


Figure 8.36 Schematic drawing of a section of a DiNOR array showing the interconnection of the minor (sub) bit lines (SBLs), and major bit lines (MBLs), and word lines (WLs).

verifying the erase operation bit by bit. A series of pulses is used to erase the desired bits. The thresholds are measured after each pulse. Those cells that are erased sufficiently, do not have the high drain voltage applied during the next erase pulse. This results in a tightened erase threshold as is illustrated in Figure 8.39.⁶⁰ The narrower erase distribution provides larger margins between the lowest erase threshold and depletion, the depletion margin, and between the largest erase threshold and the lowest read voltage, the read margin.

It should be recognized, however, that the tunneling for the DiNOR array is nonuniform, and occurs because of the potential difference between the drain and floating-gate potentials. This results in band-to-band tunneling at the edge of the drain diffusion. The current will be reduced somewhat because the voltage difference between the drain and the p well is only 6 V, but the drain current will be larger than for uniform tunneling. Hot holes generated during the erase of a cell in one cycle could render it more prone to drain disturb in a later erase cycle.

The read operation is accomplished by applying ~ 1 V to the selected bit-lines, biasing the selected word line to V_{DD} , grounding the source and array p well, and sensing the current flowing from the bit lines. Two factors facilitate operation of this

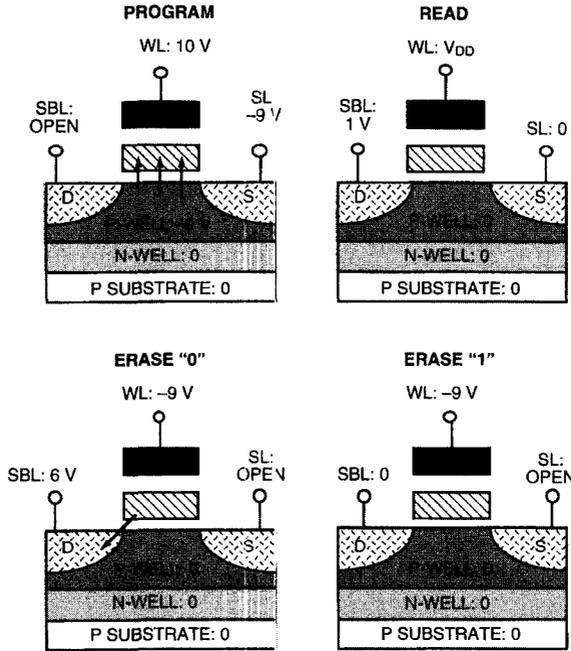


Figure 8.37 Sketches of one memory transistor illustrating the bias conditions that are employed for the various operating modes of the DiNOR array.

TABLE 8.6 Operating Conditions for the DiNOR Array

Signals	ERASE			
	READ	PROGRAM	DATA "0"	DATA "1"
MBL	1 V	open	6 V	0 V
SG	V_{DD}	-9 V	10 V	10 V
SBL	1 V	open	6 V	0 V
WL	V_{DD}	10 V	-9 V	-9 V
SL	0 V	-9 V	open	open
p well	0 V	-9 V	0 V	0 V

array at lower read voltages than is common for the T-cell flash device: (1) because drain-coupling-enhanced punchthrough during programming is not a consideration, lower erased thresholds can be accepted; and (2) the tighter erased distribution discussed above provides a lower maximum erased threshold for a given minimum erased threshold.

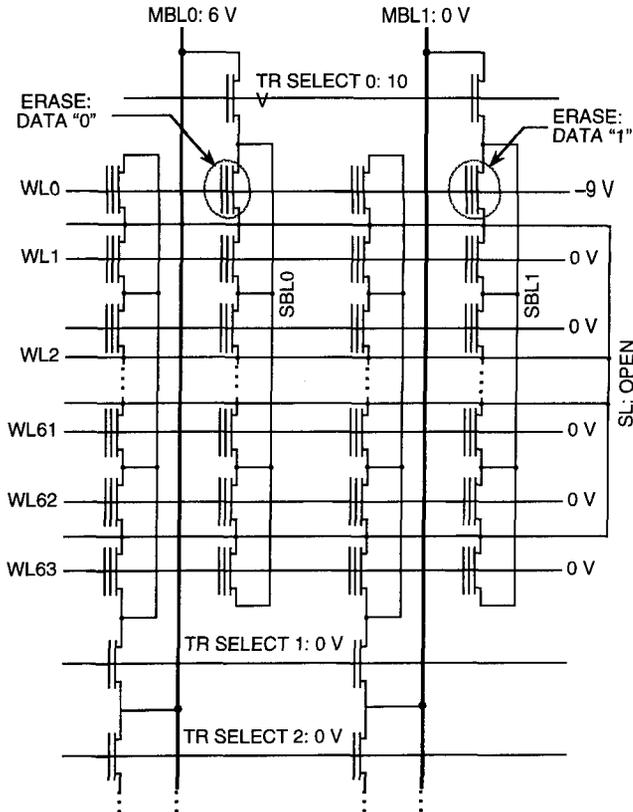


Figure 8.38 Schematic diagram of a partial array showing the bias conditions for selective erase.

8.4 RELIABILITY OF FLOATING-GATE MEMORIES

Floating-gate nonvolatile memories are heir to all the reliability hazards of other MOS integrated circuits such as electromigration and hot-electron degradation of the transconductance, although they are much less sensitive to single-event upset (soft errors) than are DRAMs or SRAMs.⁶¹ The unique hazards relate to the ability to retain data without power over long periods of time (retention), the ability to withstand repeated rewriting of the stored, nonvolatile data (endurance), and the possibility that the data in one cell may be inadvertently altered while reading that cell or writing to another cell elsewhere in the array (disturb). These three unique hazards are discussed in the following sections.

8.4.1 Retention Failures

Since the data are represented as electrons stored on isolated gates, the data can be lost if electrons leak to or from the gates. Data can also be lost if positive ions drift or diffuse to negatively charged gates.

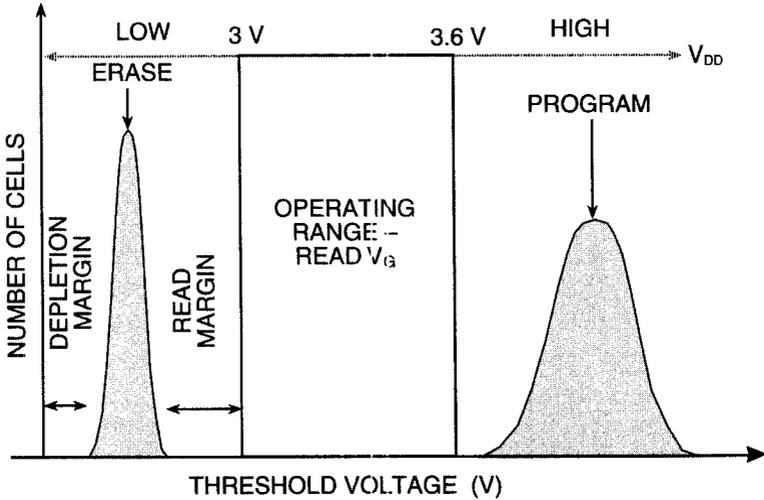


Figure 8.39 Threshold voltage distribution of the DiNOR cell.

Thermal emission of electrons over the potential barriers surrounding the floating gate (see Fig. 8.1), is described by the Richardson–Schottky equation

$$J = AT^2 \cdot \exp \left\{ \frac{-(\phi_B - \sqrt{q^3 E / 4\pi\epsilon_0})}{kT} \right\} \tag{8.6}$$

where J is the emitted current density, A is the Richardson constant, ϕ_B is the oxide barrier height, E is the self-induced electric field, and the other symbols have their usual meanings.⁶²

Because of the high SiO_2 potential barrier, charge loss via thermionic emission is expected to be very low, an expectation borne out by experiment.⁶³ Similarly, calculation of the FN tunnel current for a properly designed memory cell in retention mode shows it to be insignificantly small. Enhanced leakage has been observed through defects in the oxides surrounding the gates; this leakage exhibits apparent activation energies in the range of 0.6–0.8 eV.⁶⁴

Mobile positive ions can be either members of the alkali metals group or hydrogen. Data loss from mobile alkali ions has an apparent activation energy in the range 1.2–1.4 eV.⁶⁴ In clean processes typical of commercial IC manufacturing, alkali ions are excluded from the wafer processing environment and are able to reach the memory cells only by diffusion from the edge of the die if the seal ring around the die is inadequate or through cracks in the overlying passivation coating. Ions from either source tend to appear as electron loss in a local area that grows with time at elevated temperatures. It has as been shown that hydrogen incorporated into deposited dielectric layer may become mobile and create apparent electron loss from floating gates.⁶⁵ This effect has an apparent activation energy of ~ 1.0 eV.

All the processes responsible for retention loss in floating-gate memories discussed thus far have relatively large apparent activation energies and, consequently, can be screened effectively with a high-temperature bake. With proper screening procedures in place, manufacturers routinely produce memories for which retention failures contribute <1 FIT to the overall failure rate from these causes. (A FIT is defined as one device failure per 10^9 device hours of operation.) However, as is discussed in the following section, FN tunneling can damage the oxide so that another process that is not easily screenable can appear.

8.4.2 Endurance Failures

Endurance failure may be defined as the inability of a nonvolatile memory to meet specification as a result of data rewrites. The traditional focus of endurance discussions has been on the effects of FN tunneling on the oxides through which it takes place. This section follows that tradition, but it should be noted that the high voltages applied during the write function may cause a product to fail because of a gate oxide failure in the peripheral circuitry. It should also be noted that hot-electron injection will result in charge trapping in the gate oxide that will inhibit further injection and may also lead to a reduction in read currents.^{45,66} These effects are not important in EPROMs that are usually programmed only a few times, but can be significant in the case of flash memories that employ hot-electron injection for programming. Methods for measuring product endurance are discussed in IEEE STD-1005.

The FN tunneling endurance limit of floating-gate nonvolatile memories is a result of damage to the dielectric around the floating gate caused by the electric stresses applied to write to the cell. When current passes through this dielectric, traps are generated whose density increases with the time integral of the current density that passes through the dielectric.⁶⁷ In addition, high field damage has been observed.⁶⁸

In thin-oxide memory cells, the end of endurance is typically dielectric breakdown, which causes rapid loss of charge in the floating gate. This effect is slowed by lower temperatures and accelerated by increased temperatures.

In thick-oxide (textured poly) devices, the end of endurance is generally caused by the negative charge trapped in the dielectric, which eventually inhibits tunneling and, therefore, nonvolatile programming. Increasing temperature decreases the net trapping rate; decreasing temperature increases it.

In many of the “flash” devices, the end of endurance is caused by hot-electron trapping in the charge transport oxide. In others, effects similar to those seen in thin-oxide or thick-oxide devices have been observed.

As tunnel oxides are scaled below ~ 100 Å, they become sensitive to stress-induced leakage current (SILC). This effect is the increased electrical conduction of the gate dielectric at low electric fields that is the result of passing charge through the dielectric during programming or erasing.^{69–72} The increased low-field conduction may manifest itself as a “disturb” or, in egregious cases, as a retention failure.^{45,73,74} The SILC increases with integrated charge transported through the oxide. This effect

increases greatly as the tunnel dielectric thicknesses decrease and must be taken into account in any scaling scenario. SILC has two components: one that decreases over time after the high voltage stress, which predominates for thicker oxides, and one that is time-invariant, which becomes more important for thinner oxides.^{72,75}

8.4.3 Disturb Failures

Disturbs have been observed for almost as long as floating-gate memory elements have been used. However, the advent of flash memory has focused attention on this subject. In order to obtain smaller cell sizes, select transistors that have served to minimize the likelihood of disturbs in more established byte alterable E²PROMs have been omitted in many flash EEPROMs. This has forced more careful and systematic consideration and characterization of the disturb phenomena.

There are many ways of categorizing disturbs: according to the mode in which they occur (e. g., read disturbs and write disturbs), whether they increase or decrease the net number of electrons on the floating gate (i.e., as electron-gain disturbs and electron loss disturbs), or a grouping that tries to establish a functional discrimination (e.g., soft program, drain disturb, gate disturb).

In this section we discuss the underlying mechanisms that cause disturbs, together with a systematic approach to looking for and describing the disturb mechanisms. In studying disturbs, one must be constantly cognizant of the array architecture, which determines the voltages that appear at the terminals of the floating-gate transistor. Figure 8.40 illustrates some common paths for disturb currents.

CHE injection, case E in Figure 8.40, is always a consideration for read disturb because both the read and programming operations apply a positive voltage on the control gate and voltage of $\sim \frac{1}{2}$ the control-gate voltage on the drain. The only difference between the two modes is that the applied voltages are lower during read. Fortunately, the probability of hot-carrier injection decreases exponentially with decreasing applied bias. Nevertheless, CHE programming memories are programmed in 1–100 μ s, and a cell must be capable of being read for a minimum of 10 years without being disturbed significantly. Therefore, the injected hot-electron current during read must be reduced by 10^{13} – 10^{15} times with respect to the current injected during programming to meet this requirement if the circuit utilizes static channel current during sensing.

Since it is known that the CHE injection varies exponentially with the inverse of drain voltage, a plot of the log of the time for a fixed change in threshold voltage vs. $1/V_{\text{DRAIN}}$ for drain voltages in excess of normal read conditions allows the determination of the maximum drain voltage during the read condition for an adequate disturb margin.

In reading, current is deliberately caused to flow in the cell channel. It is also possible for current to inadvertently flow through a deselected cell during program conditions. Consider, for example, Figure 8.15. In this case, assume W/L 1 is selected for programming by biasing the control gate at a positive bias. The other rows are deselected by grounding the unselected control gates. Suppose that a

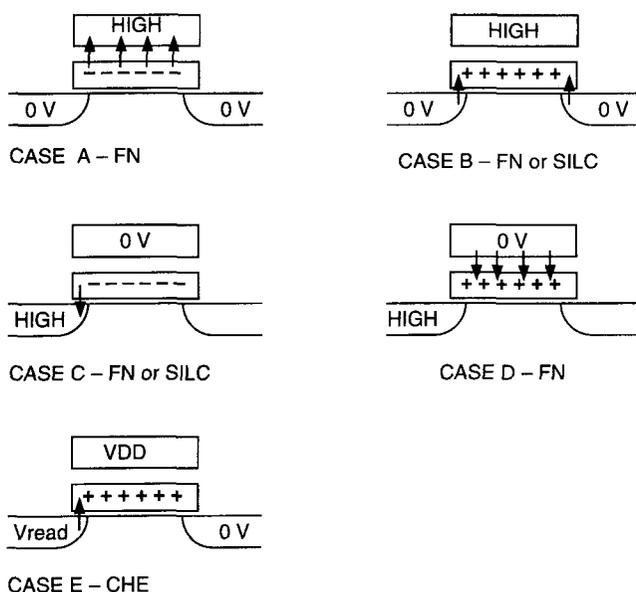


Figure 8.40 Schematic illustration of some common disturb mechanisms.

positive voltage is applied to B/L1 to cause the programming current to flow through the transistor at the intersection of W/L1 and B/L1. The voltage applied to the drain of the selected transistor is also applied to the drains of all of the transistors connected to the same bit line. This voltage may cause punchthrough of an unselected transistor. The current flowing through a punched-through transistor will generate hot electrons that will tend to program the device. A transistor with a shorter channel, whether as a result of a processing defect or as a result of normal process variation, is more susceptible to punchthrough. This effect is enhanced in floating-gate transistors because the voltage applied to the drain of such a transistor couples the floating gate to a more positive voltage, which increases the likelihood of punchthrough.

Although the discussion here was in terms of an EPROM, any array that relies on hot-electron injection from the channel is probably susceptible to this disturb condition. In particular, the T-flash cell, the FETIF flash cell, and the TPVG flash cell are all potentially susceptible. Like many high-voltage disturbs, this type can be screened for by applying a voltage in excess of that allowed in operation during production testing.

The other disturb mechanisms involve Fowler–Nordheim tunneling, which, as has been discussed, has an exponential field dependence. Cases A and D in Figure 8.40 involve electrons tunneling into states in the interpoly dielectric and then hopping to the opposing electrode. This phenomenon can be enhanced by positive curvature of sections of the emitting surface that increase the field at the emitting surface or by defects in the dielectric, which can lower the effective potential barrier. In either case, the disturbs are related to spot defects for any well-designed interpoly

dielectric and can be screened effectively by applying voltages in excess of normal operating potentials during manufacturing testing.

Cases B and C involve tunneling between the floating gate and the substrate. These can also be screened during production testing. However, there are complications in devices that experience high levels of hot-hole injection into the oxide, typically those that erase via FN tunneling into the substrate. The resultant SILC, described in the endurance section, causes greatly enhanced disturb sensitivity. It is impractical to screen for disturbs caused by SILC, but the effect can be characterized and the cell operating conditions designed to avoid failure.

The design of any nonvolatile array requires careful analysis of all the possible disturbs for the approach chosen followed by characterization of the mechanisms to assess the magnitude of the effects. Screens may be required during manufacturing to reject cells that have excessive disturb currents arising from fabrication anomalies. The interested reader is directed to IEEE STD-1005 for a detailed discussion.

8.5 FUTURE TRENDS AND SUMMARY OF FLOATING-GATE MEMORY

In a recent paper Mead studied scaling trends for MOS technology over a period of 22 years.⁷⁶ He found that empirically

$$t_{\text{ox}} \approx \max(210\ell^{0.77}, 140\ell^{0.55}) \quad (8.7a)$$

$$V_{DD} \approx 5\ell^{0.75} \quad (8.7b)$$

$$n_{\text{dop}} \approx 4 \times 10^{16}\ell^{-1.6} \quad (8.7c)$$

$$V_t \approx 0.55\ell^{0.23} \quad (8.7d)$$

where t_{ox} , V_{DD} , n_{dop} , and V_t are the gate oxide thickness, supply voltage, channel doping density, and transistor threshold voltage, respectively, and $\max(a,b)$ indicates that the larger of the two values a or b should be used. The symbol ℓ is the linear scaling parameter, the minimum feature size.

Do the structures used in floating-gate memory scale at the same rate or faster than conventional MOS technology? And, if not, what innovations can be used to allow the storage density of floating-gate nonvolatile memory to keep pace with the requirements of IC technology in the ULSI era?

The unique features involved in scaling of CHE programming is the requirement that the lateral electric fields that generate the hot electrons and the vertical electric fields that collect the electrons remain constant for constant programming requirements. For cells that depend on drain-side programming, a further constraint is that punchthrough of the unselected erased cells on a column being programmed must be avoided in the face of capacitive coupling of the of the drain, programming

voltage to the floating gate, which enhances drain turnon. It has been found necessary to add a boron halo implant around the drain junction to suppress punch through.⁷⁷

Several scaling scenarios for CHE programming have been considered. One that assumes that the lateral electric field at the junction for programming and the fields across the dielectrics are constant and that has been modified to account for the effects encountered in erase yields the following scaling laws:⁷⁸

$$x_j = \frac{0.2}{\ell} \quad (8.8a)$$

$$t_{\text{ox}} = \frac{0.2}{\ell^{0.25}} \quad (8.8b)$$

$$t_{\text{diel}} = \frac{0.2}{\ell^{0.5}} \quad (8.8c)$$

$$V_{PP} = \frac{12}{\ell^{0.5}} \quad (8.8d)$$

$$V_{PD} = \frac{6}{\ell^{0.42}} \quad (8.8e)$$

$$V_{DR} = \frac{1}{\ell^{0.42}} \quad (8.8f)$$

where x_j is the junction depth, t_{diel} is the interpoly dielectric thickness, and V_{PP} , V_{PD} , and V_{DR} are the voltage on the control gate and drain during programming and on the drain during read, respectively. The threshold shift resulting from programing scales with the programming gate voltage. This results in a lower programmed threshold in the read condition and will require a lower control-gate voltage in the read mode. The erased threshold of the memory transistor cannot be arbitrarily reduced because of the punchthrough voltage requirements. Thus, there is a tradeoff between the threshold difference of the erase and program states and the programming time. Since it is desirable to decrease the programming time of a memory element so that the time to program an array not grow linearly with the number of words, the natural consequence is a reduced read margin. Fortunately, the floating-gate cell is inherently a gain cell so that clever circuit design can overcome the decreased read margin.

The voltage that must be applied across an oxide to induce a given FN tunneling current is essentially proportional to the oxide thickness for the common case of planar electrodes. The major limitation to the reduction of the tunnel oxide thickness is imposed by the requirement that the leakage of stored charge be low enough that the data represented by the charge can be read out after some long period of time, such as 10 years. Simple calculations of the quantum-mechanical tunneling current indicate that the tunnel oxide could be reduced until direct tunneling between the two conductors surrounding the oxides becomes significant, for example, to ~ 5 nm.

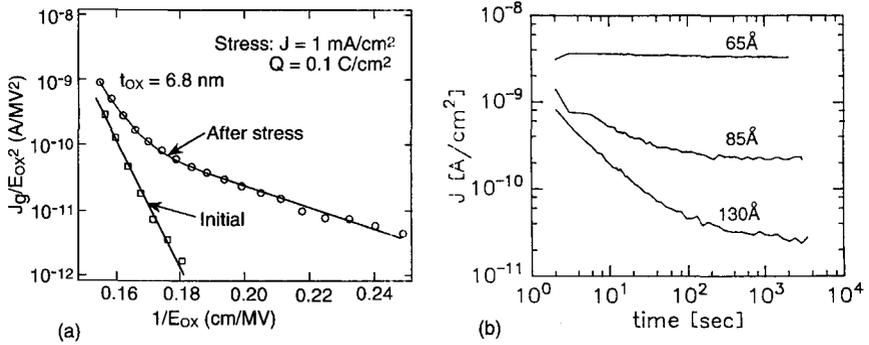


Figure 8.41 Current through oxide (a) as a function of inverse field before and after a SILC-inducing stress and (b) as a function of time after stress for constant applied bias for different oxide thicknesses.

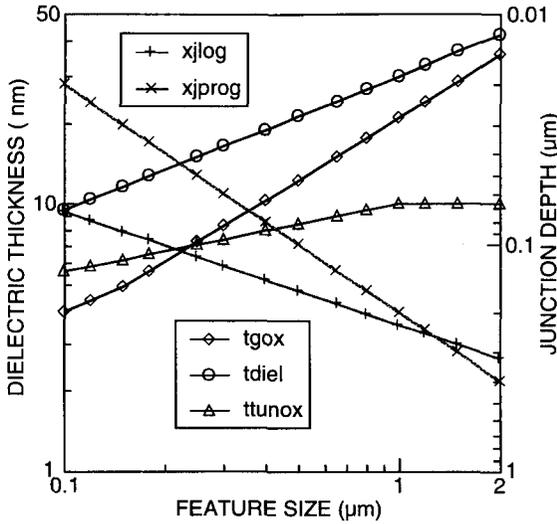
Unfortunately, stress induced leakage current, SILC, places a more stringent limitation on the lower limit of the tunnel oxide thickness. As was noted previously, the stress inherent with FN tunneling introduces traps into the oxide. On thicker oxides, for instance, >10 nm, the result is a transient current as the traps fill and empty by electrons tunneling from the electrodes. For thinner oxides, say, <7 nm, the probability that the traps filled by tunneling from the cathode are emptied by electrons tunneling to the anode is high enough that a time-independent component to the SILC becomes dominant, as shown in Figure 8.41b. As Figure 8.41a shows, the voltage dependence of SILC is much weaker than that of FN current through an unstressed oxide; therefore, SILC exceeds the FN current by many orders of magnitude at the low electric fields that typify storage and read conditions.⁷⁸ Measurement of the temperature of SILC show that it has little or no temperature dependence, demonstrating the expected behavior for a tunneling process.⁷⁹ SILC is an especially stringent limitation for cells that tunnel to the same junction that is biased during read mode, but it is in general believed to set a lower limit on the tunnel oxide thickness of 7–8 nm at the present state of oxide technology.

Probably because of the relative thickness of the oxides typically used ($> \sim 50$ nm), SILC has not been observed in geometrically enhanced tunnel emission. On the other hand, the voltage required for a given tunneling current decreases much less than linearly with tunnel oxide thickness for these structures. Operation of these structures at significantly lower voltages would appear to require smaller radii of curvature of the emitting surface; this, unfortunately, leads to increased current density near the emitter, which would increase the electron trapping rate and lower the endurance.

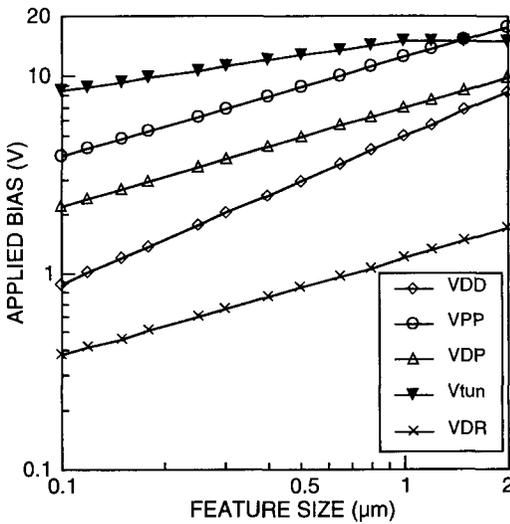
In most cases, an oxide/nitride/oxide sandwich layer is used for the interpoly dielectric. The sandwiched nitride layer acts as a trapping site for any electrons that passes through defects in the bottom oxide layer. These trapped electrons set up a retarding field that suppresses further injection. The top oxide layer serves to keep any trapped electrons from hopping to the top electrode and to suppress hole injection from the top electrode. Experiments have shown that this dielectric can be

scaled down to an effective oxide thickness of 13 nm.⁸⁰ Recent work on deposited dielectrics of SiO₂ or Al₂O₃ indicate that the use of a single material may allow an effective dielectric thickness below 13 nm.^{81,82}

The vertical scaling of structures and the voltage scaling of applied biases are shown for both logic gates and floating-gate cells in Figure 8.42. Note that the tunnel oxide thickness, t_{tunox} , scales much more slowly than does the gate oxide thickness,



(a)



(b)

Figure 8.42 Predicted scaling of (a) vertical dimensions and (b) applied biases vs. lateral feature size. Note that the vertical scale on junction depth is reversed.

t_{gox} , for logic technology and that for feature sizes on the order of $0.3 \mu\text{m}$ the logic technology gate oxide becomes thinner than the tunnel oxide. The thickness of the interpoly dielectric, t_{diel} , scales at a rate between the other two insulators. Note also that the junction depth for the drain-side CHE programming, $x_{j\text{prog}}$, scales much more rapidly than does the source–drain junction for the logic technology, $x_{j\text{log}}$, with the crossover occurring near $1 \mu\text{m}$. This suggests that the effective channel length of drain side CHE programming cells may not scale as rapidly as the effective channel length of logic technology unless the gate oxide is scaled more aggressively.

The voltage for tunneling is determined by the tunnel oxide thickness given that the control-gate coupling ratio remains constant. This means that this bias, V_{tun} , is the most slowly scaling bias voltage because the tunnel oxide thickness changes very slowly, as can be seen in Figure 8.42a. The rate of scaling of the various bias voltages is shown in Figure 8.42b. It is apparent from this figure that the logic power supply bias, V_{DD} , is decreasing much faster than either the gate bias required for CHE programming, V_{PP} , or the very slowly varying tunneling bias, V_{tun} , as the feature size decreases. The other two biases of interest, the drain bias for CHE programming, V_{dp} , and the drain bias for read, V_{DR} , decrease slightly more slowly than V_{PP} with decreasing features size.

Despite the difficulty in scaling some floating-gate features and the associated limits on voltage scaling, floating-gate cells fabricated with $0.2\text{-}\mu\text{m}$ process technology that occupy $\sim 0.25 \mu\text{m}^2$ and are suitable for the manufacture of gigabit flash memories have been reported.³² It is reasonable to believe that many more such floating-gate memory cells employing different approaches will be developed and that the technology will be pushed to 4 Gb and beyond.

8.6 SILICON NITRIDE MEMORY

8.6.1 Introduction to Silicon Nitride Memory

When the gate dielectric of an MOS transistor is modified to incorporate a layer of silicon nitride, the structure can serve as a memory device with the threshold voltage controlled by the amount of charge stored in the silicon nitride.^{83,84,85} A cross-sectional diagram of a silicon nitride device is shown in Figure 8.43. The structure of the gate dielectric is layered. Insulator I(1) is normally thermally grown silicon dioxide on the channel region and is thick enough to prevent back-tunneling of the charge stored in insulator I(2) into the silicon. Insulator I(2) is composed of silicon nitride or silicon oxynitride and is typically at least ten times the thickness of insulator I(1). It is normally deposited by chemical vapor deposition. Insulator I(3) is typically two to three times thicker than insulator I(1) and is formed by thermally oxidizing the uppermost region of insulator I(2). Insulator I(3) is optional and generally provides better retention of the charge stored in insulator I(2) at the expense of higher programming voltages. Thicknesses of insulators I(1), I(2) and I(3) are approximately 2, 20, and 4 nm, respectively.

The structure shown in Fig. 8.43 is referred to as an *SONOS device*, for silicon/oxide/nitride/oxide/silicon. If insulator I(3) is not used, the structure is referred to as

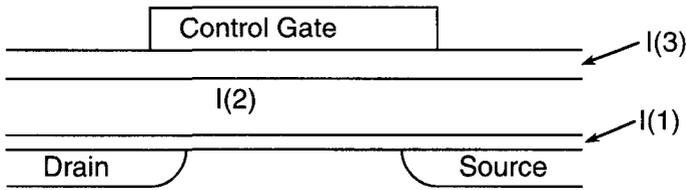


Figure 8.43 Structure for a silicon nitride nonvolatile MOSFET. I(1) and I(3) are silicon dioxide insulators, and I(2) is composed of silicon nitride or silicon oxynitride.

an SNOS device and if the gate electrode is composed of metal rather than polysilicon, the structure is referred to as an *MNOS device*.

The total thickness of the gate insulator stack is such that a field-controlled charge transport mechanism, modified Fowler–Nordheim tunneling, allows the injection of charge into the insulators. Some of the injected charge passes through the insulators, and the remainder of the injected charge is trapped in the silicon nitride layer. This trapped charge in the gate dielectric causes the threshold voltage of the device to shift similar to that in the case of a floating gate device. To erase the device, a voltage of the opposite polarity is applied across the insulator stack and the trapped charge is ejected back into the silicon or charge of opposite polarity is injected into the insulator stack. The voltage difference between the most positive threshold voltage and the most negative threshold voltage for a given programming voltage and time is referred to as the *memory window*.

To read the logic state of a silicon nitride device, a voltage between the most positive and most negative threshold voltages (within the memory window) is applied to the gate. Sensing circuitry senses whether the channel of the silicon nitride device conducts or is turned off. The read cycle is nondestructive because the state of the silicon nitride device is not changed during the read cycle.

In silicon nitride devices, the charge is stored in traps within an insulator, whereas in floating-gate devices, the charge is stored on a conductor (the floating gate) embedded within insulators. A floating gate device need be programmed only at a corner or edge of the floating gate since the charge spreads across the conductor. However, the silicon nitride device must be programmed with a voltage applied across the entire channel region of the device. This is a major disadvantage of the silicon nitride technology. However, lower programming voltages are possible with silicon nitride devices compared to those for floating gate devices.

Once the device is programmed by injecting charge into the gate insulator stack, the shifted threshold voltage of the device is empirically observed to decay with a dependence approximately linear in log time. While originally treated as a disadvantage of silicon nitride devices, this gradual decay of threshold voltage has been used as a nondestructive electric screen to discard memories with defective memory cells. This screening results in high levels of reliability for silicon nitride nonvolatile memories.

8.6.2 Physics of the Silicon Nitride Technology

Tunneling and Emission Mechanisms

The energy-band diagram corresponding to the structure similar to that in Figure 8.43, but where the top insulator, I(3), is not used, is shown in Figure 8.44.⁸⁶ Essentially, a silicon dioxide and a silicon nitride capacitor are in series. The relative dielectric constant of the silicon nitride is approximately 8 compared to 3.95 for the silicon dioxide. Therefore, the electric field across the tunnel oxide is approximately twice as large as the electric field across the silicon nitride:

$$E_{ox} = \left(\frac{\epsilon_n}{\epsilon_{ox}} \right) E_n \sim 2E_n \tag{8.9}$$

where E_{ox} is the electric field in the tunnel oxide and ϵ_n and ϵ_{ox} are the permittivities of the silicon nitride and tunnel oxide, respectively, and E_n is the electric field in the silicon nitride. This is represented by a steeper slope on the band edges corresponding to the silicon dioxide compared to the band edges corresponding to the silicon nitride.

Whereas Fowler–Nordheim (FN) tunneling, described earlier in this chapter for floating-gate devices, relates to tunneling of charge through a triangular barrier, modified FN tunneling relates to the tunneling of charge through a trapezoidal barrier. This is the case for silicon nitride devices with thin tunnel oxides.⁸⁷ Modified FN injection through the tunnel oxide and a portion of the silicon nitride,^{86,88} direct tunneling into the charge traps,⁸⁹ and Poole–Frenkel (PF) emission from the charge traps^{90,91} are significant factors in the programming dynamics of silicon nitride

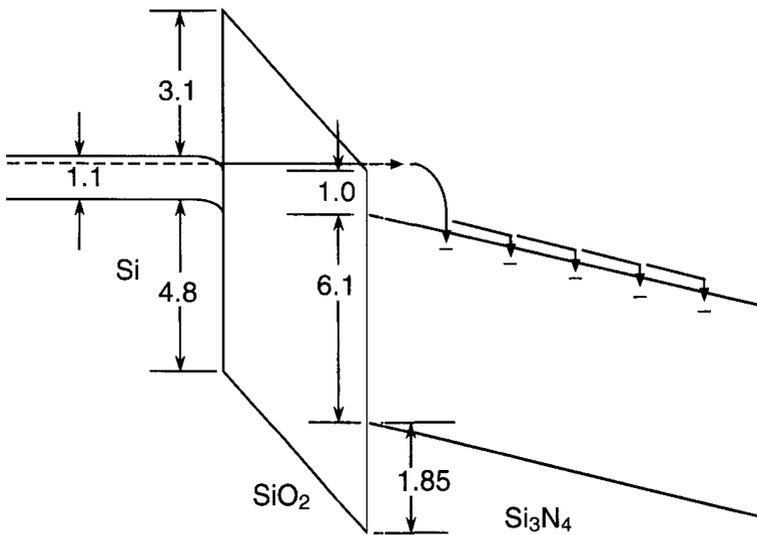


Figure 8.44 Energy-band diagram corresponding to the MOSFET structure shown in Figure 8.43 without I(3). Values of barrier heights are shown in electron volts.

devices. For both modified FN injection and PF emission, the current is exponentially dependent on the electric field. An analytical description of modified FN tunneling is beyond the scope of this text, but FN tunneling was described symbolically in Eq. 8.3.

The bulk-limited current–density expression for current transport in the silicon nitride device is

$$J = C_1 E_n \exp\left(\frac{-q\phi_1}{kT}\right) \exp\left[\left(\frac{q}{kT}\right)(\beta E_n)^{1/2}\right] + C_2 E_n^2 \exp\left(\frac{-E_2}{E_n}\right) + C_3 E_n \exp\left(\frac{-q\phi_3}{kT}\right) \quad (8.10)$$

where E_n is the electric field in the silicon nitride; ϕ_1 is the depth of the trap potential well; ϕ_3 is the thermal activation energy involved in hopping of thermally excited electrons between isolated trapping sites; C_1, C_2, C_3, E_2 , and β are characteristic constants depending on the trap level and the dielectric constant; and the other parameters have their normal meanings.⁸⁷ In Eq. 8.10, the first term is due to Poole–Frenkel emission, the second term is due to tunneling field emission of trapped electrons into the silicon nitride conduction band, and the third term is due to the hopping of thermally excited electrons between isolated trapping states. The first term limits the current transport at higher temperatures; the second term limits the current transport at lower temperatures. The third term is significant in limiting the current at low fields.

The log of the steady-state current through the silicon nitride is linear with the square root of the applied electric field, consistent with PF emission.^{92,93} An Arrhenius plot of the leakage current for a silicon nitride film deposited by chemical vapor deposition on silicon gives an activation energy of approximately 0.5 eV at an electric field of 4.5×10^6 V/cm.⁹² The activation energy increases for decreasing electric field strength.

The thickness of the thin tunnel oxide is critical in controlling the amount of charge tunneled into the silicon nitride. It must be thin enough to allow charge to tunnel into the structure during programming operations yet thick enough to prevent significant back-tunneling of charge from the insulator during read or standby conditions. Typical thicknesses of tunnel oxides used in silicon nitride devices range from 2.0 to 2.5 nm. During programming, the stored charge increases approximately linearly with time until it saturates; the injected current initially remains relatively constant and then decreases to a steady-state value. At gate-to-channel biases of 8–15 V for silicon nitride layers of the order of 20–25 nm thick, significant amounts of charge are injected into the silicon nitride and a significant fraction becomes trapped.

The location of the charge centroid and the trap depth for the trapped charge in the silicon nitride can be calculated for both constant-current charging and constant-voltage charging of the silicon nitride device.⁹⁴ Assuming that the free electrons in the silicon nitride travel at their saturation drift velocity and that the trapping rate of the injected electrons is in proportion to their free-electron density, an analytical relationship can be determined for the charge centroid under constant-current

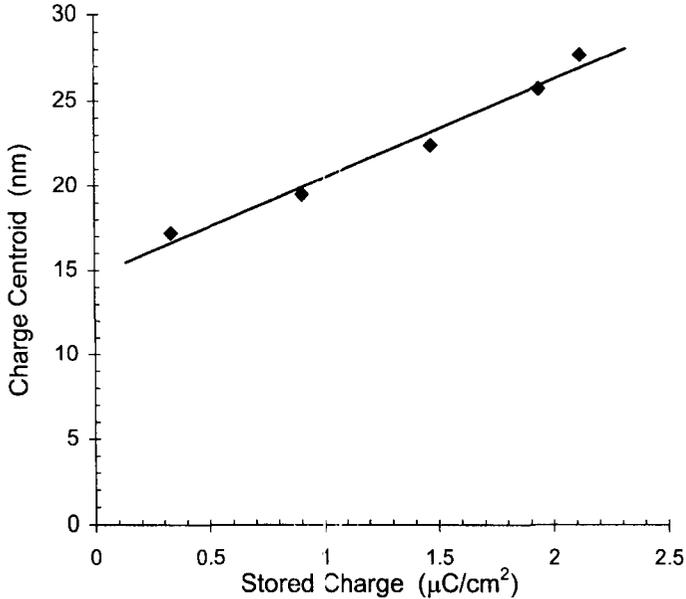


Figure 8.45 Charge centroid of the trapped charge in the silicon nitride versus quantity of trapped charge.

charging.⁹⁵ However, a simple analytic relationship has not been determined for the charge centroid under constant-voltage charging.

The charge centroid moves deeper into the silicon nitride for increased injected charge, higher temperatures, and higher programming voltage. For a 46-nm-thick nitride film with a tunnel oxide of 2 nm, the centroid was found to be approximately 14 nm into the silicon nitride from the tunnel oxide interface.⁹⁴ The trap depth was found to be 0.77 eV with an activation energy of 0.14 eV. This corresponds well with measured trap depths from steady-state current measurements of 0.5–1.3 eV.^{87,96–98} The trap density is of the order of 10^{20} – 10^{21} cm^{-3} .

Figure 8.45 shows the charge centroid for different amounts of stored charge indicating that the charge centroid is located at least 10 nm into the silicon nitride.⁹⁹ This limits the vertical scaling of the thickness of the silicon nitride film.

If the distribution of trapped charge, $\rho(x)$, is known, then the shift in threshold voltage caused by the trapped charge in the silicon nitride device is given by

$$\Delta V_t = \int_0^{d_n} x\rho(x)dx \quad (8.11)$$

where $x=0$ is at the gate electrode/silicon nitride interface and d_n is the silicon nitride thickness. The write/erase characteristics for a typical silicon nitride device are shown in Figure 8.46.¹⁰⁰ Write times of 1 ms or less can be obtained for write voltages of 15 V or less.

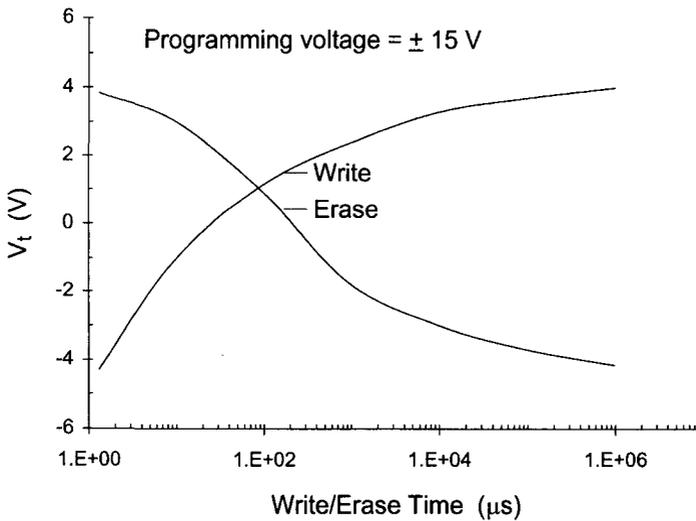


Figure 8.46 Write/erase characteristics for a silicon nitride device.

Retention

Silicon nitride devices are insensitive to pinholes in the tunnel oxide. This is because the charge is stored in an insulator and there is insufficient lateral conductivity in the silicon nitride to drain a significant amount of charge out of the silicon nitride through a pinhole in the tunnel oxide. Because thin silicon nitride films are used in almost all dynamic random access memories and considerable research and development has been applied to low pressure chemical vapor deposition (LPCVD) of high-quality silicon nitride films, electrical shorting (short-circuiting) and time-dependent dielectric breakdown characteristics of silicon nitride films are excellent.

While the dielectric rupture of silicon nitride devices is rare, the threshold voltage of a silicon nitride device shows little decay immediately after programming and then decays approximately linearly with the logarithm of time as shown in Figure 8.47.^{101,102,103}

The retention decay of a silicon nitride device can be characterized as

$$\Delta V_t = K \ln \left(\frac{t}{t_0} \right) \quad (8.12)$$

where K , the slope of the retention curve, is a function of the temperature and electric field as shown in Figures 8.48 and 8.49.¹⁰²

If the decay rate of the threshold voltage of a silicon nitride device is measured as a function of temperature and plotted against reciprocal temperature, a linear dependence results, as shown in Figure 8.50.¹⁰⁰ While the logarithmic time dependence of the decay of the threshold voltage does not strictly obey an Arrhenius

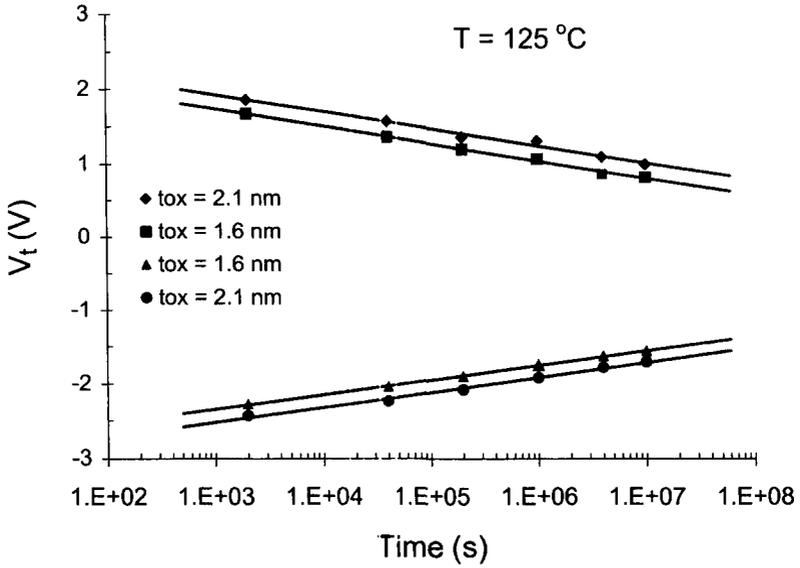


Figure 8.47 Log time retention characteristics of a typical silicon nitride device.⁹⁴ The upper curves represent a programmed device and the lower curves, an erased device.

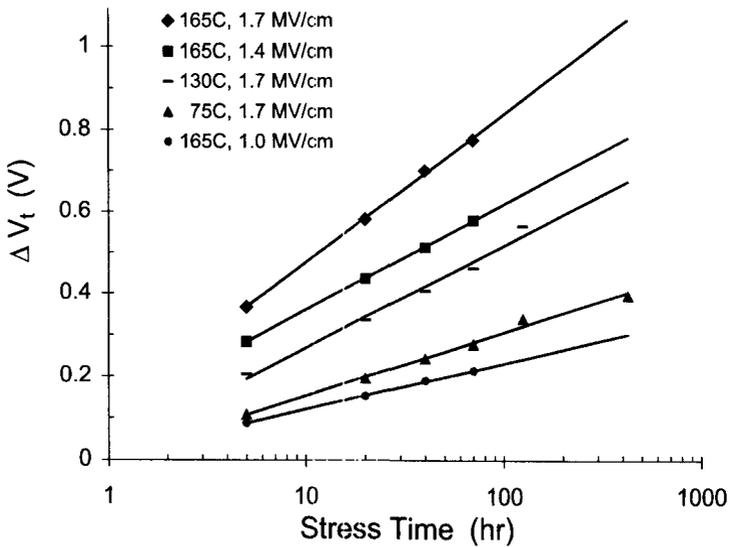


Figure 8.48 Retention dependence on temperature and electric field for a typical silicon nitride device.

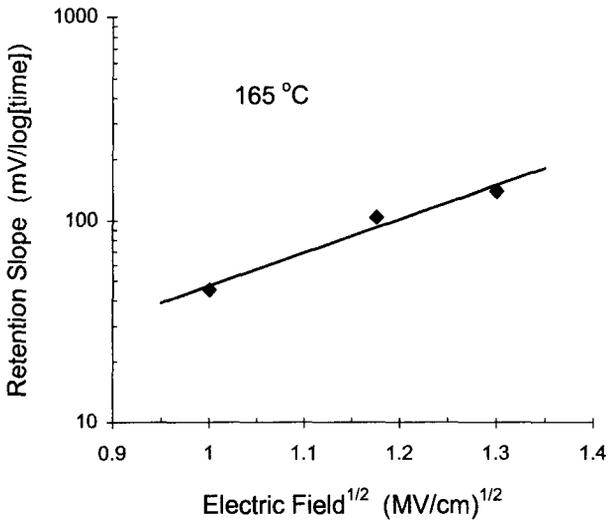


Figure 8.49 Retention curve slope as a function of the square root of the electric field at 165°C.

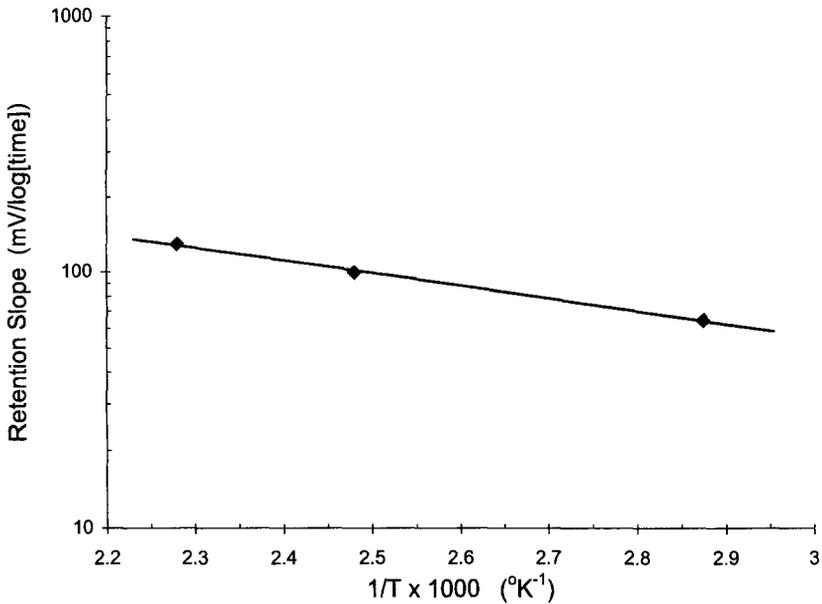


Figure 8.50 Retention curve slope as a function of reciprocal temperature.

rate-limited dependence, as given by

$$\frac{d(\Delta V_t)}{dt} = C(a - \Delta V_t)^m \quad (8.13)$$

where C and a are constants and m is the order of the Arrhenius rate equation, it may be argued that the decay mechanism for traps at the same trap depth and the same tunneling distance obeys an Arrhenius rate equation. Therefore, an activation energy can be determined for the retention since it represents the sum of individual mechanisms, each of which has Arrhenius behavior. The activation corresponding to Figure 8.49 is approximately 0.65 eV. This activation energy can be used to define accelerated tests to screen individual silicon nitride memory cells for retention.

Endurance

After a large number of write/erase programming cycles, typically 10^4 – 10^6 , the retention of silicon nitride devices degrades, which limits the endurance of silicon nitride devices.^{100–105} The exact mechanism for this degradation is not well understood, but relatively permanent charge trapping near the electrode may be responsible in part for the degradation. At write/erase cycles greater than 10^9 , the silicon nitride begins to lose its ability to trap charge and the memory window width is observed to decrease.¹⁰¹ Catastrophic failure occurs as the center of the memory window shifts, typically first to slightly more positive voltages and then dramatically to more negative voltages, placing the memory window outside the range of the sense amplifier.

The amount of retention degradation after write/erase cycling is a function of the growth conditions of the tunnel oxide. Different amounts of degradation are observed for tunnel oxides grown in the presence of water or HCl.¹⁰⁴ For given tunnel oxide and silicon nitride processes and thicknesses, if the retention slope, K , is normalized with respect to the memory window width, a single curve showing the retention degradation is obtained. This curve is shown in Figure 8.51.¹⁰³

Retention and endurance degradation are not typically observed if unipolar programming pulses are applied to the silicon nitride device. Alternating write and erase pulses must be applied to observe significant degradation. Additionally, the degradation is a function of the rise time of the programming pulses yet almost independent of the width of the programming pulses. Faster rise times result in greater degradation. Therefore, the peak electric field in the insulator stack plays a major role in the retention and endurance degradation.

Write Inhibit and Repeated Erase

There are two ways to inhibit the programming on a deselected memory cell in a memory array when the programming voltage is applied to the gate of the deselected device. In one method, the voltage on the source can be the same as the programming voltage on the gate, so no electric field is generated across the insulator stack. This method of write inhibit is referred to as *static write inhibit*.

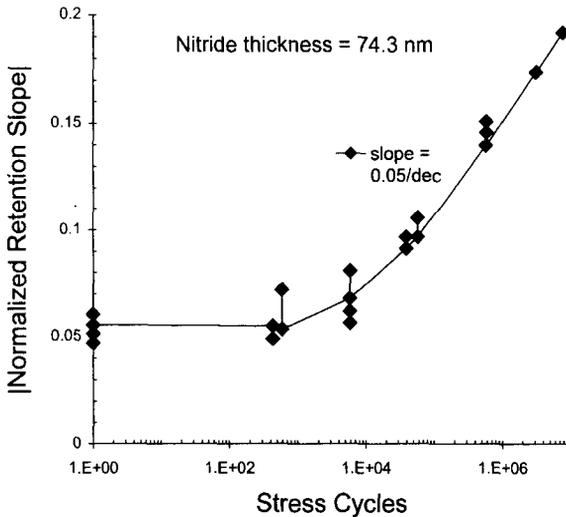


Figure 8.51 Retention curve slope as a function of the number of write/erase cycles. The decay rate has been normalized to the memory window width.

Alternatively, the source and drain of the deselected silicon nitride device can be left floating or at a potential so that they stay reverse-biased during the programming cycle.^{106,107} Provided the programming pulse is short enough, the silicon in the region under the channels of the deselected devices is driven into deep depletion for the duration of the programming pulse. Virtually all of the programming voltage drops across the depletion region formed in the silicon below the channel region, not the silicon nitride. This latter means of inhibiting the programming on deselected devices is referred to as *dynamic write inhibit*.

For silicon nitride memory devices incorporated in a memory array, during programming the voltage on the devices in the direction of a word line is increased to the programming voltage. For the devices along the word line that are deselected, the programming is inhibited by either static or dynamic write inhibit. For the device selected to be programmed, the drain is maintained at a voltage different from that of the programming voltage so as to create an injecting field across the selected device and cause programming.

However, for the erase operation, it is not possible to apply the erase voltage only across the selected device. Therefore, all memory cells along the selected word line are erased during the erase cycle. To overcome this problem, the contents of the word line are first written into a row of latches, then all devices along the word line are erased, the appropriate bits in the latches are changed, and, finally, the entire row is rewritten from the latches. Figure 8.52 shows a schematic diagram of a silicon nitride memory array and Table 8.7 shows the corresponding programming and erase voltages.

The channel region of the deselected device labeled B in Figure 8.52 is not conductive, so dynamic write inhibit prevents writing of this deselected device

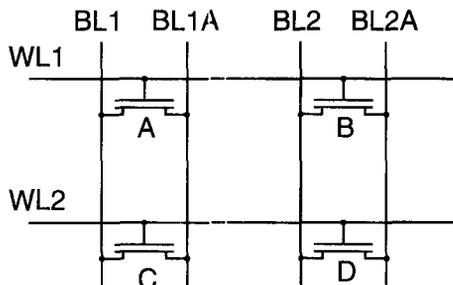


Figure 8.52 Schematic of a silicon nitride memory array. The device to be programmed is labeled A, and the deselected devices are labeled B, C, and D.

TABLE 8.7 Typical Write and Erase Voltage Conditions for Silicon Nitride Memory Array^a

Signal	Program	Erase	Read
WL1	12	-12	3
WL2	0	0	0
BL1	0	—	To sense amplifier
BL1A	0	—	3
BL2	12	—	—
BL2A	12	—	—
Well	0	0	0

^aSame memory array as depicted in Figure 8.52. Note that the entire word line, not just device A, is erased during the erase operation.

during the programming cycle. However, an erase voltage appears between both the source and gate electrode and the drain and gate electrode. Therefore, the state of the silicon nitride charge can be disturbed in the immediate source-drain vicinity, causing a gradual decrease in retention after a number of erase cycles. This degradation must be factored into the design of silicon nitride memories.

Although this erase procedure results in a single-bit erase capability, it is possible for one or more of the bits along the word line to be repeatedly erased if the data in these bits are not changed. This repeated erase drives the erased state of these bits to increasingly negative voltages. A bit that has been repeatedly erased must be capable of being written to the opposite state with a single programming pulse and then meet the retention requirement. Effectively, this repeated erase phenomenon, although caused by the arrangement of the memory cells in the memory array, behaves similarly to the imprint phenomenon observed on ferroelectric devices, discussed later in this chapter. Determination of worst-case retention after repeated erase cycles is shown in Figure 8.53.

Reliability

Early silicon nitride memories had difficulty meeting a 10-year data retention specification, a standard retention specification in the industry. This occurred

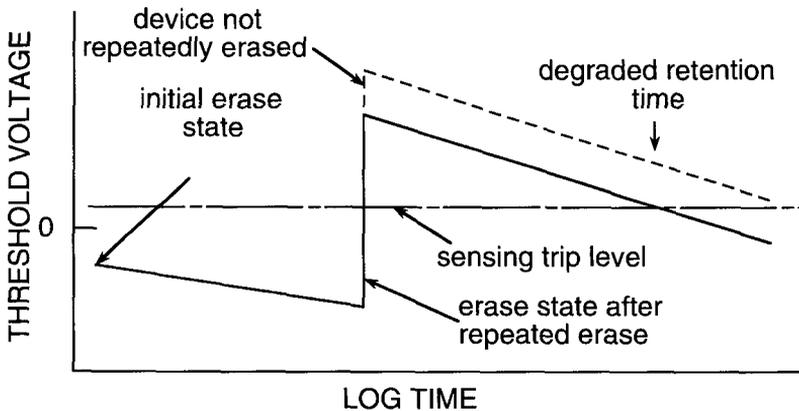


Figure 8.53 Determination of worst-case retention for a silicon nitride device.

because both the written and erased states of the silicon nitride device were enhancement mode and a nonzero voltage had to be applied to the gate of the device to read the state of the device, as shown in Figure 8.52 and Table 8.7. The electric field generated by this positive read voltage accelerated the discharge of the trapped charge in the silicon nitride. To solve this problem, memory cells were designed with the erased state as depletion mode and the written state as enhancement mode so that the state of the device could be determined without applying a read voltage to the gate. Higher reliability could then be obtained at the expense of adding a pass transistor to the memory cell to disconnect any silicon nitride devices with depletion threshold voltages from the bit line. The addition of the pass transistor increases the size of the memory cell.

A properly designed silicon nitride nonvolatile memory will meet a 10-year retention specification at an operating temperature of 125°C and exhibit almost no gate electrode shorts through the silicon nitride as a result of write/erase cycling. Nonvolatile reliability failure rates between 0 and 5 FITs have been obtained on commercial devices.

Scaling

If a nonzero voltage must be applied to the gate of the silicon nitride MOSFET during the read cycle, it is difficult to assure 10-year retention. For a gate voltage of 0 V during the read cycle, one state of the silicon nitride MOSFET must be depletion mode, necessitating the addition of a pass transistor in every memory cell. Therefore, the smallest silicon nitride memory cell requires at least two transistors, limiting the scaling. This is a disadvantage of the silicon nitride nonvolatile memory technology and limits the cost-effectiveness of silicon nitride memories because there are versions of floating gate memory cells that require only one transistor. Although attempts have been made to design a silicon nitride memory with a single transistor memory cell, they have not yet realized commercial success.

Vertical scaling, and therefore programming and erase voltage scaling, of silicon nitride devices is possible. Reducing the thickness of the silicon nitride primarily

reduces the programming voltage, and reducing the thickness of the tunnel oxide primarily reduces the erase voltage. The thickness of the tunnel oxide is limited to slightly more than 2 nm to avoid significant back-tunneling of the trapped charge and resulting unacceptable retention. Because the charge centroid of the trapped charge in the silicon nitride is located approximately 10 nm into the silicon nitride, the lower limit on the scaling of the silicon nitride is approximately 15 nm with a top dielectric of 3 nm. If the top dielectric is optimized, a silicon nitride thickness of approximately 10 nm may be used. The top dielectric must prevent significant loss of charge from the silicon nitride onto the gate electrode. This places the lower limit of the top dielectric at approximately 2.5 nm. With vertical scaling to these limits, programming voltages as low as 8 V can be realized with the silicon nitride technology.

Radiation Hardness

Silicon nitride devices exhibit superior performance in ionizing radiation environments. Ionizing radiation, such as that due to gamma rays, penetrates the structure and causes electron–hole pairs to be generated within the insulators. Because silicon nitride has many hole and electron traps, many of the excess electron–hole pairs generated by the radiation either readily recombine or become trapped in the silicon nitride and become neutralized.

At increasingly higher radiation doses, the memory window collapses.^{108,109} The collapse of one side of the memory window in silicon nitride for a constant ionizing radiation dose rate is shown in Figure 8.54.¹⁰⁸ Total doses of greater than 10^6 rads can be reached with a detectable memory window remaining.

Silicon nitride memory devices can tolerate greater ionizing radiation doses than floating-gate devices. This is because electron and hole traps are undesirable in the tunnel dielectrics of floating-gate devices and the electron–hole pairs generated by ionizing radiation are readily separated by the electric field and travel to opposing electrodes. This causes the charge on the floating gate to be disturbed. The superior radiation hardness of silicon nitride devices compared to floating-gate devices has been an advantage in space and nuclear markets.

8.6.3 Silicon Nitride Memory Cells

Several memory cells based on silicon nitride technology have been developed. These memory cells fall into three basic categories. The first category consists of memory cells that are based on a single-gate transistor containing the silicon nitride. The second category consists of memory cells that are based on a single-gate transistor containing the silicon nitride used in combination with one or more access transistors or gates. For either category the sensing may be accomplished by comparing the voltage generated on a bit line during the read operation to a reference voltage or comparing the voltage generated on a bit line to a voltage generated on a complementary bit line. The third category consists of memory cells based on SRAM memory cells where the silicon nitride devices are added to the SRAM cell to provide the nonvolatile memory feature. Either silicon nitride capacitors or silicon nitride transistors can be used for the memory cells in the third category.

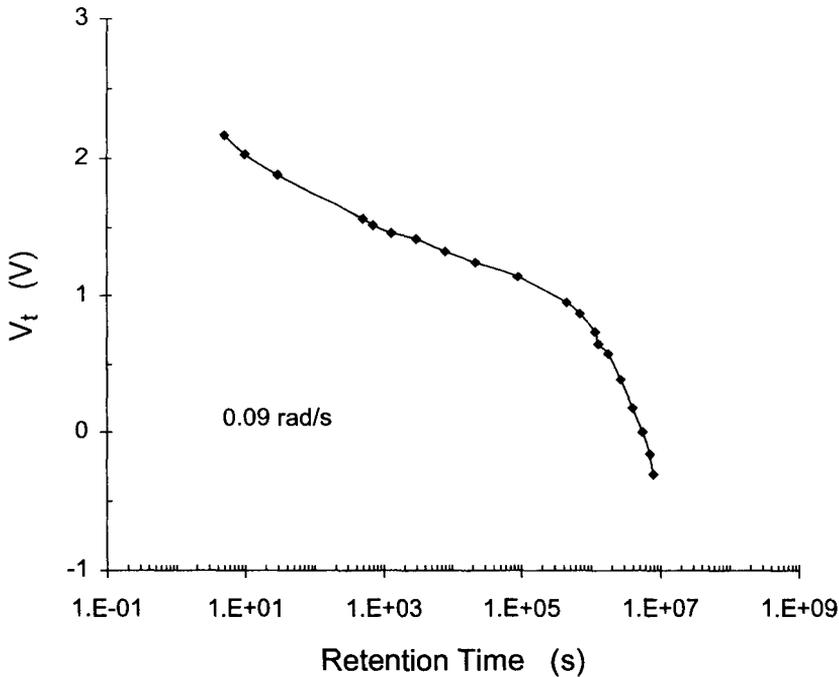


Figure 8.54 Collapse of one side of the memory window versus time for a constant ionizing radiation dose of 0.09 rad/s.

Single-Gate Silicon Nitride Memory Cells

The earliest version of a silicon nitride memory cell was the single-gate silicon nitride memory cell.^{83,98} In Figure 8.43, the cross-sectional structure of a single-gate MNOS memory cell was shown. The corresponding I–V characteristics are shown in Figure 8.55.¹⁰⁵ The reference voltage for the sense amplifier is between the two enhancement threshold voltages, allowing the logic state of the memory cell to be determined. The operation of a memory array based on this memory cell was shown in Figure 8.52 and Table 8.7.

The single-gate silicon nitride memory cell has reliability disadvantages because positive read voltages applied to the gate electrode result in increased loss of programmed charge from the silicon nitride and, consequently, poor data retention. Additionally, short circuits could form between the metal gate and the source or drain after write/erase endurance cycling. Solving these problems has resulted in development of the stepped-gate memory cell.^{110–112}

A cross section of a stepped-gate or source–drain-protected memory cell is shown in Figure 8.56. While this memory cell solves the gate electrode shorting problem by providing thicker oxide over the source and drain, it results in a larger memory cell size. The threshold voltage in the normal gate oxide regions allows the channel region to be turned off with zero gate voltage even if the silicon nitride portion of the channel is in depletion mode. The stepped-gate structure also solves

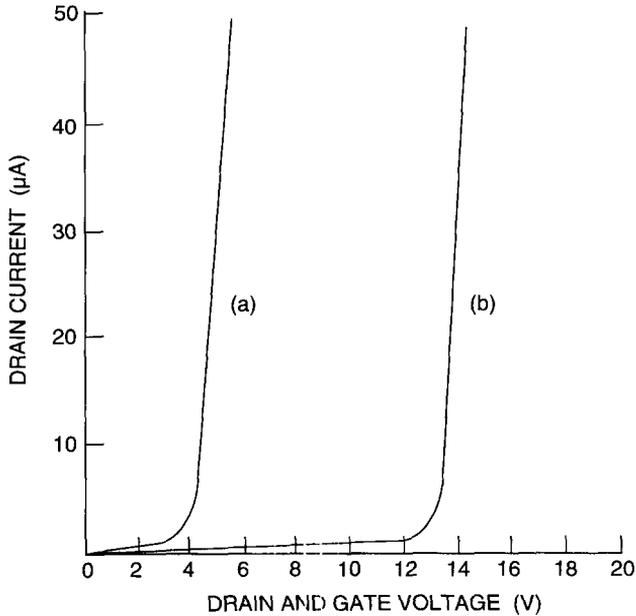


Figure 8.55 I-V characteristics of the single-gate MNOS memory cell shown in Fig 43; (a) high conductance state and (b) low conductance state.

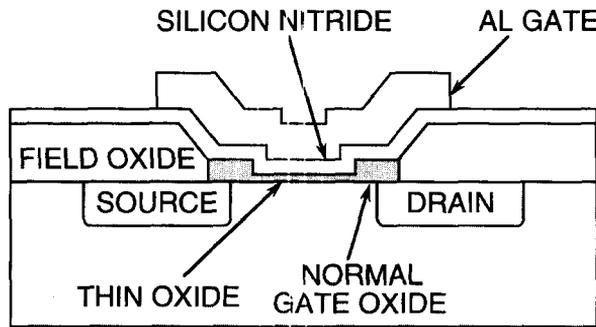


Figure 8.56 Cross-sectional view of a stepped-gate MNOS memory cell.

the disturb problem on deselected device B in Figure 8.52. However, the stepped-gate retention characteristics are satisfactory only if read voltages less than approximately 2 V are used.

Gated-Access Silicon Nitride Memory Cells

To improve the retention time of silicon nitride nonvolatile memories, gated-access silicon nitride memory cells were developed. Either one or two access gates can be used. Figure 8.57 shows a gated-access memory cell with two access gates.¹¹³ The

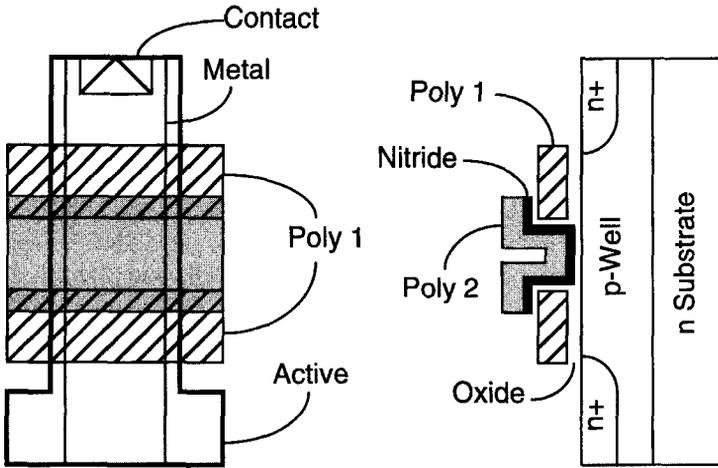


Figure 8.57 Cross-sectional and top view of a dual-gated-access SNOS memory cell.

access gates allow the threshold voltage of the silicon nitride portion of the structure to be adjusted using an implant into the channel region so that the high-conductance memory state is depletion mode and the low-conductance memory state is enhancement mode. In a memory array, the access gate between the bit line and the silicon nitride transistor isolates a deselected silicon nitride transistor along the same bit line from the bit line so that only the selected memory cell is connected to the sense amplifier via the bit line. The second access transistor on the source side of the structure allows the source region to be common for all the memory cells in the array, eliminating one decoded line across the array.

The programming and read voltages for the dual-gated-access silicon nitride memory cell are shown in Table 8.8.¹¹³ A voltage must be applied to the bit line for the read operation so that the electric field is not strictly zero across the entire channel region. This voltage can be minimized to provide for less disturb during the read cycle.

TABLE 8.8 Programming and Read Voltages for the Dual-Gated-Access SNOS Memory Cell, Where V_{hv} is the High Voltage Used for Programming and Erase

	Read	Program	Program Inhibit	Erase	Erase Inhibit
Drain	Data	0	V_{hv}	V_{hv}	V_{hv}
Select	5	V_{hv}	V_{hv}	Float	Float
Storage	0	V_{hv}	V_{hv}	0	V_{hv}
Isolate	5	0	0	5	5
Source	0	0	0	V_{hv}	V_{hv}
P well	0	0	0	V_{hv}	V_{hv}

The read voltage for gated-access silicon nitride memory cells is zero, the same as for the standby gate voltage condition. At the cost of a larger memory cell, data retention is improved by eliminating the external field generated by a positive read voltage applied to the gate of the memory device. Silicon nitride devices with more than 10 years' retention at 140°C can be fabricated using the gated-access technique.

Memory circuits have been developed without the second access transistor.^{100,114} The tradeoff in silicon area between using a dual-access-gated SNOS memory cell with one fewer decoded line per column of memory cells compared to a single-gated-access SNOS memory cell is a function of the design rules. With better deposition techniques for silicon nitride, shorting between the gate electrode and source or drain of the silicon nitride device has been reduced, allowing these single-gated-access memories to be developed.

Shadow RAM Silicon Nitride Memory Cells

Shadow RAM memory cells operate as standard SRAM memory cells, except that before power is removed, the state of the SRAM memory cell is stored in nonvolatile devices. During powerup, the nonvolatile devices restore the state that the SRAM memory cell had immediately before powerdown. The shadow RAM architecture has the advantage that the nonvolatile devices are written only during powerdown; at all other times, the memory cell is read and written as a standard SRAM cell. This architecture greatly increases the endurance of the memory since the endurance relates only to the number of powerdown cycles, not the number of write cycles on a given memory cell. This architecture also gives fast read/write times since the read access times are those of an SRAM. During powerdown, the nonvolatile devices have no voltage across them, so the retention time is long.

Successful commercial shadow RAMs have been implemented using either capacitor or transistor storage devices for both floating-gate or nitride technologies.^{106,115,116} The concepts are discussed here in the context of silicon nitride-based nonvolatile devices, but are applicable to either technology.

The disadvantages of the capacitance shadow RAM SNOS memory cell are that the cell size is large and the restored data are in the complement state. Therefore, special circuitry is required to perform a complement operation to restore the original data after powerup.

The nvSRAM silicon nitride memory cell is similar to the capacitive shadow RAM memory cell, except instead of capacitors providing differential loading on the SRAM latch nodes, SNOS transistors feed differential currents into the latch nodes during powerup.^{106,116} A schematic of an nvSRAM silicon nitride memory cell is shown in Figure 8.58, where dual-gated silicon nitride devices are used.¹⁰⁶ Figure 8.59 shows the operation of the nvSRAM memory cell.¹⁰⁶

The advantages of the nvSRAM memory cell compared to the capacitive shadow-RAM memory cell are that the silicon nitride transistors are smaller than the silicon nitride capacitors, resulting in a smaller memory cell size and that the memory powers up in the same state that existed during powerdown. By adjusting the layout so that some area is shared by the high value resistors and the silicon nitride

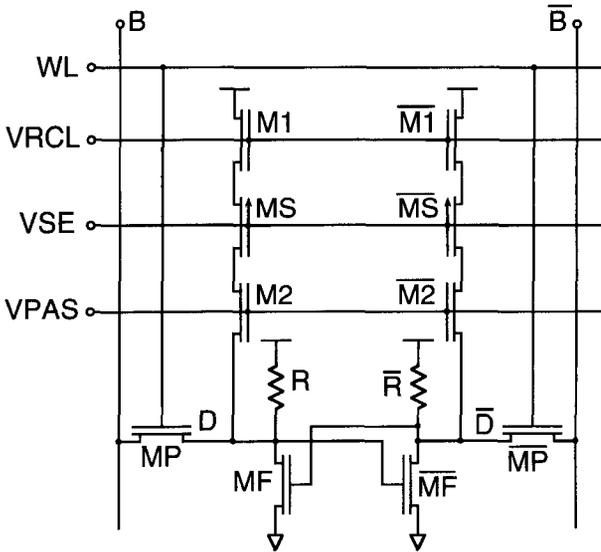


Figure 8.58 Circuit schematic for an nvSRAM memory cell.

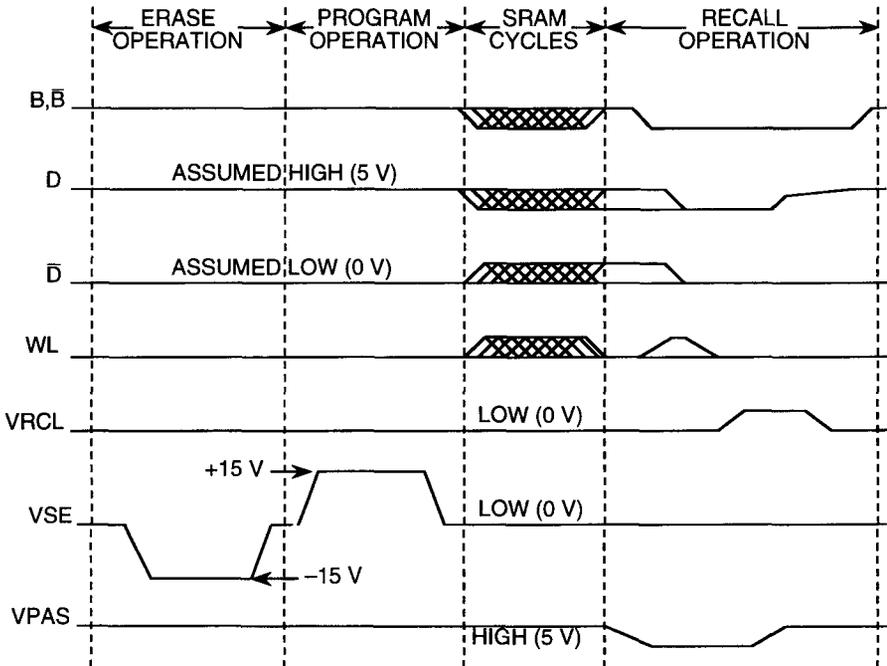


Figure 8.59 Operation of an nvSRAM memory cell. The operation labeled “recall” is also referred to as the “restore” operation.

transistors, a memory cell only approximately 25% larger than a standard SRAM cell can be realized.

8.6.4 Reliability of Silicon Nitride Nonvolatile Memory

Memories utilizing the silicon nitride technology can have high reliability. This is because every bit in a memory array can be nondestructively electrically screened for retention by measuring the slow log time decay of threshold voltage.

To test the long-term retention of silicon nitride memory cells, all memory cells in the array are programmed to a given memory state. The memory is then subjected to a thermal bake of less than 24 h similar to that used for testing of EPROMs. Using circuitry on the chip designed to allow the voltage on the gate electrode above the silicon nitride to be varied, the threshold voltage of every memory cell in the array can be determined. Those memory cells that have unacceptable retention decay rates can be readily identified, and the chips containing them can be discarded. The remaining population of chips is very reliable, and it is difficult to find even a single memory cell failure for a nonvolatile function during stringent product qualification procedures.

Because the charge is stored within an insulator in the silicon nitride technology, the technology is relatively immune to leakage paths through the insulator. Rarely are gate-to-substrate shorts observed in the silicon nitride technology.

8.6.5 Summary of Silicon Nitride Memory

The use of silicon nitride memory cells is limited by the relatively large memory cell size because two or three transistor memory cells have been required to meet data retention requirements. Because of this and the relatively low volume of production of silicon nitride memories, silicon nitride nonvolatile memories have not been cost-competitive with flash memories. However, silicon nitride memory technology has a proven advantage in applications requiring radiation hardness, which is of increasing importance with the proliferation of communication satellites. Additionally, silicon nitride nvSRAMs have replaced battery-backed SRAMs in certain applications.

The simplicity of the silicon nitride process makes this technology attractive for some embedded applications. The scalability of the voltages needed for write operations and the relative immunity of the nitride technology to point defects suggest that it may find application in deep-submicrometer-scale technologies, although perhaps in the form of new cell configurations.

8.7 FERROELECTRIC MEMORY

8.7.1 Introduction to Ferroelectric Memory

When a capacitor incorporates a dielectric consisting of a ferroelectric material, the capacitor structure can serve as a memory device. The ferroelectric material has the

property that certain atoms in the crystal lattice can reside in each of two stable positions, resulting in a spontaneous polarization that can be reversed by application of an applied electric field.^{117–120} Reversal of the polarization by applying the appropriate electric field to the capacitor results in the generation of a switched current that can be detected by a sense amplifier. Such switching of the ferroelectric capacitor results in destructive readout of the data, after which the original data must be rewritten to the capacitor, which is essentially the same as for magnetic core or DRAM memories. Alternatively, since the capacitance–voltage characteristic of a device is different for different polarization states, nondestructive read schemes can be used as well, although these are more difficult to implement than the destructive readout schemes.

Ferroelectric materials have been known for many decades. The earliest memories used ferroelectric capacitors at the cross-points of the rows and columns of the memory,¹²¹ but voltages on the deselected capacitors caused disturb problems. The disturb problems were solved by using back-to-back diodes as access devices to the word lines.¹²² Despite additional ferroelectric research activity in the 1950s and 1960s, materials incompatibility problems between the elements contained in the ferroelectric and typical semiconductor processes prevented their use in integrated circuits. More recently, isolating the ferroelectric capacitors from the silicon and locating them within the interlevel dielectric layers has resulted in the fabrication of successful IC ferroelectric memories.^{123–126}

Because ferroelectric materials typically have high dielectric constants, contacting schemes have involved noble metals, such as platinum, or conductive oxides such as iridium oxide or ruthenium oxide. Therefore, special semiconductor processing is required to include ferroelectric materials in typical semiconductor processes. Concern about cross-contamination of the elements used in ferroelectric materials in semiconductor fabrication facilities has necessitated the use of separate semiconductor facilities to perform the ferroelectric processing and all subsequent processing steps. More studies are required to determine whether, in fact, the elements used in the ferroelectric process have an adverse effect on other products built in the same semiconductor fabrication facility.

Ferroelectric memories have certain advantages over floating gate and silicon nitride memories. Typically, ferroelectric memories can be programmed at lower voltages, eliminating the need for write voltage charge pumps and high-voltage junction technology. In addition, ferroelectric memories can be programmed orders of magnitude faster than other types of nonvolatile memories and they have significantly higher levels of endurance after write/erase cycling.¹²⁷ Certain types of ferroelectric materials have fatigue-free characteristics, or infinite endurance.¹²⁸

8.7.2 Physics of Ferroelectric Technology

Ferroelectric Materials and Structure

Typical ferroelectric materials that are used for semiconductor devices have the perovskite crystal structure shown in Figure 8.60.¹³³ These materials have the chemical structure ABO_3 , where B is a relatively small cation located at the center of

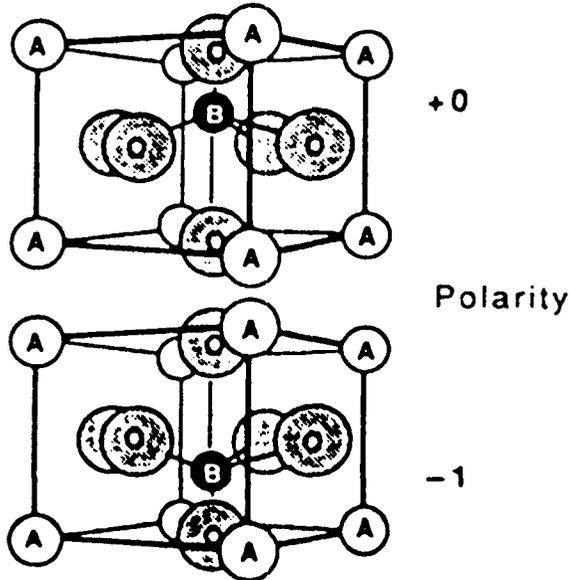


Figure 8.60 Perovskite crystal structure. ABO_3 , of a ferroelectric material. The two representations show the B cation in its two stable states.

the unit cell, A are relatively large cations located at the corners of the unit cell, and O are oxygen anions located at the face-center positions. Above a critical temperature, known as the *Curie temperature*, the perovskite structure assumes a cubic symmetry and is paraelectric; that is, it exhibits no ferroelectric behavior. Although certain ferroelectric materials exhibit high-temperature ferroelectric behavior and low-temperature paraelectric behavior, these materials are not discussed here since they are not very useful in integrated circuits. Likewise, certain ferroelectric materials, such as Rochelle salts, exhibit ferroelectric behavior over a range of temperatures with both high- and low-temperature ferroelectric phases, but they are not discussed here.

Below the Curie temperature, the large A cations cause the lattice to distort to orthorhombic, rhombohedral, or tetragonal structures. The orthorhombic structure forms as the cube stretches along a face diagonal. The rhombohedral structure forms as the cube stretches along a body diagonal. The tetragonal structure forms as the cube stretches along one side of the cube, forming a longer axis (by definition, the *c* axis), and shorter axes (by definition, the *a* axes). In each case, the polarization of the B cation occurs in the direction of the stretched axis.

Since the ferroelectric properties do not exist above the Curie temperature, it is important that the Curie temperature for a ferroelectric material be sufficiently high for the material to have wide application in semiconductor circuits. This is particularly true when high-temperature, accelerated testing is used to screen products for defects and reliability. There is evidence that the Curie temperature for a thin-film ferroelectric material may be different from that for a bulk ferroelectric

material, but this is not well documented. Therefore, the ferroelectric behavior of a material should be evaluated on thin-film structures to determine its applicability to integrated circuits.

Typical perovskite ferroelectric materials are BaTiO_3 , PbTiO_3 , PZT ($\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$), PLZT ($\text{Pb}_{1-x}\text{L}_x\text{ZrO}_3$), PMN ($\text{PbMg}_{1-x}\text{Nb}_x\text{O}_3$), SBT ($\text{SrBi}_2\text{Ta}_2\text{O}_9$), SBN ($\text{SrBi}_2\text{Nb}_2\text{O}_9$), and SBTN [$\text{SrBi}_2(\text{Ta}_{1-x}\text{Nb}_x)_2\text{O}_9$], to name a few. The B cations are Ti, Zr, Mg, Ta, and Nb. The latter three ferroelectric materials have the interesting property that they spontaneously layer themselves at the atomic level, as shown in Figure 8.61.¹²⁸ For SBT (strontium bismuth tantalate), a single atomic layer of

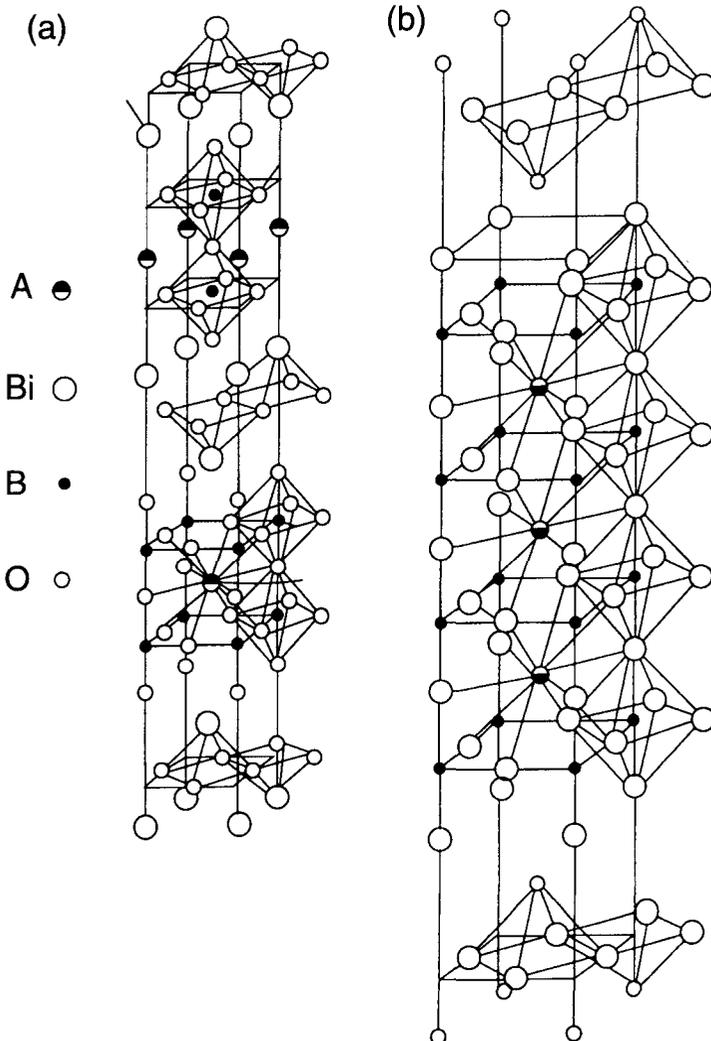


Figure 8.61 Crystal structure for a layered perovskite ferroelectric. For SBT (strontium bismuth tantalate), A = Sr and B = Ta.

Bi_2O_3 forms, then two layers of SrTaO_3 unit cells and then another layer of Bi_2O_3 , and so forth. This layering causes these latter ferroelectric materials to show virtually fatigue-free behavior.

Hysteresis and Retention

A typical hysteresis curve of a ferroelectric capacitor is shown in Figure 8.62.¹³⁰ This curve depicts the response of the polarization, P , to the externally applied electric field, E . The hysteresis curve saturates at P_{sat} when the maximum alignment of the spontaneous polarization occurs. When the electric field is removed instantaneously after reaching P_{sat} , the electronic polarization associated with the linear capacitance component decreases to zero, and the spontaneous polarization, P_S , remains. Then, within milliseconds usually, the polarization decays to the remnant polarization, P_r . For much longer times, the polarization is observed to decay linearly with the log of time for many orders of magnitude of time, similar to the retention decay of silicon nitride devices. A retention curve for SBT is shown in Figure 8.63.¹²⁵ The typical decay rate for layered perovskite, ferroelectric materials is considerably smaller than that for silicon nitride devices, so retention only under worst-case conditions, such as after imprint, is of concern.

From Maxwell's equations, the displacement charge density, D , is related to the polarization and applied electric field by

$$D = \epsilon_0 E + P \quad (8.14)$$

where ϵ_0 is the permittivity of free space. For most ferroelectric materials used in

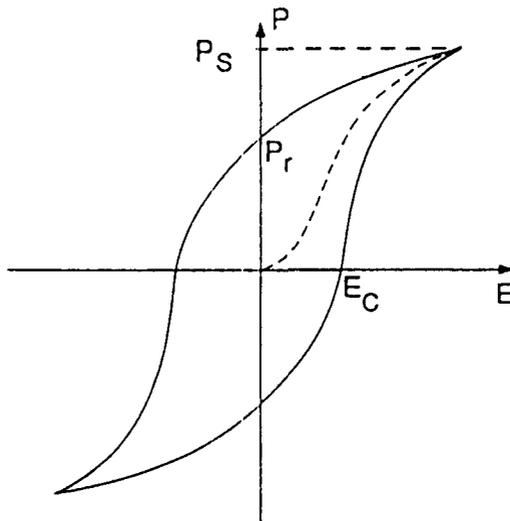


Figure 8.62 Typical hysteresis curve for a ferroelectric material. The electric field, E_c , is the coercive field, the polarization P_S is the spontaneous polarization, and the polarization P_r is the remnant polarization.

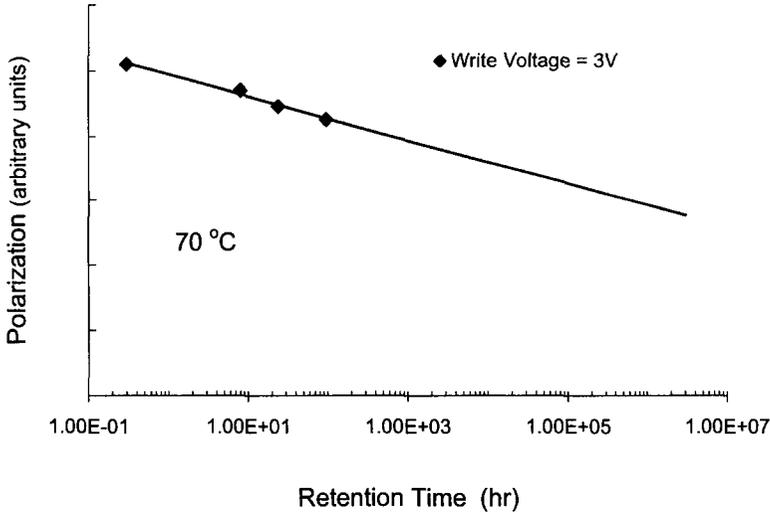


Figure 8.63 Typical log time retention decay for an SBT ferroelectric capacitor.

semiconductor devices, $\epsilon_0 E \ll P$ and therefore $D \sim P$. Consequently, the hysteresis curve shown in Figure 8.62 is also approximately the displacement charge versus electric field.

Ferroelectric Switching Time and Current

Various models have been presented to explain the time dependence of the switching of the hysteresis curve. The switching dynamics of ferroelectric devices are complex and depend on the orientation of the grains in the polycrystalline film, defects within the grains, at grain boundaries, and at the electrodes, and the pinning of domains. One of the models for the switching current and switching time of ferroelectric devices is that of Ishibashi.¹³¹ The switching current versus time is

$$I(t) = 2P_s A \left(\frac{n}{t_0}\right) \left(\frac{t}{t_0}\right)^{n-1} \exp\left[-\left(\frac{t}{t_0}\right)^n\right] \tag{8.15}$$

where P_s is the spontaneous polarization that depends in part on the applied voltage, A is the electrode area, n is a parameter (that may have a noninteger value) related to the number of reversed domains that come from latent nuclei and thermally activated nuclei per unit area, and t_0 is a characteristic time related to the same latent nuclei and thermally activated nuclei. Equation 8.15 fits polarization reversal reasonably well for a few different ferroelectric materials. However, different values of n ranging between 1 and 2 are required to fit the data, and it is not understood why the specific noninteger values are obtained.

It is difficult to determine theoretically the dependence of the switching time on applied field. Empirically, the switching time has been determined for certain

ferroelectric materials. For example, for the switching time t_s , in BaTiO₃, it has been found that an exponential law,¹³²

$$t_s \sim t_0 \exp\left(\frac{E_0}{E}\right) \quad (8.16)$$

where t_0 and E_0 are constants or a power law, in the case of high field polarization reversal¹³³

$$t_s \sim E^{-3/2} \quad (8.17)$$

fit the switching-time dependence on applied electric field, E .

The switching times of ferroelectric materials can be less than one nanosecond. It is difficult, in practice, to measure the true switching times of ferroelectric capacitors, so the switching models are difficult to confirm.

The switching current is important for ferroelectric memories that utilize destructive readout schemes.¹³⁴ Figure 8.64 shows the switching currents of a ferroelectric capacitor for a step function in voltage applied across the capacitor. The switching current is given by

$$I(t) = C_L \frac{dV}{dt} + \frac{dQ_s}{dt} \quad (8.18)$$

where C_L is the linear capacitance and Q_s is the switched charge associated with the polarization change of the capacitor. Usually the resistive components are small for ferroelectric capacitors, so they are not included in Eq. 8.18. The upper curve shown in Figure 8.64 represents switching the capacitor from the opposite polarization state. The lower curve shown in Figure 8.64 represents partial switching of the capacitor from the relaxed same-state polarization. The leading edges of both curves are related to the rise time of the applied voltage pulse. The difference between the two curves represents the signal that is used to detect the state of the ferroelectric capacitor.

Polarization Temperature Dependence

Typical polarization of a ferroelectric material useful for semiconductor devices versus temperature is shown in Figure 8.65, and the corresponding dependence of the permittivity, ϵ_r , is shown in Figure 8.66.¹³⁰ As the ferroelectric material is heated through the Curie temperature, T_c , the ferroelectric material typically undergoes a first-order phase transition into the paraelectric phase. The permittivity becomes large at the Curie temperature and typically exhibits Curie–Weiss behavior at higher temperature as given by

$$\epsilon_r = \frac{C}{T - T_c} \quad (8.19)$$

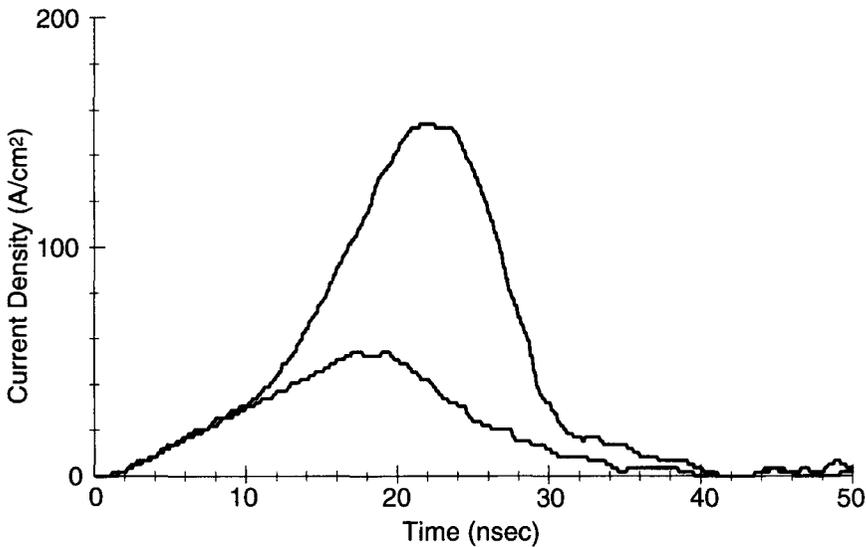


Figure 8.64 Switching currents for a ferroelectric capacitor. The upper curve represents the switching current for a capacitor where switching occurs from the opposite state. The lower curve represents the switching current for a capacitor where partial switching occurs from the relaxed same state. (Data courtesy of Symetrix Corp.)

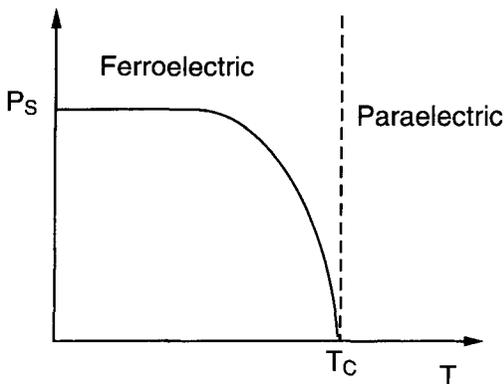


Figure 8.65 Polarization versus temperature for a typical ferroelectric material used in integrated circuits. T_C is the Curie temperature.

where C is the Curie constant. Since the capacitance is directly proportional to the dielectric constant, the C - V curves have the same basic shape as the curve shown in Figure 8.66.

Fatigue and Imprint

Many ferroelectric thin films show degradation of the polarization as a function of the number of write/erase cycles where the polarization is repeatedly reversed. This

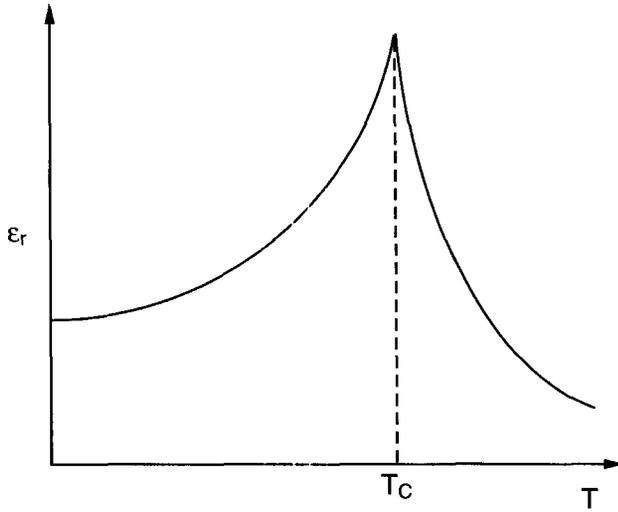


Figure 8.66 Permittivity of a ferroelectric material versus temperature.

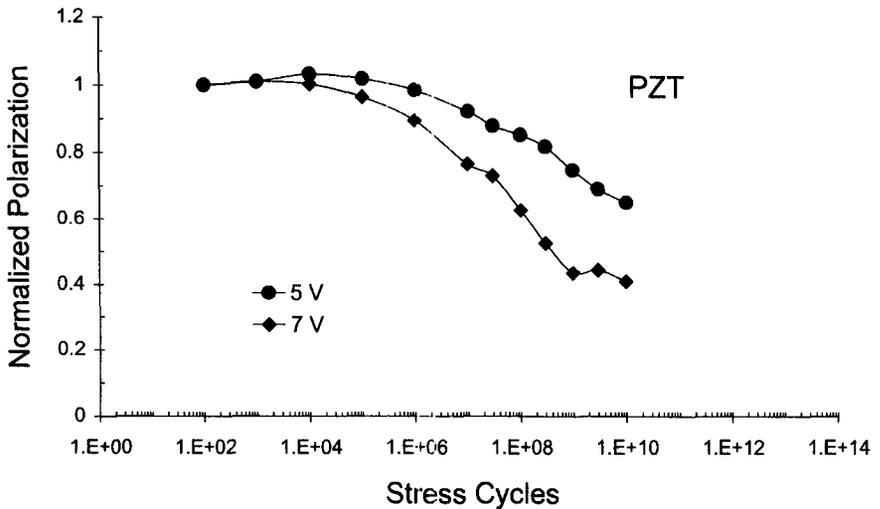


Figure 8.67 Fatigue behavior of PZT.

fatigue degradation is not observed if unipolar pulses are applied to the ferroelectric film where the polarization is not reversed. The remnant and maximum polarization of the film decreases and the coercive field may increase or decrease.¹³⁵⁻¹³⁸ The squareness of the hysteresis loop may also decrease.¹³⁹ Layered-perovskite ferroelectric films have the unique property that fatigue degradation is greatly

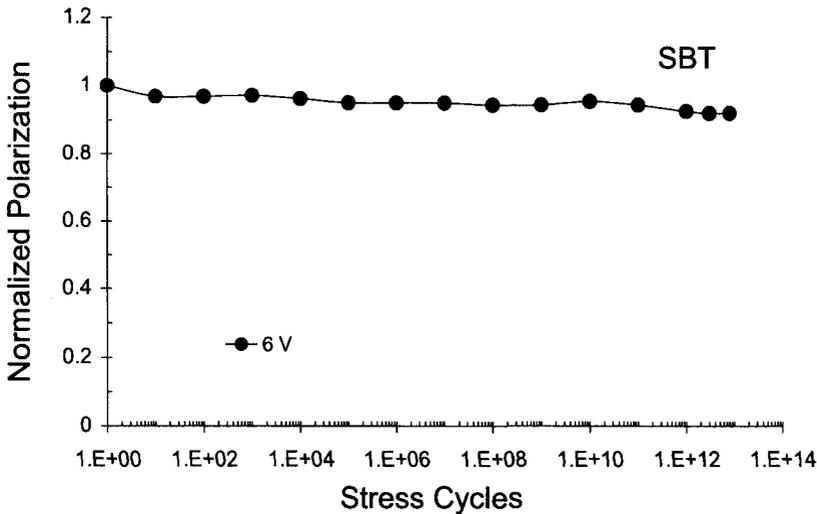


Figure 8.68 Fatigue behavior of SBT.

reduced or nonexistent.¹²⁸ Figure 8.67 shows fatigue curves for PZT for three different stress levels.¹⁴⁰ In Figure 8.68 a fatigue curve is shown for SBT for a stress electric field higher than any of those shown in Figure 8.67.¹²⁵ After fatigue, the hysteresis curves for PZT lose their squareness. In contrast, the hysteresis curves for SBT are virtually unchanged from those before fatigue. The fatigue and imprint degradation in PZT can be reduced by using certain conductive oxide electrodes, such as lanthanum strontium cobalt oxide (LSCO).¹⁴¹ The fatigue in PZT can also be reduced by doping the PZT with lanthanum or niobium.¹⁴² However, the doped PZT films still exhibit hysteresis loops that lose their squareness after write/erase cycling.

If a ferroelectric capacitor is subjected to programming pulses of the same polarity, the ferroelectric state corresponding to that polarity is reinforced. This phenomenon is called *imprint*. The manifestation of imprint is similar to that of repeatedly erasing silicon nitride devices discussed earlier in this chapter. If a ferroelectric film is repeatedly pulsed with unipolar voltage pulses, it may become more difficult to switch the capacitor to the opposite state, and once switched to the opposite state, the retention time of the capacitor may be degraded. Therefore, to ensure that a ferroelectric memory meets the retention time specification, it is important to perform the tests for worst-case imprint conditions.

Imprint is a function of the ferroelectric material used in the ferroelectric capacitor. SBT, a layered perovskite, has lower imprint than PZT. If niobium is added to SBT, the imprint of the resulting SBTN capacitor is approximately one-third that of SBT. It is not understood why the substitution of approximately 30% of the Ta atoms with Nb atoms causes this effect. It is believed that the imprint is due, in part, to charge injection at the electrodes, the behavior of the depletion regions that form near the electrodes, and the Debye or screening length in the ferroelectric material.

Reliability

Since ferroelectric memories are an emerging technology, statistical reliability levels have not been published. The reliability of ferroelectric devices can be affected by leakage current through the ferroelectric material, catastrophic dielectric breakdown, time-dependent dielectric breakdown, and imprint. Since ferroelectric materials used in integrated circuits are polycrystalline, typical electrical integrity is poorer than for amorphous materials, such as silicon nitride of the same film thickness. The electrical integrity of the ferroelectric thin films is dependent on the specific ferroelectric material, processing sequences, thickness, and thermal annealing and crystallization temperatures. With proper process design, ferroelectric memories can have high reliability levels.

Scaling

Scaling of ferroelectric devices is related to the minimum ferroelectric-layer thickness and the inherent switching capability of the ferroelectric film. The switching capability is relatively independent of the film thickness, but the electrical integrity is a strong function of the film thickness. For ferroelectric films, such as PZT, the depletion layers near platinum electrodes can penetrate into the ferroelectric material by as much as 100 nm. Polarization switching does not occur in these depletion regions. Therefore, to maintain a polarization switching characteristic, the film must be at least 250–300 nm thick. This places a lower limit of approximately 4 V on the programming voltage for a PZT film using platinum electrodes.

In SBT, the film can be scaled below 100 nm and still maintain its ferroelectric switching properties. This corresponds to a lower limit on the programming voltage of approximately 1 V. With such low-voltage switching, it is not necessary to use charge pumps to generate the programming voltage on chip. However, caution must be exercised in the design of the integrated circuit for powerup and powerdown sequences so that small voltages are not inadvertently applied to the ferroelectric capacitors and the data disturbed.

Radiation Hardness

The ionizing radiation hardness of ferroelectric devices is approximately the same as that for silicon nitride devices. This is because of the relatively high trap densities in the ferroelectric films. Radiation failure of the ferroelectric film is related to retention failures as in the case of silicon nitride films.¹⁴³

8.7.3 Ferroelectric Memory Cells

Ferroelectric memory cells can utilize any of the asymmetrical or nonlinear properties of a ferroelectric material. The three types of ferroelectric memory cells are those with DRAM-like architectures that incorporate destructive data readout sensing, those with shadow SRAM-like architectures, and those with ferroelectric transistor architectures, that incorporate nondestructive readout data sensing. In the DRAM-like configuration, the state of the memory cell is determined

by whether a ferroelectric capacitor switches when a voltage is applied across the capacitor.

Ferroelectric transistors have been difficult to fabricate because of the material incompatibilities between the ferroelectric materials and silicon substrate. However, theoretically such memory cells should be possible. As in silicon nitride devices, conventional ferroelectric transistors must be programmed across the entire channel length. This necessitates the addition of a pass transistor to each memory cell.

Destructive Readout Ferroelectric Memory Cells

Several versions of destructive readout ferroelectric memory cells have been proposed. Perhaps the first of these was the ferroelectric memory invented by Anderson in 1956 and shown in Figure 8.69.¹²² This memory consisted of an array of ferroelectric capacitors, a switching device connected to a bit line, and a switched plate line. Since transistors were not commonly available at the time of this invention, back-to-back diodes were used as the switching device between the ferroelectric capacitor and word lines. The switching devices prevented the half-program disturb voltages from appearing across the deselected memory transistors because the back-to-back diodes turned on at a higher voltage than the half-voltage. Turning off the switching device before the programming voltage was removed allowed the programming voltage to remain on the plate of the capacitor until it leaked off through the surrounding insulators or the ferroelectric film itself. This provided a reinforced polarization of the memory capacitor, resulting in improved retention characteristics.

The more modern ferroelectric memory cell consisting of one transistor and one ferroelectric capacitor (1T/1C), as shown in Figure 8.70,¹²⁶ is a direct descendant of

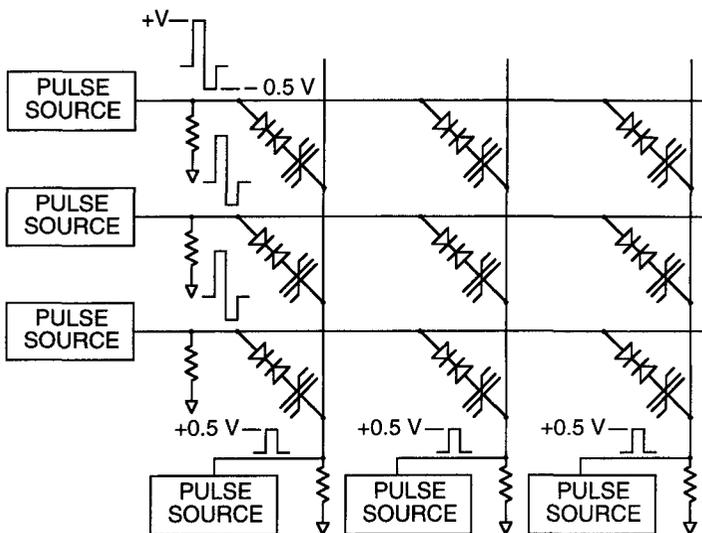


Figure 8.69 Diagram of the Anderson ferroelectric memory array.

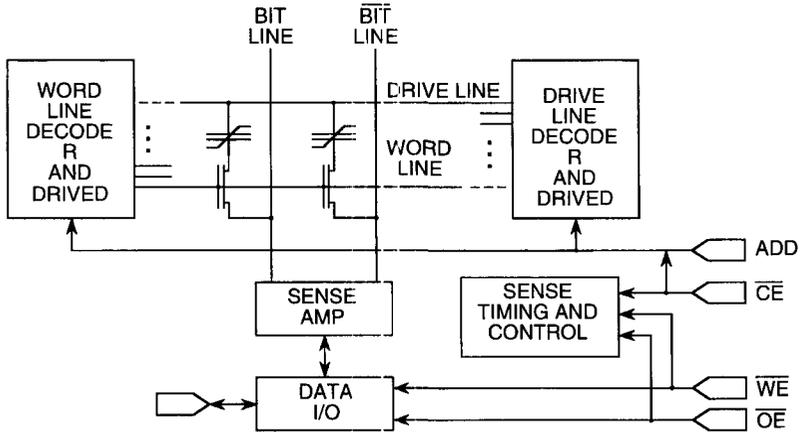


Figure 8.70 Chip architecture of a one transistor, one capacitor ferroelectric memory cell. The capacitor may be fabricated above the CMOS transistors either over the field oxide or over the CMOS pass transistor itself.

the Anderson ferroelectric memory cell in that the back-to-back diodes are simply replaced by a transistor.^{129,144} It is similar to a DRAM memory cell with the exception that the plate line is decoded through the array so that the voltage on it can be varied, whereas in a DRAM architecture the plate line is connected to either V_{DD} or V_{SS} . Ferroelectric memory cells that do not require switching of the plate line have also been developed.¹⁴⁵

Operation of the 1T/1C ferroelectric memory cell is shown in Figure 8.71.¹²⁹ During read sensing, when the sense amplifier is latched, the bit lines are automatically driven to the appropriate voltages to restore the data in the ferroelectric capacitors.

One design difficulty of the 1T/1C ferroelectric memory cell is determination of the appropriate cell:bit line capacitance ratio. While this calculation is straightforward for a DRAM, the ferroelectric capacitance changes dramatically with voltage as shown in Figure 8.72. Since accurate ferroelectric models for circuit simulation are not readily available, the cell:bit line capacitance ratio has generally been determined empirically.

Because of materials incompatibility problems between the platinum electrodes for ferroelectric capacitors and polysilicon, ferroelectric 1T/1C memory cells have been fabricated by using either local interconnect or aluminum straps to connect the capacitor. This results in large memory cell sizes compared to DRAM memory cells. Stacked PZT capacitor ferroelectric memory cells have been realized by using a titanium nitride barrier between the platinum electrode and the polysilicon plug as shown in Figure 8.73.¹⁴⁶ When an appropriate barrier material is developed that withstands the higher crystallization temperatures and gas ambients required for layered-perovskite ferroelectric films, stacked-capacitor architectures with sizes similar to stacked DRAMs will be possible.

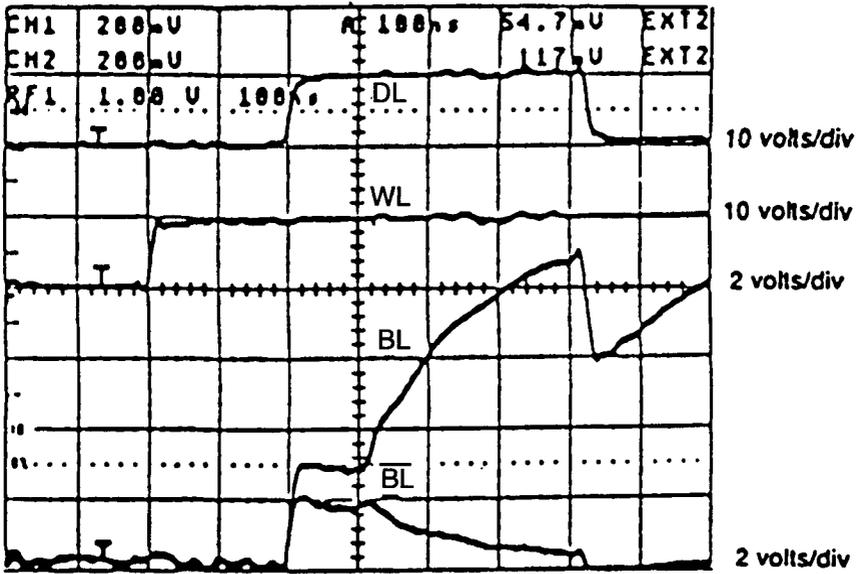


Figure 8.71 Signals for the 1T/1C switched plate ferroelectric memory cell for the read operation.

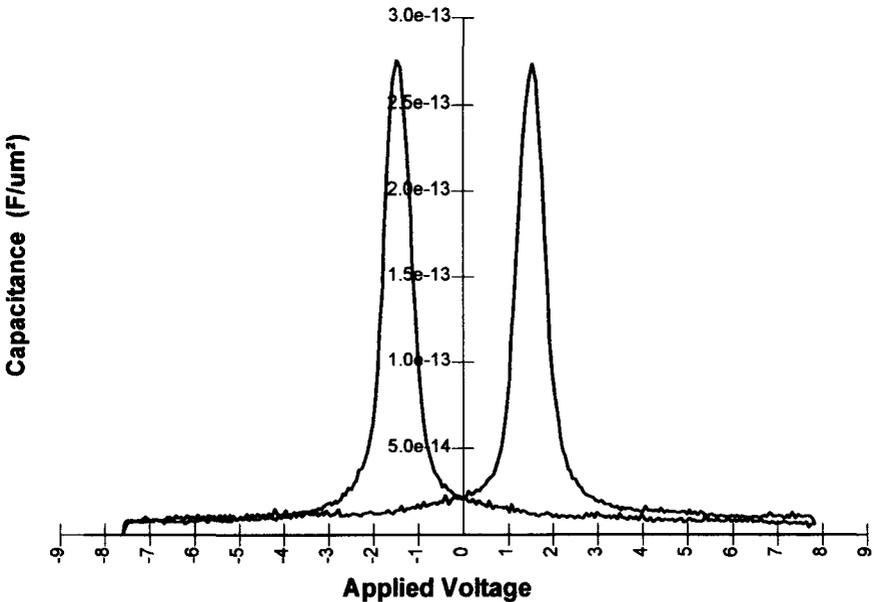


Figure 8.72 Large-signal capacitance/voltage curve for an SBT ferroelectric capacitor. The magnitude of the voltage ramp rate is constant for the measurement. (Data courtesy of Symetrix Corp.)

2. C. Hu, "Lucky Electron Model of Channel Hot Electron Emission," *IEDM Tech. Digest*, 1979, p. 22
3. T. Urai et al., "Simulation of EPROM Programming Characteristics," *Electron. Lett.* **24**, 716 (1990).
4. C. Fiegna et al., "Simple and Efficient Modeling of EPROM Writing," *IEEE Trans. Electron Devices* **ED-38**, 603 (1991).
5. A. Concannon et al., "The Numerical Simulation of Substrate and Gate Currents in MOS and EPROMs," *IEDM Tech. Digest*, 1995, p. 289
6. M. V. Fischetti and S. E. Laux, "Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices Including Band-Structure and Space-Charge Effects," *Phys. Rev. B* **38**, 9721 (1988).
7. T. H. Ning et al., "Emission Probability of Hot Electrons from Silicon into Silicon Dioxide," *J. Appl. Phys.* **48** 286 (1977).
8. E. H. Nicollian and C. N. Berglund, "Avalanche Injection of Electrons into Insulating SiO₂ Using MOS Structures," *J. Appl. Phys.* **41** 3052 (1970).
9. D. Frohman-Bentchkowsky, "A Fully-Decoded 2048-Bit Electrically-Programmable MOS ROM," *ISSCC Digest Tech. Papers*, 1971, p. 80
10. D. Frohman-Bentchkowsky, "FAMOS—a New Semiconductor Charge Storage Device," *Solid-State Electron.* **17**, 517 (1974).
11. Simon Tam et al., "Lucky-Electron Model of Channel Hot-Electron Injection in MOSFETs," *IEEE Trans. Electron Devices* **ED-31**, 1116 (1984).
12. A. T. Wu et al., "A Novel High Speed, 5-Volt Programming EPROM Structure," *IEDM Tech. Digest*, 1986, p. 584.
13. M. Kamiya et al., "EPROM Cell with High Gate Injection Efficiency," *IEDM Tech. Digest*, 1982, p. 741.
14. S. Keeney et al., "Simulation of Enhanced Injection Split Gate flash EEPROM Device Programming," *Microelectron. Eng.* **18**, 253 (1992).
15. J. van Houdt et al., "Study of the Enhanced Hot-Electron Injection in Split Gate Transistor Structure," *Proc. 20th Solid State Research Conf. (ESSDERC 90)*, 1990, p. 261.
16. C. J. Hemink et al., "High Efficiency Hot Electron Injection for EEPROM Applications Using a Buried Injector," *Ext. Abstracts 21st Conf. Solid State Devices and Materials*, 1989, p. 133
17. I. C. Chen et al., "Band-to-band Tunneling Induced Substrate Hot-Electron (BBISHE) Injection: A New Programming Mechanism for Nonvolatile Memory Devices," *IEDM Tech. Digest*, 1989, p. 263.
18. R. H. Fowler and L. Nordheim, "Electron Emission in Intense Electric Fields," *Proc. Roy. Soc. London*, **A119**, 173 (1928).
19. M. Lenzlinger and E. H. Snow, "Fowler-Nordheim Tunneling into Thermally Grown SiO₂," *J. Appl. Phys.* **40**, 278 (1969)
20. J. Suñé et al., "Temperature Dependence of the Fowler-Nordheim Injection from Accumulated n-type Silicon into Silicon Dioxide," *IEEE Trans. Electron Devices* **ED40**, 1017 (1993).
21. J. Maserjian, "Tunneling in Thin MOS Structure," *J. Vacuum. Sci. Technol.* **11**, 996 (1974).

22. G. Lewicki and J. Maserjian, "Oscillations in MOS Tunneling," *J. Appl. Phys.* **48**, 3032 (1975).
23. G. Moglesture "Self-Consistent Calculation of Electron and Hole Inversion Charges at Silicon-Silicon Dioxide Interfaces," *J. Appl. Phys.* **59**, 3176 (1986).
24. J. Suñé et al., "Self-Consistent Solution of the Poisson and Schrödinger Equations in Accumulated Semiconductor-Insulator Interfaces," *J. Appl. Phys.* **70**, 337 (1991).
25. M. Lazoni et al., "Advanced Electrical-Level Modeling of EEPROM Cells," *IEEE Trans. Electron Devices* **ED-40**, 951 (1993).
26. D. J. DiMaria and D. R. Kerr, "Interface Effects and High Conductivity in Oxides Grown from Polycrystalline Silicon," *Appl. Phys. Lett.* **27**, 505 (1975).
27. R. M. Anderson and D. R. Kerr, "Evidence for Surface Asperity Mechanism of Conductivity in Oxide Grown on Polycrystalline Silicon," *J. Appl. Phys.* **48**, 4834 (1977).
28. H. R. Huff et al., "Experimental Observation on Conduction through Polysilicon Oxide," *J. Electrochem. Soc.* **127**, 2482 (1980).
29. R. K. Ellis et al., "Electron Tunneling in Non-Planar Floating Gate Memory Structure," *IEDM Tech. Digest*, 1982, 749.
30. H. A. R. Wegener "Endurance Model for Textured-Poly Floating Gate Memories," *IEDM Tehc. Digest*, 1984, 480
31. A. Roy et al., "Electron Tunneling from Polysilicon Asperities into Poly-Oxides," *Solid-State Electron.* **32**, 655 (1989).
32. R. B. Marcus and T. T. Sheng, "The Oxidation of Shaped Silicon Surfaces," *J. Electrochem. Soc.* **129**, 1278 (1982).
33. D. B. Kao et al., "Two-Dimensional Thermal Oxidation of Silicon—I. Experiments," *IEEE Trans. Electron Devices* **ED-34**, 1008 (1987).
34. D. B. Kao et al., "Two-Dimensional Thermal Oxidation of Silicon—I. Experiments," *IEEE Trans. Electron Devices* **ED-35**, 25 (1988).
35. R. B. Marcus and T. T. Sheng, "Polysilicon/SiO₂ Interface Microtexture and Dielectric Breakdown," *J. Electrochem. Soc.* **129**, 1282 (1982).
36. R. D. Katznelson and D. Frohman-Bentchkowsky, "An Erase Model for FAMOS EPROM Devices," *IEEE Trans. Electron Devices* **ED-27**, 1744 (1980).
37. D. C. Guterma et al., "X-series Approach to High Density 128K and High Speed 32K EPROMs," *ISSCC Digest Tech. Papers*, p. 154.
38. R. Kazerounian et al., "Alternate Metal Virtual Ground EEPROM Array Implemented in a 0.8 μm Process for Very High Density Applications," *IEDM Tech. Digest*, 1991, p. 311.
39. S. Ali et al., "A New Staggered Virtual Ground Array Architecture Implemented in a 4 Mb CMOS EPROM," *Symp. VLSI Circ. Digest Tech. Papers*, 1989, p. 35.
40. A. Gupta et al., "A 54-only 16K EEPROM Utilizing Oxynitride Dielectrics and EPROM Redundancy," *ISSCC. Digest Tech. Papers*, 1982, p. 184.
41. W. Johnson et al., "A 16K Electrically Erasable Nonvolatile Memory," *ISSCC Digest Tech. Papers*, 1980, p. 152.
42. G. Yaron et al., "A 16K EEPROM Employing a New Array Architecture and Designed-in Reliability Features," *IEEE J. Solid-State Circ.* **SC-17**, 833 (1982).
43. S. Mukherjee et al., "A Single Transistor EEPROM Cell and its Implementation in a 512 K CMOS EEPROM," *IEDM Digest Tech. Papers*, 1985, p. 616.

44. S. Tam et al., "A High Density CMOS 1-T Electrically Erasable Non-Volatile (Flash) Memory Technology," *Symp. VLSI Technology*, 1988, p. 31.
45. G. Verma and N. Mielke, "Reliability Performance of ETOX based Flash Memories," *Proc. Int. Relative Physics Symp.* 1988, p. 158.
46. H. Kume et al., "A Flash EEPROM Cell with an Asymmetrical Source and Drain Structure," *IEDM Digest Tech. Papers*, 1987, p. 560.
47. K. Yoshikawa et al., "Comparison of Current Flash EEPROM Erasing Methods: Stability and How to Control," *IEDM Digest Tech. Papers*, 1992, p. 595.
48. M. Van Buskirk et al., "Flash Array with Negative Voltage Gate Erase Operation," U. S. Patent, 5,077,691 (1991).
49. T. C. Ong et al., "Erratic Erase in ETOX Flash Memory Array," *Digest Symp. VLSI Technology*, 1993 p. 83.
50. P. Cappelletti et al., "Failure Mechanisms of Flash Cell in Program/Erase Cycling," *IEDM Tech. Digest*, 1994, p. 291
51. F. C. Schmidlin. "Enhanced Tunneling through Dielectric Films Due to Ionic Defects," *J. Appl. Phys.* **37**, 2823 (1966).
52. C. Dunn et al., "Flash EPROM Disturb Mechanisms," *Proc. 32nd Int. Relative Physics Symp.* 1991, p. 299.
53. Sohrab Kianian et al., "A Novel 3 Volts-Only Small Sector Erase, High Density Flash E²PROM," *VLSI Technol. Symp. Digest Tech. Papers*, 1994, p. 71.
54. S. Mehrotra et al., "Serial 9Mb Flash EEPROM for Solid State Disk Applications," *Digest Tech. Papers, VLSI Circuits Symp.*, 1992, p. 24.
55. D. J. Lee et al., "An 18Mb Serial Flash EEPROM for Solid-State Disk Applications," *Digest Tech. Papers, VLSI Circuits Symp.*, 1994, p. 59.
56. Y. Iwata et al., "A High-Density NAND EEPROM with Block-Page Programming for Microcomputer Application," *IEEE J. Solid State Circ.* **SC-25**, 417 (1990).
57. F. Matsuoka et al., "New Ultra High Density EPROM and Flash EEPROM Cell with NAND Structure Cell," *IEDM Tech. Digest*, 1987, p. 552.
58. M. Momodomi et al., "An Experimental 4-Mbit CMOS EEPROM with a NAND Structure Cell," *IEEE J. Solid-State Circ.* **SC-24**, 1238 (1989).
59. S. Kobayashi et al., "A 3.3 V-Only 16Mb DINOR Flash Memory," *IEEE J. Solid-State Circ.* **SC-29**, 454 (1994).
60. H. Onoda et al., "A Novel Cell Structure Suitable for 3 Volt Operation, Sector Erase Flash Memory," *IEDM Tech. Digest*, 1992, p. 599.
61. J. M. Caywood and B. L. Prickett. "Radiation Induced Soft Errors and Floating Gate Memory," *Proc. 21st Annual Int. Relative Physics Symp.*, 1983, p. 167.
62. C. R. Crowell, "The Richardson Constant for the Thermionic Emission in Schottky Barrier Diodes," *Solid-State Electron.* **8**, 395 (1965).
63. T. Uetsuki, "Study of the Degradation of the Data Retention Characteristics of Floating Gate Type Nonvolatile Memory," *Trans. Inst. Electron. Inform. Commun. Eng. C-II J74C-II*, 218 (1991).
64. R. E. Shiner et al., "Data Retention in EPROMs," *18th Annual Proc. Int. Relative Physics Symp.*, 1980, p. 238.
65. G. Crisenza et al., "Charge Loss in EPROMs Due to Ion Generation and Transport in Interlevel Dielectrics," *IEDM Tech. Digest*, 1990, p. 107.

66. F. C. Hsu and S. Tam, "Relationship between MOSFET Degradation and Hot-Electron-Induced Interface-State Generation," *IEEE Electron Device Lett.* **EDL-5**, 50 (1984).
67. C. S. Jenq et al., "High Field Generation of Electron Traps and Charge Trapping in Ultra-Thin SiO₂," *IEDM Tech. Digest*, 1981, p. 223.
68. Prickett, B. L. et al., "Trapping in Tunnel Oxides Grown on Textured Polysilicon," *Proc. 21st Annual Int. Relative Physics Symp.*, 1983, p. 114.
69. D. A. Baglee and M. C. Smayling, "The Effects of Write Cycling on Data Loss in EEPROMs," *IEDM Tech. Digest*, 1985, pp. 624–626.
70. K. Naruke et al., "Stress Induced Leakage Current Limiting to Scale Down EEPROM Tunnel Oxide Thickness," *IEDM Tech. Digest*, 1988, p. 424.
71. R. Moazzami and C. Hu, "Stress-Induced Leakage," *IEDM Tech Digest*, 1992, p. 139
72. D. J. DiMaria and E. Cartier, "Mechanism for Stress-Induced Leakage Currents in Thin Silicon Dioxide Films," *J. Appl. Phys.* **78**, 3883 (1992).
73. S. Aritome et al., "A Reliable Bi-Polarity Write/Erase Technology in Flash EEPROMs," *IEDM Tech. Digest*, 1990, p. 111.
74. Adam Brand et al., "Novel Read Disturb Failure Mechanism Induced by FLASH Cycling," *Proc. 31st Annual Int. Relative Physics Symp.*, 1993, pp. 127–132.
75. E. F. Runnion et al., "Limitations on Oxide Thickness in FLASH EEPROM Applications," *Proc. 34th Annual Int. Relative Physics Symp.*, 1996, p. 93.
76. C. A. Mead, "Scaling of MOS Technology to Submicron Feature Sizes," *J. VLSI Signal Process.* **8**, 1 (1994).
77. K. Yoshikawa et al., "Flash EEPROM Cell Scaling Based on Tunnel Oxide Thinning Limitations," *Symp. VLSI Technology Digest Tech. Papers*, 1991, p. 79.
78. J. de Blauwe et al., "SILC-Related Effects in Flash E²PROM's-part I: A Quantitative Model for Steady-State SILC," *IEEE Trans. Electron Devices* **45**, 1745 (1998).
79. J. De Blauwe et al., "High Temperature Reliability Behavior of SSI-Flash E²PROM Devices," *IEDM Tech. Digest*, 1997, p. 93.
80. Y. Yamaguchi et al., "ONO Interpoly Dielectric Scaling Limit for Nonvolatile Memory Devices," *Symp. VLSI Technology Digest Tech. Papers*, 1993, p. 85
81. W.-H. Lee et al., "A Novel High K Inter-Poly Dielectric (IPD), Al₂O₃ for Low Voltage/High Speed Flash Memories:Erasing in msec at 3.3 V," *Symp. VLSI Technology Digest Tech. Papers*, 1997, p. 117.
82. T. Kobayashi et al, "A 0.24- μm^2 Cell Process with 0.18- μm Width Isolation and 3-D Interpoly Dielectric Films for 1-Gb Flash Memories," *IEDM Tech. Digest*, 1997, p. 275.
83. H. A. R. Wegener et al., "A New Electrically Alterable Non-Destructive Read-Out (EANDRO) Memory Element," *IEEE IEDM Tech. Digest*, 1967, p. 58
84. J. F. Verwey, "Nonvolatile Semiconductor Memories," *Advances in Electronics and Electron Physics*, Vol. 41, Academic Press, 1976, Chapter VI, p. 249
85. Y. Nishi and H. Iizuka, "Theory of the Switching Behavior of MIS Transistors," in *Nonvolatile Memories*, Appl. Solid State Suppl 2A, Academic Press, 1981, p. 366.
86. I. Lundstrom and C. Svensson, "Properties of MNOS Structures," *IEEE Trans. Electron Devices*, **19**, 826 (1972).
87. D. Frohman-Bentchkowsky and M. Lenzlinger, "Charge Transport and Storage in Metal-Nitride-Oxide-Silicon (MNOS) Structures," *J. Appl. Phys.* **40**, 3307 (1969).

88. C. Svenson and I. Lundstrom, "Theory of the Thin Oxide M.N.O.S. Memory Transistor," *Electron. Lett.* **6**, 645 (1970).
89. E. C. Ross and J. T. Wallmark, "Theory of the Switching of MIS Transistors," *RCA Rev.* **30**, 366 (1969).
90. J. Frenkel, "On Pre-Breakdown Phenomena in Insulators and Electronic Semi-Conductors," *Phys. Rev.* **54**, 647 (1938).
91. F. A. Sewell, Jr. et al., "A Charge Storage Model for Variable Threshold FET Memory Element," *Appl. Phys. Lett.* **14**, 45 (1969).
92. A. K. Sinha and T. E. Smith, "Electrical Properties of Si-N films Deposited on Silicon from Reactive Plasma," *J. Appl. Phys.* **49**, 2756 (1978).
93. B. E. Deal et al., "Electrical Properties of Vapor-Deposited Silicon Nitride and Silicon Oxide Films on Silicon," *J. Electrochem. Soc.* **115**, 300 (1968).
94. K. Lehovc and A. Fedotowsky, "Charge Centroid in MNOS Devices," *J. Appl. Phys.* **48**, 2955 (1977).
95. P. C. Arnett, "Transport Conduction in Insulators at High Fields," *J. Appl. Phys.* **46**, 5236 (1975).
96. S. M. Sze, "Current Transport and Maximum Dielectric Strength of Silicon Nitride Films," *J. Appl. Phys.* **32**, 2951 (1967).
97. E. J. M. Kendall, *Com. J. Phys.* **46**, 2509 (1968).
98. D. Frohman-Bentchowsky, "The Metal-Nitride-Oxide-Silicon (MNOS) Transistor—Characteristics and Applications," *Proc. IEEE* **58**, 1207 (1970).
99. C. M. Svensson, "The Conduction Mechanism in Silicon Nitride Films," *J. Appl. Phys.* **48**, 329 (1977).
100. Y. Yatsuda et al., "Hi-MNOS II Technology for a 64-kbit Byte-Erasable 5-V-Only EEPROM," *IEEE J. Solid-State Circ.* **JSSC-20**, 144 (1985).
101. M. H. White and J. R. Cricchi, "Characterization of Thin-Oxide MNOS Memory Transistors," *IEEE Trans. Electron Devices* **ED-19**, 1280 (1972).
102. P. K. Chaudhari, "Threshold Voltage Degradation of MNOS FET Devices," *J. Electrochem. Soc.* **125**, 1657 (1978).
103. C. A. Neugebauer and J. F. Burgess, "Endurance and Memory Decay of MNOS Devices," *J. Appl. Phys.* **47**, 3182 (1976).
104. M. H. White et al., "Endurance of Thin-Oxide Nonvolatile MNOS Memory Transistors," *IEEE Trans. Electron Devices* **ED-24**, 577 (1977).
105. Y. Hsia, "MNOS LSI Memory Device Data Retention Measurements and Projections," *IEEE Trans. Electron Devices* **ED-24**, 568 (1977).
106. C. E. Herdt and C. A. Paz de Araujo, "Analysis, Measurement, and Simulation of Dynamic Write Inhibit in an nvSRAM Cell," *IEEE Trans. Electron Devices* **ED-39**, 1191 (1992).
107. R. Kondo et al., "Dynamic Injection of MNOS Memory Devices," *Jpn. J. Appl. Phys.* (Suppl.) **19-1**, 231 (1980).
108. P. J. McWhorter et al., "Retention Characteristics of SNOS Nonvolatile Devices in a Radiation Environment," *IEEE Trans. Nucl. Sci.* **NS-34**, 1652 (1987).
109. G. J. Brucker, "Interaction of Nuclear Environment with MNOS Memory Device," *IEEE Trans. Nucl. Sci.* **NS-21**, 186 (1974).

110. J. R. Chricci et al., "The Drain-Source Protected MNOS Memory Device and Memory Endurance," *IEEE IEDM Tech. Digest*, 1973, p. 126
111. C. T. Naber and G. C. Lockwood, *Semiconductor Silicon*, Electrochemical Society, 1973, p. 401.
112. J. E. Brewer, *Proc. Natl. Aerospace Elec. Conf. Record*, 1974, p. 32.
113. A. Lancaster et al., "A 5-V-Only EEPROM with Internal Program/Erase Control," *ISSCC Digest Tech. Papers*, 1983, p. 164
114. T. Hagiwara et al., "A 16 kbit Electrically Erasable PROM Using n-Chanel Si-Gate MNOS Technology," *IEEE J. Solid-Circ.* **JSSC-15**, 346 (1980).
115. D. D. Donaldson et al., "SONS 1K \times 8 Static Nonvolatile RAM," *IEEE J. Solid-State Circ.* **JSSC-17**, 847 (1982).
116. C. E. Herdt et al., "Non-Volatile RAM with Integrated Compact Static RAM Load Configuration," U. S. Patent 5,065,362 (1991).
117. B. Jaffe et al., *Piezoelectric Ceramics*, Academic Press, London, 1971.
118. M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, Clarendon, Oxford, 1977.
119. J. C. Burfoot and G. W. Taylor, *Polar Dielectrics and their Applications*, Univ. Calif. Press, Berkeley, 1979.
120. J. M. Herbert, *Ceramic Dielectrics and Capacitors*, Gordon & Breach, New York, 1985.
121. W. J. Metz and J. R. Anderson, "Ferroelectric Storage Devices," *Bell Lab. Rec.* 335 (1955).
122. J. R. Anderson, "Electrical Circuits Employing Ferroelectric Capacitors," U. S. Patent 2,876,436 (1959).
123. T. Sumi et al., "A 256 kb Nonvolatile Ferroelectric Memory at 3 V and 100 ns," *ISSCC Digest Tech. Papers*, 1994, p. 268.
124. T. Sumi et al., "A 0.9 V Embedded Ferroelectric Memory for Microcontrollers," *ISSCC Digest Tech. Papers*, 1995, p. 52.
125. T. Otsuki and K. Arita, "Quantum Jumps in FeRAM Technology and Performance," *Int. Ferroelectrics* **17**, 31 (1997).
126. R. E. Jones, "Integration of Ferroelectric Nonvolatile Memories," *Solid-State Technol.* **40**, 201 (1997).
127. J. F. Scott and C. A. Paz de Araujo, "Ferroelectric Memories," *Science* **246**, 1400 (1989).
128. C. A. Paz de Araujo et al., "Fatigue-Free Ferroelectric Capacitors with Platinum Electrodes," *Nature* **374**, 627 (1995).
129. J. T. Evans and R. Womack, "An Experimental 512-bit Nonvolatile Memory with Ferroelectric Storage Cell," *IEEE J. Solid-State Circ.* **JSSC-23**, 1171 (1988).
130. L. H. Parker and A. F. Tasch, "Ferroelectric Materials for 64 Mb and 256 Mb DRAMs," *IEEE Circ. Devices* **6**, 17 (1990).
131. Y. Ishibashi, in C. P. de Araujo et al., eds., *Ferroelectric Thin Films: Synthesis and Basis Properties*, Gordon & Breach, The Netherlands, 1996, Chapter 5, p. 135.
132. W. J. Merz, "Switching Time in Ferroelectric BaTiO₃ and Its Dependence on Crystal Thickness," *J. Appl. Phys.* **27**, 938 (1956).

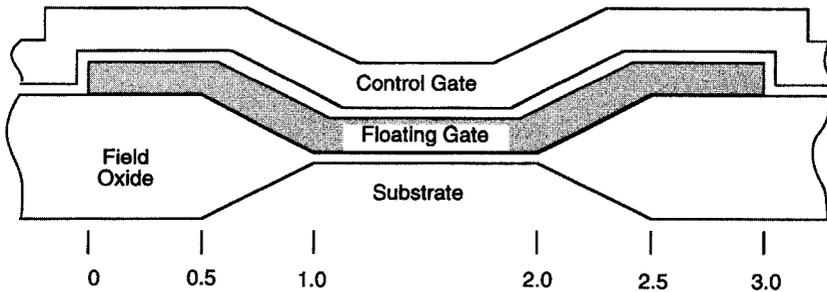
133. H. L. Stadler, "Ferroelectric Switching Time of BaTiO₃ Crystals at High Voltages," *J. Appl. Phys.* **29**, 1485 (1958).
134. T. Mihara et al., "Process Dependent Electrical Characterization and Equivalent Circuit Model of Sol-gel based PZT Capacitors," *Int. Ferroelectrics*, **1**, 269 (1992).
135. G. W. Taylor, "Electrical Properties of Niobium-Doped Ferroelectric Pb(Zr, Sr, Ti)O₃ Ceramics," *J. Appl. Phys.* **38**, 4697 (1967).
136. W. C. Stewart and L. S. Cosentino, *Ferroelectrics* **1**, 149 (1970).
137. D. B. Fraser and J. R. Maldonado, "Improved Aging and Switching of Lead Zirconate-Lead Titanate Ceramics with Indium Electrodes," *J. Appl. Phys.* **41**, 2172 (1970).
138. W. R. Salaneck, *Ferroelectrics* **4**, 97 (1972).
139. J. R. Anderson et al., "Effects of Ambient Atmosphere on the Stability of Barium Titanate," *J. Appl. Phys.* **26**, 1387-1388 (1955).
140. T. D. Hadnagy et al., "The Use of Voltage to Accelerate the Endurance Degradation of PZT Capacitors," *Int. Ferroelectrics* **16**, 219 (1997).
141. R. Dat et al., "Imprint Testing of Ferroelectric Capacitors Used for Non-volatile Memories," *Intl. Ferroelectrics* **5**, 275 (1997).
142. M. Shimizu et al., "Effects of La and Nb Modification on the Electrical Properties of Pb(Zr,Ti)O₃ Thin Films by MOCVD," *Intl. Ferroelectrics* **14**, 69 (1997).
143. J. F. Scott et al., "Radiation Effects on Ferroelectric Thin-Film Memories: Retention Failure Mechanisms," *J. Appl. Phys.* **66**, 1444 (1989).
144. S. S. Eaton et al., "A Ferroelectric Nonvolatile Memory," *ISSCC Digest Tech. Papers*, 1988, p. 130.
145. H. Koike et al., "A 60-n 1-Mb Nonvolatile Ferroelectric Memory with a Nondriven Cell Plate Line Write-Read Scheme," *IEEE J. Solid-State Circ.* **JSSC-31**, 1625 (1996).
146. S. Onishi et al., "A Half-Micron Ferroelectric Memory Cell Technology with Stacked Capacitor Structure," *IEDM Tech. Digest*, 1994, p. 843.
147. S. L. Miller and P. J. McWhorter, "Physics of the Ferroelectric Nonvolatile Memory Field Effect Transistor," *J. Appl. Phys.* **72**, 5999 (1992).
148. T. A. Rabson et al., "Ferroelectric Gate Transistors," *Int. Ferroelectrics* **6**, 15 (1995).

PROBLEMS

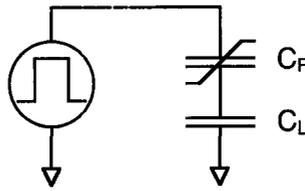
8.1 Let's model an EPROM cell. Consider the EPROM cell whose cross section is shown below. The distances shown are in microns. Assume that the bird's beak has the linear wedge shape illustrated. Assume that the gate oxide thickness is 350 Å, that the interpoly dielectric is oxide of 500 Å thickness, and that the field oxide thickness is 0.6 μm. Assume that the physical gate length is 1.2 μm, that the metallurgical junctions lie 0.15 μm under the gate and that the effective channel length is 0.7 μm. Finally, assume that the floating-gate poly is 0.3 μm thick.

Calculate the approximate values of (a) the control gate-floating gate capacitance, (b) the drain-floating gate capacitance, (c) the floating gate-substrate capacitance, and (d) the floating gate-source capacitance. Neglect

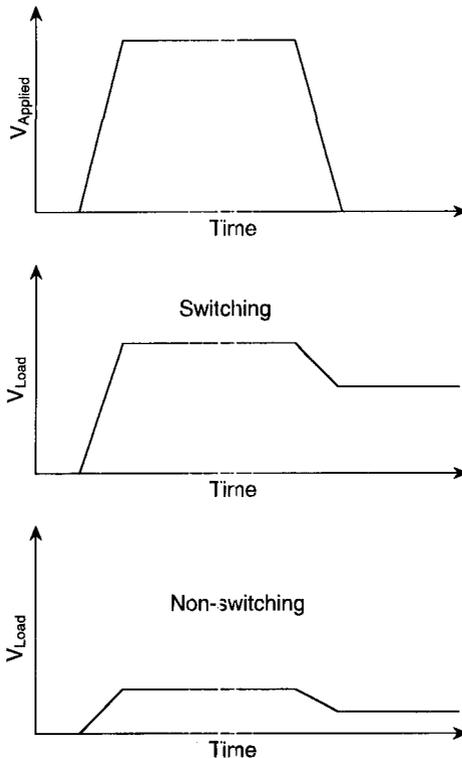
the voltage dependence of the capacitors. (Assign $\frac{1}{2}$ of the channel capacitance to the source and the other half to the drain.)



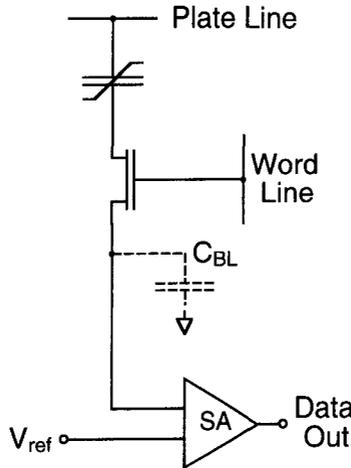
- 8.2 Calculate the control gate–floating gate coupling ratio, R_{CG} , and the drain–floating gate coupling ratio, R_D .
- 8.3 How many electrons are needed to shift the measured threshold by 1 V?
- 8.4 Assume that when the control gate and floating gate are shorted together, the transistor threshold is measured to be 1.5 V. If the control gate is at 12 V and the drain is at 7 V during programming, to what potential can the floating gate be charged while the programming voltages are present? What threshold would be observed after programming under these biases for a drain bias during a reading of 2 V?
- 8.5 Speculate on the possible reasons for the variation in program threshold seen in practice.
- 8.6 A Sawyer Tower circuit consisting of the series combination of a ferroelectric capacitor (C_F) and a linear load capacitor (C_L) is typically used to measure ferroelectric polarization as shown on the right. If a voltage is applied to the circuit, the change in polarization of C_F can be determined by measuring the voltage on C_L since the charge displaced in one is about the same as that displaced in the other. A large enough C_L is chosen that most of the voltage (>90%) drops across C_F . The change in polarization of C_F consists of a linear component due to the device's linear capacitance and a nonlinear component arising when ferroelectric dipoles change states. If C_F is polarized by a voltage pulse (putting most of its dipoles in the same state), a subsequent voltage pulse of the same polarity will reinforce the existing polarization (the nonswitching case) and a pulse of the opposite polarity and sufficient amplitude will reverse the polarization (the switching case).
- (a) Write an equation expressing the relationship between the polarization of the ferroelectric capacitor and the voltage on the load capacitor. (*Hint:* Remember that polarization has units of dipole moment per unit volume, which is the same as charge per unit area.)



- (b) Suppose that a voltage pulse is applied to the Sawyer Tower circuit. Qualitatively draw a graph the shape of the voltage on the load capacitor as a function of time for the switching and nonswitching cases. Draw a graph of the applied pulse for reference. As you think about this, be careful to account for the effect of both the linear and nonlinear polarization. (For help on this one, see Fig. 8.64.)
- (c) Suppose we have a Sawyer Tower circuit with a $10,000\text{-}\mu\text{m}^2$ ferroelectric capacitor and a 10-nF load capacitor. Assume that the remnant polarization (P_r) is $20\ \mu\text{C}/\text{cm}^2$ and that when a 5-V pulse is applied during the switching case, the resulting linear polarization of the ferroelectric is about 25% of the P_r value. Calculate the maximum load voltage when a 5-V pulse is applied to the circuit for both the switching and nonswitching cases.



- 8.7** A basic 1T1C ferroelectric memory cell is shown in the schematic on the right. This is essentially a Sawyer Tower circuit where the parasitic bit line capacitance serves as the load. During a read operation, the word line and the plate line go high. If the ferroelectric capacitor switches, the bit line goes to a relatively high voltage; if it does not switch, it goes to a relatively low voltage. Thus the bit line voltage indicates the state of the cell. Suppose that we have a $2\text{-}\mu\text{m}^2$ ferroelectric capacitor with a worst-case P_r of $4\ \mu\text{C}/\text{cm}^2$ and a best-case P_r of $20\ \mu\text{C}/\text{cm}^2$. If we need a bit line voltage difference between 100 and 500 mV, what would be an acceptable bit line capacitance?



- 8.8** A silicon nitride memory has a data retention characteristic approximated by $V_t = -0.4 + 0.025 \log(t)$ volts for $V_t < 0$ V.
- Assuming that the memory transistor has an $I-V$ characteristic of $I = 58 \cdot (V - V_t)^2\ \mu\text{A}$, the bit line capacitance is 1 pF, the read pulse width is 25 ns, the voltage applied to the memory transistor gate is 0 V, the reference voltage to the sense amplifier is 0 V, the sense amplifier is capable of detecting a 25-mV signal, and the bit line is precharged (starts out) at 0 V, what is the retention time of the memory? Assume that the current is limited by the memory transistor and not the pass gates.
 - Assuming the sense amplifier instead is capable of detecting only a 100-mV signal, what is the retention time of the memory?
 - After 10^6 write erase cycles, the slope of the retention curve doubles. Calculate the retention time for the 25-mV sense amplifier case.
 - Assuming a freak memory cell has a data retention characteristic approximated by $V_t = -0.4 + 0.1 \log(t)$ volts for $V_t < 0$ V, what is the retention time for the 25-mV sense amplifier cases?

PART III

CIRCUIT BUILDING BLOCKS AND SYSTEM-IN-CHIP CONCEPT

CMOS Digital and Analog Building Block Circuits for Mixed-Signal Applications

DAVID PEHLKE

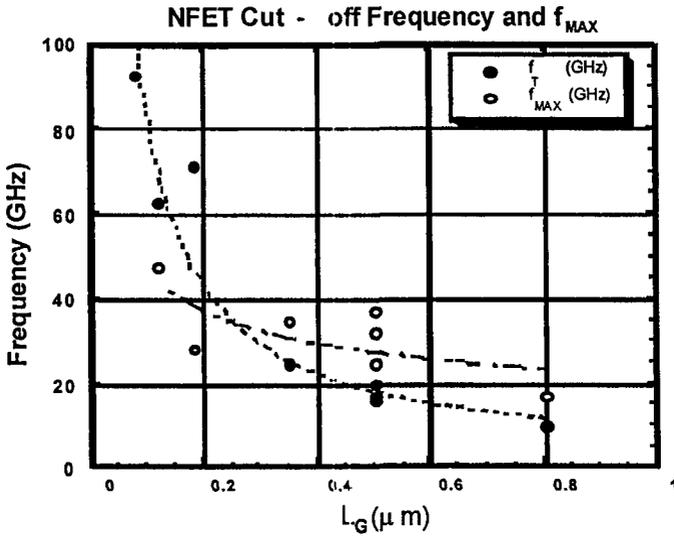
Rockwell Science Center
Thousand Oaks, CA

MAU-CHUNG FRANK CHANG

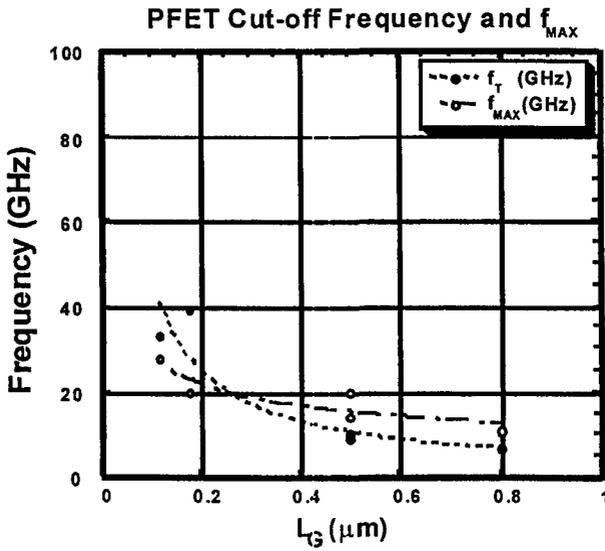
University of California, Los Angeles
Los Angeles, CA

9.1 INTRODUCTION

The development of CMOS technology, driven largely toward higher and higher integration levels of digital circuits, has expanded in capability and density at an exponential rate of advancement. The level of CMOS integration, first observed by Gordon Moore, has been doubling every 18 months, and continues to progress as technological hurdles are simply overrun by the large infrastructure, financial resources, and innovative work being done to meet the ever expanding demands of the semiconductor markets. This technological advancement in CMOS has meant the shrinking of gate lengths gradually down toward the deep-submicrometer dimensions of 0.5, 0.35 and 0.25 μm and now approach 0.18 μm for standard technologies. The lithographic processes to define high-yielding patterns of these smaller dimensions involves not only gate lengths, but also contacts and vias, active areas defining diffusion regions, and interconnect metal linewidths and spacings. The result is a significantly smaller transistor footprint that allows circuits to be made in a smaller area, taking up less “real estate” on a given silicon wafer, with associated reductions in manufacturing cost, even with the added expense of higher-resolution tools to achieve this. This financial incentive toward the shrinking of a given die area has continued to push the limits of to the point now where the gate length and parasitic capacitances of the transistors are so small that their *analog*



(a)



(b)

Figure 9.1 The cutoff frequencies (f_T and f_{MAX}) of (a) NMOS and (b) PMOS as a function of gate lengths.

performance and *digital* switching speed have been significantly improved. The cutoff frequencies for NMOS devices with 0.15 μm gate length has been reported to achieve an f_T of 93 GHz (Fig. 9.1).¹ These reports of state-of-the-art transistor performance demonstrate an evolving trend toward the capability of active CMOS devices for higher- and higher-frequency mixed-signal applications. In this chapter, we first review the progress made in high-speed digital building block circuits and follow that with high-performance analog circuits made by the same CMOS process. This is consistent with the current trend that system-on-a-chip will be realized by multifunction CMOS circuits implemented with a common mixed-signal process.

9.2 CMOS FOR DIGITAL APPLICATIONS

In the 1990s, CMOS became the most favored technology for digital circuit implementation because of its high packing density, low static-power dissipation, and increasingly high operating speed. CMOS inverters and basic logic gates are the building blocks for digital circuits. Once their operations are clearly characterized and modeled, designing more sophisticated structures such as NAND and NOR gates, flip-flops, adders, multipliers, and microprocessors becomes greatly simplified. The electrical behavior of these complex circuits can be fully derived by extrapolating the results obtained for CMOS inverters and basic logic gates. In this chapter, we first discuss the static and dynamic behavior of the CMOS inverter. With that, we then analyze the various differential logic gates that are available to build modern mixed-signal CMOS circuits.

9.2.1 Static Characteristics of CMOS Inverter

The basic CMOS inverter is shown in Figure 9.2a. Transistor Q1 is the n-channel unit and Q2 is the p-channel unit. The two devices are in series, with their drains joined together and gates also connected. The power supply voltage V_{dd} is applied from the p-channel to n-channel source and the output is taken at the common drain. Since we have grounded the source of the NMOS, the input voltage V_i swings between ground and a positive V_{dd} .

The electrical function of an inverter is best expressed by its voltage transfer characteristics (VTC), which plots the output voltage (V_o) as a function of the input voltage (V_i), as shown in Figure 9.2b. Assuming the threshold voltage of NMOS to be $V_t(N)$ and the threshold voltage of PMOS to be $V_t(P)$, we have that for $V_i < V_t(N)$, Q1 is OFF, Q2 is ON, and $V_o = V_{dd}$. Similarly for $V_i > V_{dd} - V_t(P)$, Q2 is OFF, Q1 is ON, and V_o is 0 V. Furthermore, Q1 is saturated when $V_{ds1} > V_{gs1} - V_t$, or

$$V_t(N) < V_i < V_o + V_t(N) \quad (9.1)$$

and Q2 is saturated when $V_{sd2} > V_{sg2} - V_t$, or

$$V_{dd} - V_t(P) > V_i > V_o - V_t(P) \quad (9.2)$$

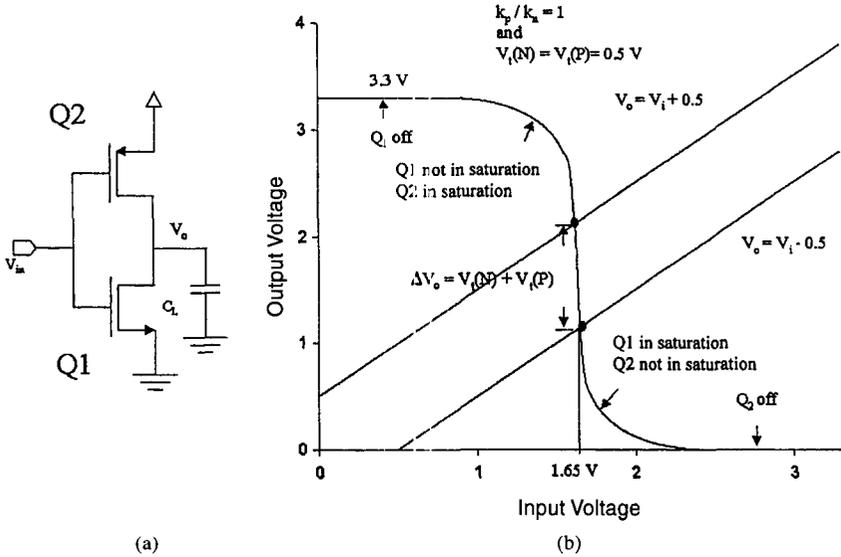


Figure 9.2 (a) A static CMOS inverter and (b) its voltage transfer characteristics (VTC).

For 0.35- μm CMOS technology, provided $V_t(N) = V_t(P) = 0.5\text{ V}$ and $V_{dd} = 3.3\text{ V}$, we would have Q1 saturated when $0.5 < V_i < V_0 + 0.5$ and Q2 saturated when $V_0 - 0.5 < V_i < 2.8$. In other words, Q1 would be saturated when

$$V_0 > V_i - 0.5 \tag{9.3}$$

And Q2 would be saturated when

$$V_0 < V_i + 0.5 \tag{9.4}$$

Since the two devices are in series, the drain currents for NMOS, $I_{ds}(N)$, and PMOS, $I_{sd}(P)$, are always equal. When Q1 is in saturation and Q2 is not (remained in the triode region), we have²

$$k_n[V_i - V_t(N)]^2 = k_p\{2[V_{dd} - V_i - V_t(P)](V_{dd} - V_0) - (V_{dd} - V_0)^2\} \tag{9.5}$$

where k_n and k_p are the gain factors for n- and p-channel devices, respectively.

Similarly, we find that when Q2 is in saturation and Q1 is not, we have

$$k_p[V_{dd} - V_i - V_t(P)]^2 = k_n\{2[V_i - V_t(N)]V_0 - V_0^2\} \tag{9.6}$$

When both transistors are in saturation, we find that

$$k_n[V_i - V_t(N)]^2 = k_p[V_{dd} - V_i - V_t(P)]^2 \tag{9.7}$$

On the basis of Eqs. (9.5)–(9.7), in Figure 9.2*b* we have plotted the VTC of a 0.35- μm -gate CMOS inverter for $V_i(N) = V_i(P) = 0.5\text{ V}$, $V_{dd} = 3.3\text{ V}$, and $k_n/k_p = 1$. Above the line $V_0 = V_i + 0.5$ (Eq. 9.5), Q2 is in saturation. Below the line $V_0 = V_i - 0.5$, Q1 is in saturation. In the region between the two lines, both transistors are in saturation.

Note also in Figure 9.2*b* that the simultaneous saturation of both NMOS and PMOS defines a unique switching threshold voltage V_M for the inverter, calculated from Eqs.9.5–9.7 to be

$$V_M = \frac{(k_n/k_p)^{1/2}[V_{dd} - V_i(P)] + V_i(N)}{1 + (k_n/k_p)^{1/2}} \quad (9.8)$$

With $V_i(N) = V_i(P) = 0.5\text{ V}$, $V_{dd} = 3.3\text{ V}$, and $k_n/k_p = 1$, this threshold voltage for inverter switching is midway between 0 V and V_{dd} because we have selected $k_n = k_p$ and $V_i(P) = V_i(N)$. If k_n is not equal to k_p or $V_i(P)$ is not equal $V_i(N)$, the inverter would not be symmetrical. In any case, we find from Eqs. 9.1 and 9.2 that the magnitude of the abrupt transition is given by

$$\Delta V_0 = V_i(P) + V_i(N) \quad (9.9)$$

The infinite slope displayed in Figure 9.2*b* results from our assumption that in the saturation region the device current is independent of drain–source voltages (zero output conductance). This, of course, is not precisely correct in real devices, and, hence, the gate switching in a physical situation would be sharp but not with an infinite slope.

9.2.2 Dynamic Characteristics of CMOS Inverter

Consider the situation represented in Figure 9.2. One of the major advantages of CMOS is that there is always a conduction path to charge and discharge a capacitive load across the gate output. Under high-speed operation, the propagation delay of the CMOS inverter is determined by the time it takes to charge or discharge the load capacitance C_L through the P- or NMOS device. This load capacitance may be the input capacitance of a (or multiple) succeeding gates and the interconnect capacitance. Both of these parasitics must be minimized to obtain high-speed CMOS operation.

The propagation delay of the CMOS inverter can be computed by integrating the capacitor (dis)charging times. Suppose that Q1, which is originally OFF, is turned ON by applying a gate voltage $V_{gs} > V_t$. Initially, the transistor operates in its saturation region at a maximum drain–source voltage $V_{ds} = V_{CM}$. When the load capacitor discharges, its voltage decreases and the transistor Q1 makes an excursion from the saturation to the triode region, when the voltage of the load capacitor $V_C(t)$ drops to the level $V_{gs} - V_t$. Let t_{sat} be the discharging time in the saturation region and t_{triode} the time in the triode region, then the total discharging time t_d can be

calculated as³

$$t_d = t_{\text{sat}} + t_{\text{triode}} = \frac{C_L V_{CM} - V_{gs} + V_t}{k_n (V_{gs} - V_t)^2} + \frac{1.15 C_L / k_n}{V_{gs} - V_t} \quad (9.10)$$

Generally speaking, $t_{\text{triode}} \gg t_{\text{sat}}$, and the average $V_{gs} = 0.5 V_{ss}$. We can further simplify the C_L discharge time:

$$t_d = \frac{C_L / k_n}{V_{gs} - V_t(N)} = \frac{C_L / k_n}{0.5 V_{dd} - V_t(N)} \quad (9.11)$$

Using the same principle, a similar relation can be derived for the C_L capacitor charging time t_c :

$$t_c = \frac{C_L / k_p}{V_{gs} - V_t(P)} = \frac{C_L / k_p}{0.5 V_{dd} - V_t(P)} \quad (9.12)$$

Since the total propagation delay of the CMOS inverter is defined as $\frac{1}{2}(t_d + t_c)$, we then have

$$t_p = \frac{C_L}{2} \left[\frac{1}{k_p(0.5 V_{dd} - V_t(P))} + \frac{1}{k_n(0.5 V_{dd} - V_t(N))} \right] \quad (9.13)$$

For the 1.0- μm CMOS technology, the loading capacitance C_L is approximately 40 fF (assuming fanout = 1). Suppose that $V_{dd} = 5 \text{ V}$, $V_t = 0.8 \text{ V}$, and $k_n = k_p = 120 \times 10^{-6} \text{ A/V}^2$; the propagation delay per-inverter-gate is estimated to be about 200 ps. For more advanced CMOS technologies, the gate delay times are shorter, approximately in proportion to their gate lengths: $t_p = 100 \text{ ps}$ for 0.5- μm CMOS and 50 ps for 0.25- μm CMOS.

Because it is the most widely used logic gate, the static CMOS inverter is a good choice for low-power and moderate-speed integrated circuits. Ideally, the static power consumption of the CMOS inverter is zero. In reality, however, there is always a leakage current present, flowing through the reverse-biased diode junctions between the source, drain, and substrate of the transistors. The subthreshold current of the transistors is another source of the leakage current. For both sources of leakage, the resulting static power dissipation is expressed by

$$P_{\text{static}} = I_{\text{leakage}} V_{dd} \quad (9.14)$$

For the device size of interest, the leakage current ranges typically from 0.1 nA to a few nanoamperes at room temperature. For a die with 10^6 transistors, operated at a supply voltage of 3 V, the leakage current results in a few milliwatts in power consumption, which is obviously not a serious concern. The majority of the power is still consumed during the switching. When the capacitor C_L gets charged during the

low-to-high transition through the PMOS, the voltage of the load capacitor rises from 0 to V_{dd} . While part of the energy is stored in the load capacitor, the remainder of the energy drawn from the power supply is dissipated into the PMOS device. During the high-to-low transition, the capacitor is discharged and the stored energy is dissipated through the NMOS transistor. If the gate is switched ON and OFF f times per second, the power consumption equals

$$P_{\text{dynamic}} = C_L V_{dd}^2 f \quad (9.15)$$

Advances in CMOS technology result in increasingly high values of f as the f_T of devices increase and the gate propagation time decreases. At the same time, the total capacitance on the chip (C_L) continues to increase as more gates are integrated on the same chip. The dynamic power dissipation has become very significant in today's high-speed (>100-MHz) ULSI chips.

As we mentioned earlier, the assumption of the zero rise and fall times of the CMOS inverter gate is incorrect. During the gate switching (low-to-high or high-to-low) there is a period when both P- and NMOS transistors conduct and draws dc current (Fig. 9.2). It contributes extra power dissipation besides $C_L V_{dd}^2 f$ and generates a fair amount of switching-noise (Fig. 9.3). Under the assumption that the conducting current spikes can be approximated as triangles, we can estimate the power consumed per switching period,

$$P_{\text{direct-path}} = \frac{1}{2} (T_{\text{direct-path}} V_{dd} I_{\text{peak}} f) \quad (9.16)$$

where $T_{\text{direct-path}}$ represents the total time when both PMOS and NMOS conduct. By adding Eqs. 9.14–9.16, we have the total power consumption of the

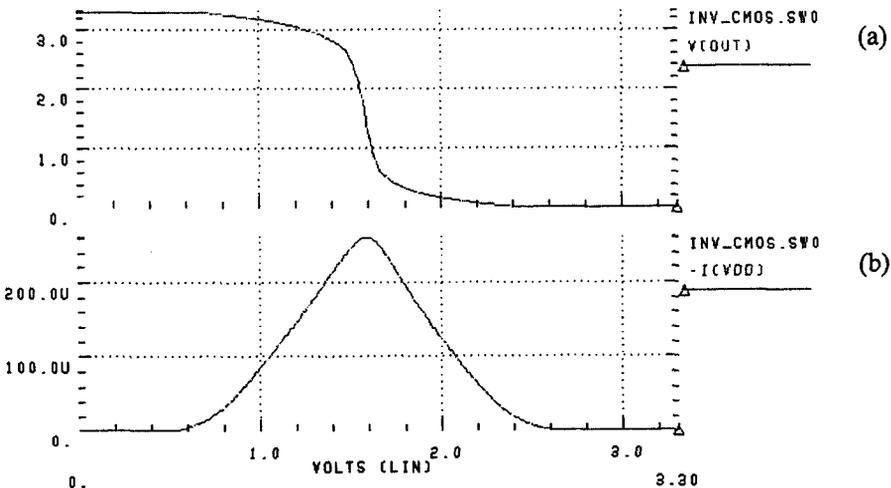


Figure 9.3 SPICE simulated static 0.5- μm CMOS inverter (a) output voltage and (b) power supply current versus input voltage.

CMOS inverter

$$P_{\text{total}} = C_L V_{dd}^2 f + I_{\text{leakage}} V_{dd} + \frac{1}{2} T_{\text{direct-path}} V_{dd} I_{\text{peak}} f \quad (9.17)$$

The switching noise generated by direct-path leakage can be easily coupled through power supply buses or semiconductor substrate. It often causes serious crosstalk between digital and analog circuits, especially when both circuits are laid out in close proximity. Note that the time period for simultaneous P- and NMOS conduction takes up a more substantial portion in a clock cycle with increased clock speed, making this effect more serious in high-speed circuits. However, by carefully arranging the threshold voltages of both P- and NMOS and their power supply voltage, one can reduce the input voltage range to minimize the simultaneous current conduction. Figure 9.3 shows dc transfer curves of output voltage and power supply current flow versus input voltage. A peak leakage current of $260 \mu\text{A}$ is observed during the ON-OFF transient of the invert gate. Figure 9.4 shows the transient behavior of the same CMOS inverter at 1 GHz. The leakage current now peaks at a different value as a result of the transient behavior of the inverter. Generally speaking, the direct-path power consumption is lower with faster rise and fall times.

9.2.3 Differential Logic

In the previous section, we discussed the static and dynamic characteristics of the most widely used CMOS inverter gate. The CMOS inverter has major advantages in its simple circuit architecture and low static power consumption. However, its speed is somewhat limited by the relatively large signal swing, and its switching noise is typically high because of its unbalanced single-ended nature. CMOS circuit

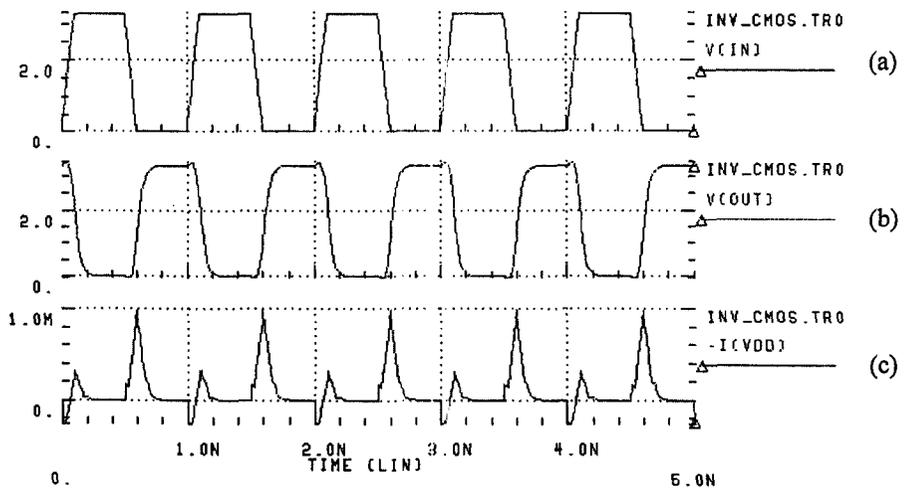


Figure 9.4 Simulated (a) input voltage, (b) output voltage, and (c) power dissipation waveform of a static CMOS inverter operating at 1 GHz.

designers must pay special attention to minimize the switching noise at the maximum operating speed for mixed-signal design. This problem becomes more severe because of the increased amounts of on-chip digital circuitry and sensitive analog circuits, especially in the design of oversampled A/D converter and high-speed frequency synthesizers. In this section, we discuss differential logical gates that are designed to provide lower switching-noise for high-speed mixed-signal circuit applications.

Current Mode Logic (CML)

In a conventional CMOS inverter, an overlapping current pulse flows from V_{dd} to ground during the logic transition (see Fig. 9.3). When many similar gates switch states, the resulting large current pulse flow through the resistors and inductors associated with the supply lines, bonding pads, and substrate. This generates a large switching noise voltage on the supply lines that could reach several hundred millivolts.^{4,5} The accuracy of the mixed-signal system is significantly degraded with the propagation of digital switching noise into the sensitive analog circuitry through supply lines and the common substrate. Several MOSFET logic gates have been designed to counter the switching noise problem and to operate at very high speed. The most widely used is current mode logic (CML).

Figure 9.5 illustrates the CML inverter for implementation in a p-well NMOS technology. This differential topology has been used as a basic block for analog amplifiers.⁶ As a primary topology used in differential logic, the CML inverter, typically, uses diode-connected NMOS as its active loads (Fig. 9.5).⁷ The NMOS differential-pair input stage is biased with constant current I and the loads are biased at constant voltage V_d . The logic gate operation is based on the current steering technique, similar to that of current mode bipolar logic, by applying differential input voltages. For example, with an input voltage applied such that Q1 is ON and Q2 is OFF, the current I flows through Q1 and the current through Q2 is almost zero. The load devices Q3 and Q4 with $k_{n3} = k_{n4}$ are biased by V_d in the saturation region to

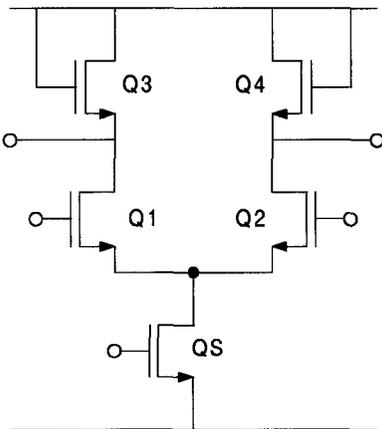


Figure 9.5 Current mode logic (CML) inverter in a p-well NMOS technology.

provide the differential output voltages at a desired level as follows:

$$V_{LO} = V_d - V_{\text{ciode}} - \left(\frac{2I}{k_{n3}}\right)^{1/2} \tag{9.18}$$

and

$$V_{HI} = V_d - V_{\text{diode}} \tag{9.19}$$

By properly sizing the output devices, the output logic swing is

$$\Delta V_{\text{out}} = \left(\frac{2I}{k_{n3}}\right)^{1/2} \tag{9.20}$$

where ΔV_{out} is designed to range, typically, from 0.5 to 1 V. The bias V_d is used to adjust the trigger voltage to a desired value of $V_{dd}/2$. This design reduces not only chip size but also the voltage swing at its outputs. As shown in Figure 9.6, the reduced output voltage swing can be translated into faster switching time due to shorter charging and discharging times and decreased dynamic power consumption ($P = C_L \Delta V_{\text{out}}^2 f$). The switching noise is also reduced because of the smaller output logic levels and the current mode operation of the circuit. The switching noise of the CML is generated only by the displacement current during the switching.

Compared to single-ended logic families, the switching noise of CML is minimal because of its differential circuit topology. Figure 9.7 compares the propagation delays and power consumption of the standard CMOS versus the CML gates.⁷ It indicates that CML is much faster, especially at low V_{dd} bias conditions.

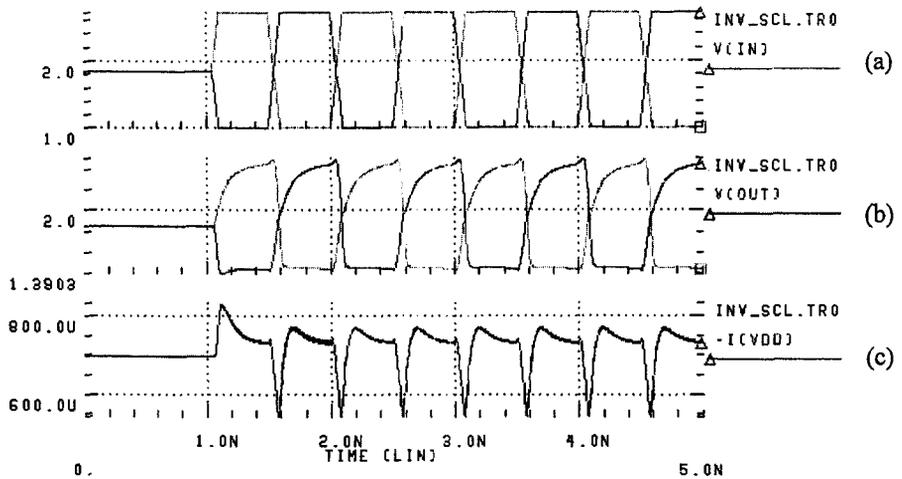


Figure 9.6 Simulated (a) input voltage, (b) output voltage, and (c) power dissipation waveform of a source-coupled logic inverter at 1 GHz.

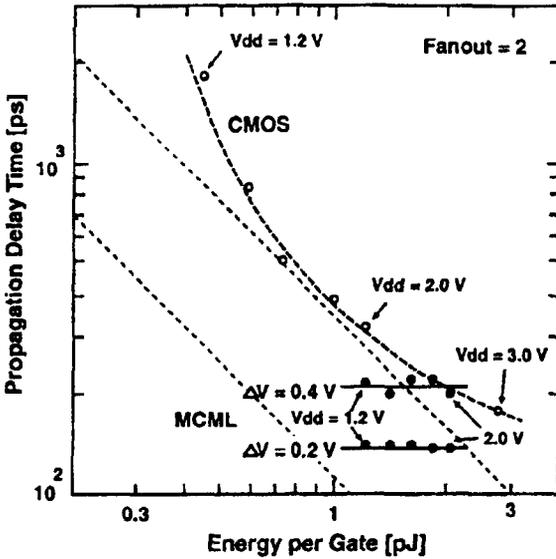


Figure 9.7 Propagation delay time versus energy per gate⁷ for standard CMOS and CML gates.

The fully differential CML topology generates complementary outputs. This leads to a reduction in the number of devices required to implement complex logic gates.^{8,9} Figure 9.8 shows a complex logic gate implementation of CML for the NAND/AND logic functions. The CML logic family may be used when maximum speed and very low noise circuit operation are needed and the circuit size is not a major concern.

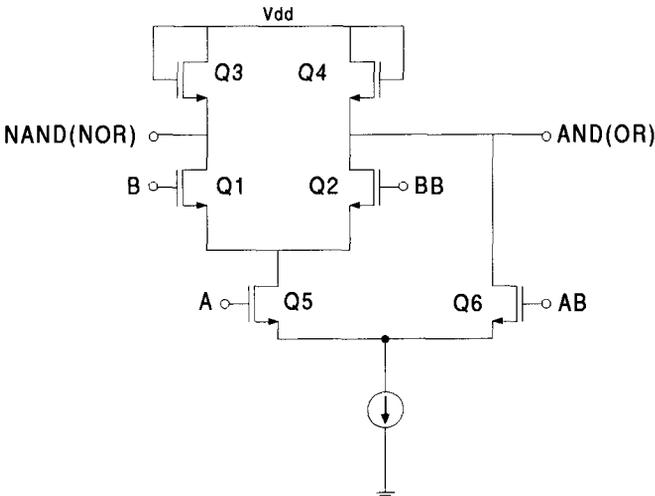


Figure 9.8 A CML NAND/AND gate.

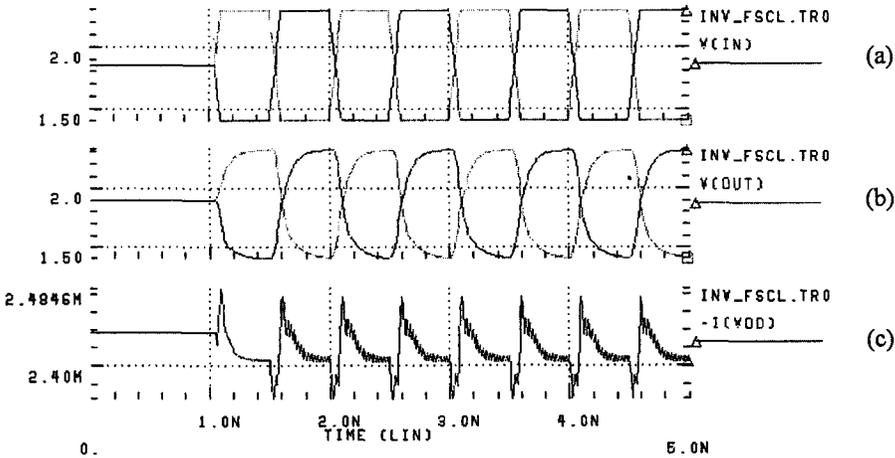


Figure 9.10 (a) Simulated input voltage, (b) output voltage, and (c) power dissipation waveform of a FSCL inverter at 1 GHz.

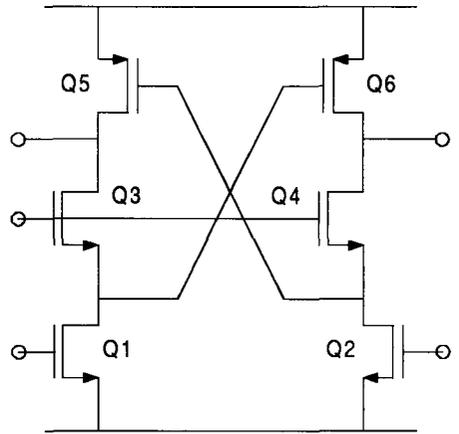
which is usually designed to range from 0.5 to 1 V. To have an adequate noise margin, the voltage gain of the FSCL stage is generally greater than 2.

The major benefit of FSCL (over CML and SCL) is that it is very fast and has extremely low switching noise, which makes it very desirable in a noise-sensitive circuit such as the phase-locked loop (PLL) in a wireless receiver circuit. Figure 9.10 shows the simulated output waveform and power dissipation of a FSCL 0.5- μm CMOS inverter operating at 1 GHz. The switching noise is about 100 μA , which is significantly lower than CML and static CMOS.

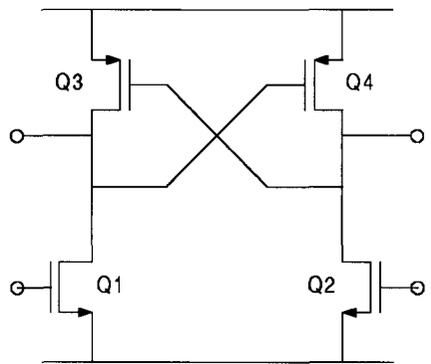
Differential Split-Level Logic (DSL) and Cascode Voltage Switch Logic (CVSL)

Differential split-level logic (DSL) and cascode voltage switch logic (CVSL) are very attractive alternatives to traditional static CMOS logic. Both logic circuits offer fully differential inputs and outputs while simplifying the circuit complexity by not having current sources as in SCL and FSCL. The differential input/output are provided through a pair of cross-coupled inverters. Because of the complementary outputs, circuit topology can be simplified especially for complex logic functions.

Detailed explanations and analysis of DSL CMOS logic have been published.¹¹ The operating principle can be explained briefly with the aid of Figure 9.11a. The transistors Q3 and Q4 represent complementary NMOS transistor switch structures, which implement a logic function and its inverse. A cross-coupled PMOS and NMOS load, comprising Q1, Q2, Q5, and Q6, converts the drain outputs of Q3 and Q4 (V_F and V_{FN}) into CMOS logic levels. The switching behavior of this logic circuit is shown in Figure 9.12. The propagation delay t_d of the DSL circuit can be defined as the time difference between when V_i (input voltage) and V_o (output



(a)



(b)

Figure 9.11 (a) DSL CMOS logic circuit; (b) CVSL logic circuit.

voltage) go through $0.5 V_{dd}$. Mathematically, it can be represented as follows:

$$t_d = \frac{0.5C_1V_{dd}}{i_{C1}} \tag{9.24}$$

where C_1 is the capacitance at output and i_{C1} is the current, which charges it during the propagation delay.

As shown in Figure 9.11b, CVSL is a variation of DSL but with a simpler architecture.¹² Without buffering transistors Q5 and Q6 of Figure 9.11a, CVSL has a larger signal swing at the expense of lower operating speed. The switching behavior of this logic circuit is shown in Figure 9.13. To benchmark the high-speed performance of different logic circuits, we have compared SCMOS, FSCL, CML, DSL, CVSL, and NMOS-only ring oscillators in terms of maximum oscillation

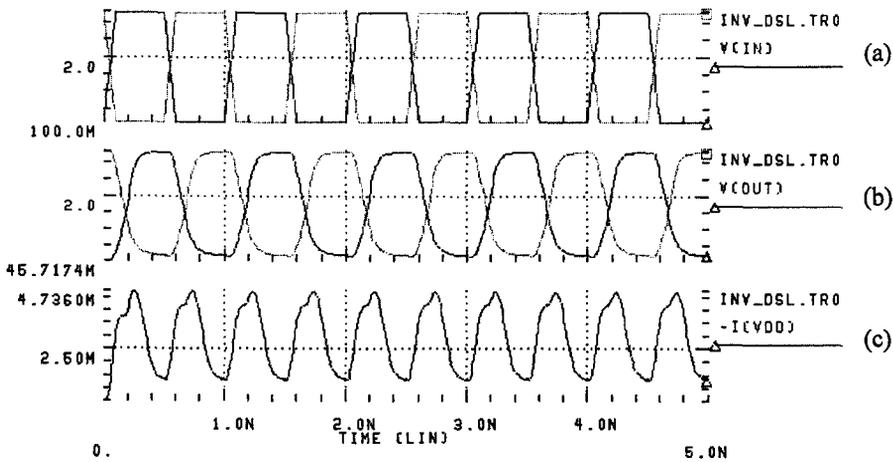


Figure 9.12 (a) Simulated input voltage, (b) output voltage, and (c) power dissipation waveform of a DSL inverter operating at 1 GHz.

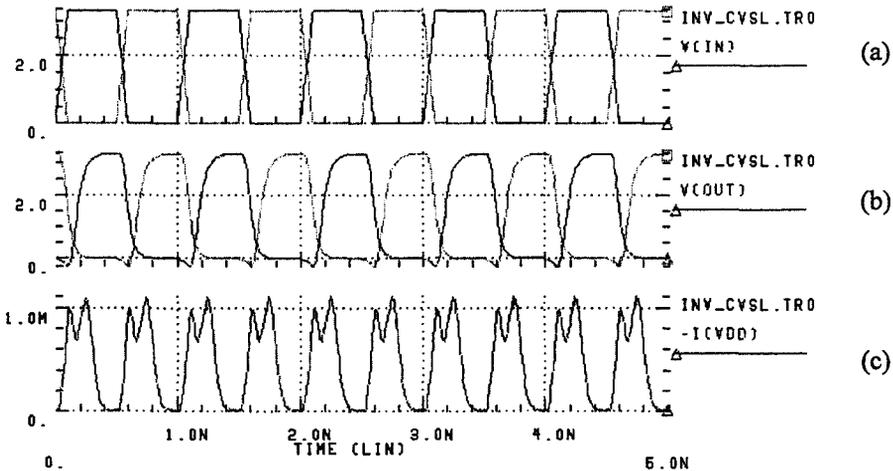


Figure 9.13 Simulated input voltage (a), output voltage (b), and power dissipation (c) waveform of a CVSL inverter operating at 1 GHz.

frequency (f_{osc}), power consumption (P_d), switching noise (i_{pulse}), and peak-to-peak voltage swings (V_{p-p}). The results are summarized in Table 9.1, where **boldface** indicates better performance and *underline* indicates worse performance in a category. The ring oscillator simulation is based on a standard 0.5- μm CMOS technology, operating at $V_{dd} = 3.3\text{ V}$. As noted in Table 9.1, the SCMOS offers very high speed and the lowest power consumption among all, but has significant switching noise. The performance of NMOS is similar to that of SCMOS with slightly higher power consumption and relatively lower switching noise. The

TABLE 9.1 Benchmark of CMOS Logic Circuits (0.5- μ m CMOS; $V_{dd} = 3.3$ V)

	SCMOS	FSCL	CML	DSL	CVSL	NMOS
f_{OSC} (MHz)	411	477	587	290	<u>152</u>	386
P_d (mW)	2.5	<u>43.8</u>	23.2	17.2	2.8	9.7
i_{pulse} (μ A)	276	19.8	75.1	<u>715</u>	130	115
V_{p-p} (V)	3.36	0.92	1.39	3.06	3.42	2.11
t_{PLH} (ps)	108	73.2	91	117	<u>466</u>	180
t_{PHL} (ps)	109	114.7	64	196	128	51

differential-logic circuits such as FSCL is excellent for switching noise but generally worse in power consumption (except CVSL). The CML technology is widely used in modern mixed-signal circuits because of its outstanding speed and moderate power consumption.

9.3 CMOS TECHNOLOGY FOR ANALOG AND RF APPLICATIONS

The advance of CMOS technology has resulted in “standard” devices in conventional digital technologies that are highly suitable for high-frequency *analog* and RF applications. For example, standard devices have made the radio-on-a-chip possible by integrating RF front-end circuits (in the 1–6-GHz range) with the baseband digital circuits for wireless transceivers. Significant industrial and academic effort has recently focused on the applications of CMOS for the RF/analog front end up to the antenna and the integration of these analog components with the very noisy back-end digital systems. In this section, we concentrate on the analog building blocks required for highly integrated CMOS-based ULSI wireless transceivers, their design, and state-of-the-art performance.

Although the diverse wireless standards and modulation schemes require very different circuit architectures for the RF/analog front end, there are a consistent set of analog building blocks that are required in any given wireless transceiver implementation. These are illustrated in the example architecture of Figure 9.14. There are many different applications with different tradeoffs and design specifications, yet they all share the same analog building blocks. The transceiver contains the low-noise amplifiers (LNA) and downconversion mixers in the receiver chain and contains upconversion mixers and power amplifiers (PA) in the transmitter chain. The system also contains frequency-controlling components (often shared between the receiver and transmitter chains) of voltage-controlled oscillators (VCO), phase-locked loops (PLLs), and frequency synthesizers. The generic schematic of a simplified wireless transceiver in Figure 9.14 highlights the components of interest here along with the off-chip components such as antennas, filters, and transceiver switches (T/R). The design of these building blocks and the present state-of-the-art efforts in advancing their performance in CMOS technology for highly integrated transceiver systems is described in the next section.

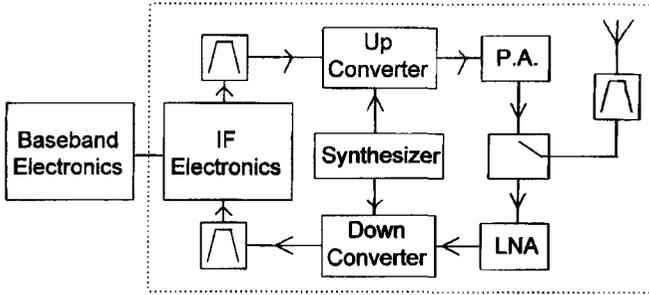


Figure 9.14 A generic transceiver used in modern wireless communications.

9.4 CMOS LOW-NOISE AMPLIFIERS

The low-noise amplifier (LNA) requires significant gain of 10–20 dB over the narrow frequency band of interest coming in from the antenna, a low additive noise, and the capability to maintain linearity under large-signal conditions where distortion can degrade the performance. The requirements for wide dynamic range CMOS LNAs and their design issues are discussed as follows.

9.4.1 Low-Noise Amplifier (LNA) Fundamentals

The function of LNAs in a transceiver is to amplify the incoming small-amplitude signal from the antenna (on the order of microvolts) to a larger value of roughly 10–100 times or equivalently 10–20 dB. This is so that the following stages demodulating that signal may operate on an incoming signal, whose value is significantly above the noise level inherently surrounding the signal. Therefore, the LNA must amplify the incoming signal while avoiding the addition of its own noise to the amplified output. For such an application, a specific figure of merit denoted as the “noise figure” is defined for such amplifiers and describes the input-referred additive noise of the LNA. The noise factor, F , is related to the noise figure, NF , by

$$NF = 10 \log_{10} F \quad (9.25)$$

where F is defined as the amount that the amplifier degrades the signal-to-noise ratio (SNR) from input to output. For example, if the signal to noise ratio is 30 dB at the input, and the additive noise of the device under test (DUT) degrades the signal-to-noise ratio at the output to 27 dB, the noise figure is 3 dB. This is illustrated in Figure 9.15 and may be expressed in a number of ways, as the ratio of the input signal-to-noise ratio (SNR_i) to the output signal-to-noise ratio (SNR_o).

$$F = \frac{\left(\frac{S_i}{n_i}\right)}{\left(\frac{S_o}{n_o}\right)} = \frac{n_o}{G_T n_i} \quad (9.26)$$

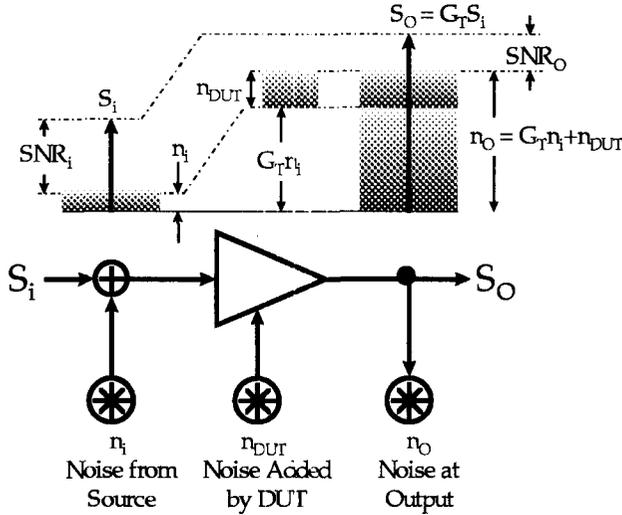


Figure 9.15 Schematic definition of the noise figure of a LNA.

Equivalently, this may be described as the ratio of the total output noise power of the amplifier to the output noise power of an equivalent but “noise-free” amplifier that simply amplifies the noise power from the source but adds no noise of its own.

A two-port representation of the noise properties of any amplifier is shown in Figure 9.16a. The two equivalent noise sources are referred to the input of the device and represent the noise voltage and current source magnitudes that when configured with a noise-free version of the device as shown in Figure 9.16b, will generate the output noise power of the actual device. This equivalent representation is useful in that it relates the equivalent input noise levels by which the device is inherently self-limited to the incoming-signal levels to be amplified. For the receiver, this defines a critical “noise floor” that incoming signals must exceed to be detected.

One important point is that *all* resistive losses generate noise, including substrate losses, radiation loss, and the intrinsic resistive loss of reactive components such as inductors and capacitors. An ideal pure reactance without the resistance does not

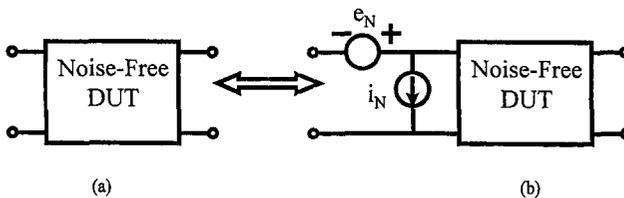


Figure 9.16 (a) A two-port representation for the noise property of an amplifier; (b) an equivalent-noise property represented by a noise-free amplifier with equivalent noise sources referred to the input of the device.

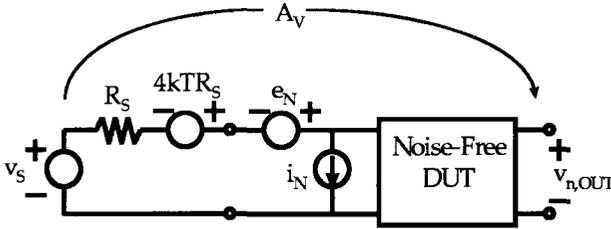


Figure 9.17 Equivalent circuit to calculate noise figure directly from voltage gain and cumulative noise sources.

generate thermal noise. A second fundamental source of noise is the shot-noise component that occurs when discrete electronic charges are injected over a potential barrier and drift to a collection point. Each arrives at random discrete intervals, causing a series of uncorrelated current fluctuations. For CMOS devices, this component from channel to drain may be neglected for low-noise amplification.

To calculate the equivalent noise sources e_n and i_n from a given equivalent circuit containing all the thermal and shot-noise components, a simple linear superposition is used to calculate the contribution of each source separately. To calculate e_n , the input of the amplifier or device is shorted (short-circuited) and then the voltage spectral density is calculated at the output as a result of each source. It is then referred to the input by dividing by the total voltage gain to represent e_n . To calculate i_n , the input is open circuited and the current spectral density at the output is calculated. It is then referred to the input by dividing by the total current gain.

Equivalently, the noise behavior of any two-port device may be represented¹⁴ as in Figure 9.17 by the spot noise factor defined to be the ratio of the available noise power per unit bandwidth (Δf) at the output to that portion that is due to the input termination, R_S . This may be calculated as shown in Figure 9.17 by dividing the open-circuit mean-square noise voltage measured at the output port ($v_{n,OUT}^2$) divided by the voltage gain squared (A_V^2) times e_S^2 , where $e_S^2 = 4kT\Delta fR_S$ resulting in $F = v_{n,OUT}^2 / (A_V^2 \cdot 4kTR_S)$. This is equivalent to the more conventional expression of the noise factor in terms of the input noise current and voltage representation as shown in Figure 16b and is calculated as

$$F = 1 + \left(\frac{e_N^2}{4kT\Delta fR_S} \right) + \left(\frac{i_N^2R_S}{4kT\Delta f} \right) \quad (9.27)$$

In general, the two equivalent noise sources e_N and i_N will each be the result of several internal noise sources within the two port, and because these internal noise sources contribute to both, they are said to be at least partially correlated.

We see from this expression that the noise factor is a strong function of the source resistance, and that the minimum noise factor, F_{MIN} , occurs at an optimal value of R_S

expressed as

$$R_{\text{SCPT}} = \frac{e_n}{i_n} \quad (9.28)$$

$$F_{\text{MIN}} := \frac{1 + e_n i_n}{2kT} \quad (9.29)$$

The previous discussion focused on the noise floor and minimum signal level that may be reasonably amplified by the LNA. A second design consideration of the low-noise amplifier is its ability to amplify larger signals without significant distortion. The larger signal level input to the low-noise amplifier eventually causes the amplifier to be nonlinear and generates harmonics at the output. Although harmonics are integer multiples of the fundamental frequency, they are fairly easily filtered out in the receiver architecture following the LNA. However, signals that are both large and in band at f_1 and $f_2 = f_1 + \Delta f$ will mix together and generate intermodulation distortion products at $2f_1 - f_2 = f_1 - \Delta f$ and $2f_2 - f_1 = f_2 + \Delta f$ as shown in Figure 9.18.

These third-order intermodulation products are critical because they are very close to the original signal and are, therefore, very difficult to filter out, and may significantly interfere with desired signals in band to be detected and demodulated. The output power at the fundamental frequency increases with a 1 : 1 linear slope versus the input power. The third-order intermodulation, however, increases with a much higher slope of 3 : 1 versus the input power, because it is the result of a multiplication of three signals (f_1, f_1 , and, f_2). Therefore, it generates a logarithmic, cubic dependence on input power. The rapid increase in third-order intermodulation product (IMD_3) with input power causes a critical degradation in the in-band distortion characteristics of the amplifier, and defines the upper amplitude limit of signal levels into the LNA.

This specification, shown in Figure 9.19, is the spurious free dynamic range (SFDR) and defines the input power range over which these third-order intermodulation products are below the minimum detectable signal level. Another large-signal limitation involves gain compression as the input signal becomes so large as to “saturate” or “desensitize” the amplifier and cause its gain to roll off

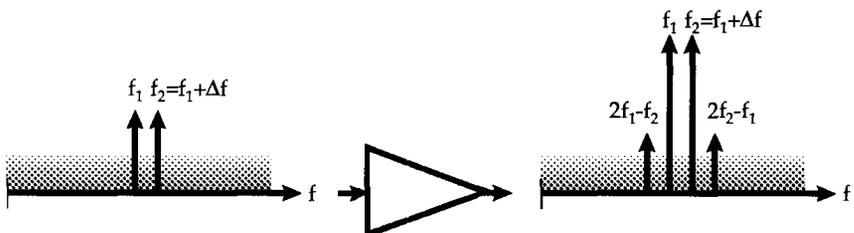


Figure 9.18 In-band intermodulation products caused by amplifier's nonlinearity.

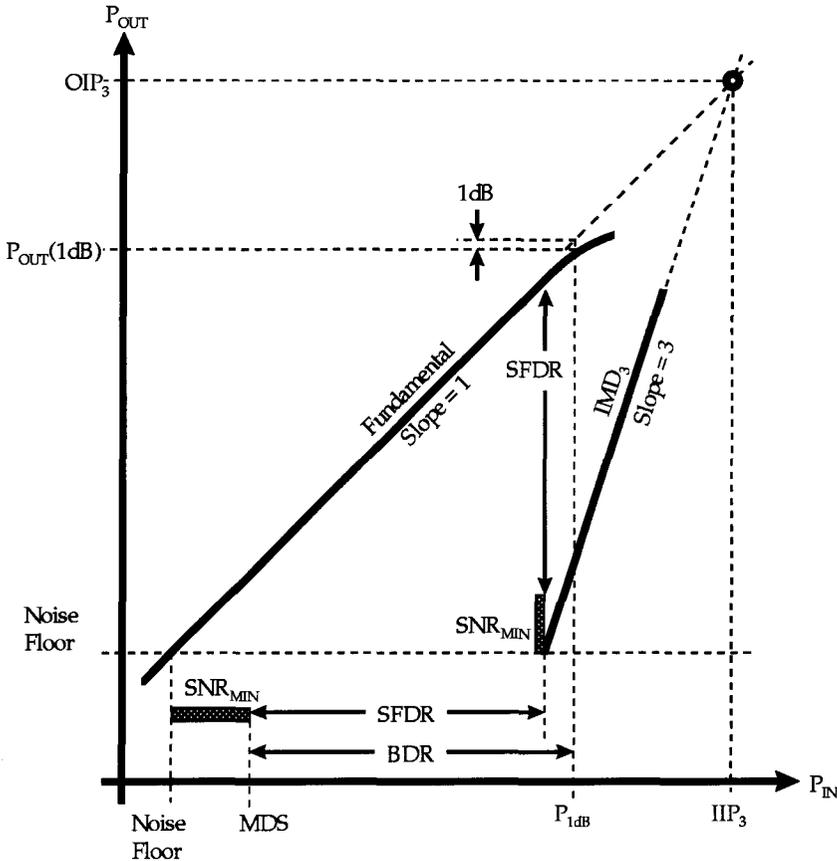


Figure 9.19 Output power versus input power for an amplifier showing the minimum and maximum signals that may be amplified linearly based on the noise floor and negligible third order intermodulation products, respectively.

from the small-signal value. This may be caused by in-band interference or *blocking signals* that cannot be filtered out before reaching the LNA input. This point is usually defined at the input power for which the gain of the amplifier has rolled off 1 dB, and sets the *blocking dynamic range* (BDR).

Figure 9.19 also shows the useful figure of merit called the *intermodulation intercept point* (IP3) that is calculated as the point at which the extrapolated third-order intermodulation product would become equal to the output power. This figure of merit describes the output power of the fundamental relative to the third-order intermodulation over the entire curve up to where the fundamental starts to compress (at the P_{1dB} point). The input-referred intercept point (IIP_3) is the input power at which the extrapolated IMD_3 is equal to the output power, and the output-referred intercept point (OIP_3) is equal to that value of output power where the two curves meet. The usefulness of this figure of merit is that the ratio of output power to third

order intermodulation product at any input power (over which the 3:1 slope is maintained) may be expressed according to

$$(P_{\text{OUT}} - \text{IMD}_3) = 2 \cdot (\text{IIP}_3 - P_{\text{IN}}) \quad (9.30)$$

The region of usable input-level signal amplitudes between the two extremes of minimum signal levels limited by noise and the maximum signal levels limited by spurious in-band intermodulation products, gain desensitization and compression, is called the *dynamic range* of the amplifier. The noise floor, shown in Figure 9.19, of the receiver is defined as

$$\text{NoiseFloor} = -174 \text{ dBm} + \text{NF} + 10 \log B \quad (9.31)$$

where -174 dBm is the minimum noise level for any system according to the kT thermal noise limit at room temperature. NoiseFloor is the noise figure of the entire receiver in decibels and sets the signal level at the input that the incoming signal must exceed in order to generate an output signal at least equal to the noise level after amplification, and B is the bandwidth of the band of interest in hertz over which the noise is integrated. This noise floor defines the minimum levels that must be seen at the input so that the output signal is merely equal to the noise level. In practical receivers, however, additional margin is required so that the output signal exceeds the noise and may be reasonably detected. This is termed the *sensitivity* or *minimum detectable signal level* (MDS) of the receiver and is defined as the minimum signal level that the system can detect with acceptable signal to noise ratio SNR_{MIN} :

$$\text{MDS} = \text{NoiseFloor} + \text{SNR}_{\text{MIN}} = -174 \text{ dBm} + \text{NF} + \text{SNR}_{\text{MIN}} + 10 \log B \quad (9.32)$$

The spurious free dynamic range is defined as the ratio of the maximum allowable input that maintains IMD levels below the noise floor to this sensitivity (MDS) and is shown in Figure 9.19. This may be expressed mathematically in terms of the IIP_3 by

$$\text{SFDR} = \frac{2}{3} (\text{OIP}_3 - \text{NoiseFloor}) - \text{SNR}_{\text{MIN}} \quad (9.33)$$

The blocking dynamic range up to the 1dB compression point or P_1 dB is defined as:

$$\text{BDR} = P_{1\text{dB}} - \text{NoiseFloor} - \text{SNR}_{\text{MIN}} \quad (9.34)$$

The critical limitations over which the LNA will function adequately are a strong function of the practical design, discussed in detail below specifically for state-of-the-art CMOS implementations.

A critical issue for highly integrated transceiver design is the noise figure of more than one stage in cascade, which is often required to attain the gain specification. Friis' law may express this as

$$\text{NF}_{\text{TOT}} = 1 + (\text{NF}_1 - 1) + \frac{(\text{NF}_2 - 1)}{G_1} + \frac{(\text{NF}_3 - 1)}{G_1 G_2} + \dots + \frac{(\text{NF}_n - 1)}{\prod_{i=1}^{n-1} G_i} \quad (9.35)$$

where NF_n is the noise figure of the n th stage and G_n is the power gain of the n th stage under the conditions of matching provided by the preceding and following stages. This is an important point, as the NF_n is a critical function of the output impedance of the preceding stage, $R_{OUT(n-1)}$. The overall noise figure of a cascade amplifier such as this for highly integrated designs is fundamentally limited by the noise figure of the first stage, as long as the gain of the first stage is large enough to minimize the noise figure contributions of subsequent stages. Those subsequent stages may then be optimized for gain as needed, and not depend too much on an optimal noise match to achieve the noise figure.

The detailed discussion above concerning the optimum source impedance for minimum noise figure is critically traded off with the requirement for good input impedance match for conjugate power transfer as well as for proper termination of the filter preceding the LNA and its low ripple passband characteristics. It is rarely the case that optimum noise matching corresponds to the conjugate match condition for power transfer, and it is certainly possible to have a great noise figure with a very large input power mismatch and small amplifier gain.

To match the input of the device for a given power level and device size, whether for optimum noise figure or power gain, there are matching techniques specifically for CMOS LNAs. First, to match impedance of any given level to an arbitrary impedance of a second value, important techniques of reactance transformation need to be discussed. For any given series impedance, a narrowband equivalent, parallel impedance may be synthesized and vice versa as shown in Figure 9.20.

The relationship between these two equivalent topologies for a narrowband centered around a given frequency is

$$R_S + jX_S = \left[\frac{1}{R_P} + \frac{1}{jX_P} \right]^{-1} \quad (9.36)$$

The objective is to match a complex impedance (device input–output) to a standard purely real resistance value ($50\ \Omega$ for most RF systems). This involves the transformation of the real part of a given impedance up or down to $50\ \Omega$ and the cancellation of all other reactances in narrowband. To convert a given complex impedance to a purely real resistance, the equivalence of series and parallel linear networks is used to represent the input impedance

$$Z_{IN} = R_S + jX_S = \left[\frac{1}{R_P} + \frac{1}{jX_P} \right]^{-1} \quad (9.37)$$

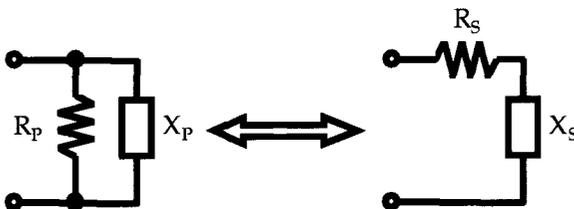


Figure 9.20 Equivalence of series and parallel linear networks.

The real parts on both sides of the equation must be equal, as must be the imaginary parts, which sets up relationships between the equivalent topologies. These relationships allow the conversion of a series representation to a parallel one, for example, according to

$$R_P = \frac{(R_S^2 + X_S^2)}{R_S} \tag{9.38}$$

$$X_P = \frac{(R_S^2 + X_S^2)}{X_S} \tag{9.39}$$

These equations may be used to transform a given network's real-part resistive component to a larger target value R_P . By adding additional series reactance (either a positive inductive or negative capacitive reactance) in the form of X_{SM} to the existing series topology and then adding a parallel reactance, X_{PM} , to cancel out the imaginary component and zero out X_P as shown in Figure 9.21a, we may effectively transform the real resistance R_S up to the desired target R'_P (say, $50\ \Omega$) according to

$$R'_P = \frac{R_S^2 + (X_S + X_{SM})^2}{R_S} \rightarrow X_{SM} = \sqrt{R'_P R_S - R_S^2} - X_S \tag{9.40}$$

and zero out the reactive impedance by placing an additional parallel reactance X_{PM} to net a zero reactive impedance for Z_{IN}' :

$$X_{PM} = \frac{R_S^2 + (X_S + X_{SM})^2}{X_S + X_{SM}} \tag{9.41}$$

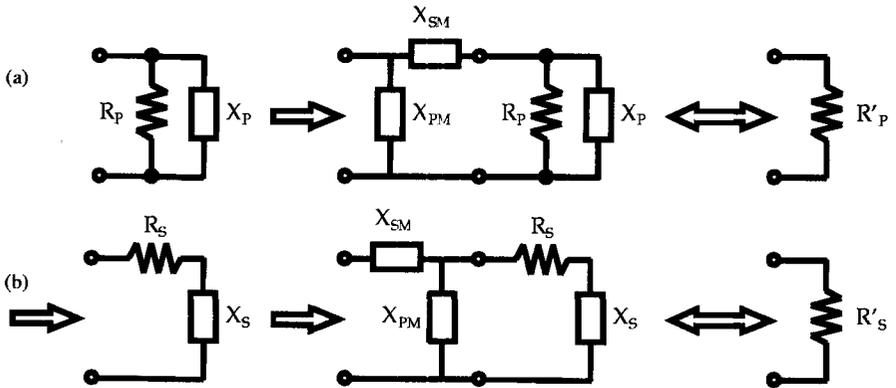


Figure 9.21 Impedance transformation to match a target resistance R'_P or R'_S by transforming (a) up or (b) down to that target.

Similarly, the impedance transformation down from a given real-part resistance to a lower target-resistance value requires the conversion from a parallel representation to a series one according to Figure 9.21*b* and

$$R_S = \frac{R_P X_P^2}{R_P^2 + X_P^2} \quad (9.42)$$

$$X_S = \frac{R_P^2 X_P}{R_P^2 + X_P^2} \quad (9.43)$$

By adding a parallel reactance X_{PM} , the net real-part resistance may be transformed down to the target value R'_S according to

$$R'_S = \frac{R_P (X_P + X_{PM})^2}{R_P^2 + (X_P + X_{PM})^2} \rightarrow X_{PM} = \sqrt{\frac{R'_S R_P^2}{R_P - R'_S}} - X_P \quad (9.44)$$

and by the addition of a series reactance to zero out the net reactive impedance according to

$$X_{SM} = -\frac{R_P^2 \cdot (X_P + X_{PM})^2}{R_P^2 + (X_P + X_{PM})^2} \quad (9.45)$$

resulting in the matched topology of Figure 9.21. This general approach to matching can be applied in all areas of transceiver circuit design, but specifically for CMOS implementation, several approaches have been developed to match a given transistor input. These are summarized in Figure 9.22*a, b* for both common gate and common source configurations.¹⁴ The common gate topology uses the inductance L to resonate with transistor input capacitances to provide an impedance match to 50Ω , according to

$$Z_{IN} = \frac{R_{LS} + j\omega L_S}{1 + g_m R_{LS} - \omega^2 (C_{GS} + C_{SB} + C_{PAD}) L_S + j\omega [g_m L_S + (C_{GS} + C_{SB} + C_{PAD}) R_{LS}]} \quad (9.46)$$

At the frequency of resonance and assuming R_{LS} is very small, Eq. 9.46 reduces to

$$Z_{IN} \approx \frac{1}{g_m} \quad (9.47)$$

For a complete treatment of the voltage gain, we find

$$\frac{V_2}{V_1} = \frac{-g_m R_L (R_{LS} + j\omega L_S)}{R_{LS} + R_S (1 + g_m R_{LS}) - \omega^2 R_S (C_{GS} + C_{SB} + C_{PAD}) L_S + j\omega [L_S (1 + g_m R_S) + (C_{GS} + C_{SB} + C_{PAD}) R_{LS}]} \quad (9.48)$$

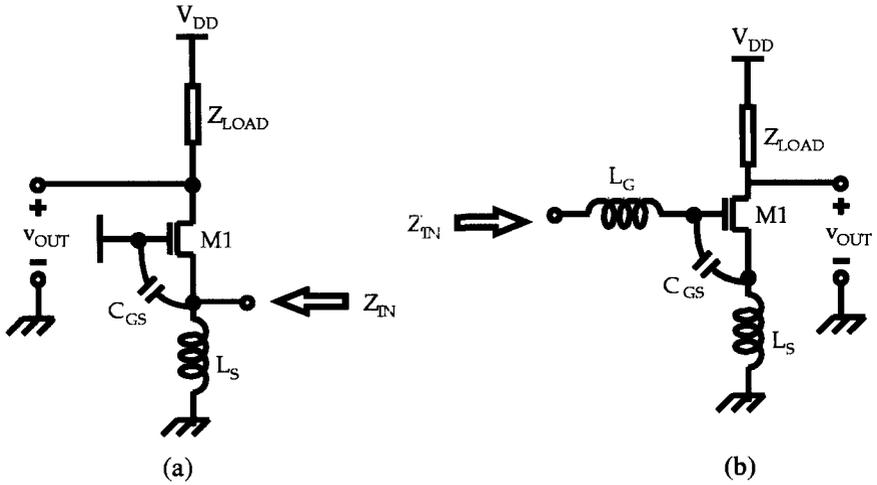


Figure 9.22 Matching techniques for low-noise amplifiers with (a) common gate and (b) common source configurations.

where R_S and R_L are the source and load resistances, respectively. For the case where only the channel thermal noise is considered, and we therefore represent the single noise source to be

$$I_n^2 = 4kT\gamma g_m \tag{9.49}$$

we may use Eq. 9.27 to calculate the approximate expression for the noise figure (given the match for input impedance described above) to be

$$F = 1 + \frac{4kT\gamma}{g_m} \cdot \frac{1}{4kTR_S} \approx 1 + \gamma \tag{9.50}$$

The noise figure is therefore fundamentally limited to 2.2 dB for long-channel devices in saturation with $\gamma \approx \frac{2}{3}$, but typically is higher for short-channel FETs with γ anywhere up to 5 as a result of short-channel effects. This topology may improve its noise match at the expense of input impedance matching and power consumption by further increasing g_m .

The common source FET in Figure 9.22b has a capacitive input impedance with some small parasitic resistance from gate and source. To match it to a much larger real part of 50Ω using the general series-to-parallel impedance-matching technique described above may require extremely large inductance values in series with the gate. Consequently, these devices may be difficult to integrate on chip with any reasonable Q value. To improve our inductive matching, source degeneration may be used. By inserting an inductance, L_S , at the source terminal as shown in Figure 9.22b, the impedance looking into the FET gate is supplemented with a fairly

broadband real part, $g_m L_S / C_{GS}$, according to

$$Z_{IN} \approx \left[R_G + R_{LS} + R_{LG} + \frac{g_m L_S}{C_{GS}} \right] + j \left[\omega(L_G + L_S) - \frac{(1 + g_m R_{LS})}{\omega C_{GS}} \right] \quad (9.51)$$

where R_G is the gate resistance, R_{LG} is the series resistance of the gate inductor, and R_{LS} is the series resistance of the source inductor. The remaining reactive components may then be resonated out as a narrowband net-zero reactance through the series inductor L_G at the operating frequency ω_0 , where

$$\omega_0 = \sqrt{\frac{1 + g_m R_{LS}}{(L_G + L_S) C_{GS}}} \quad (9.52)$$

The real-part matching then requires that

$$R_S = R_G + R_{LS} + R_{LG} + \frac{g_m L_S}{C_{GS}} \quad (9.53)$$

Or, effectively, that the source inductance relates to the unity current gain cutoff frequency, $f_T = g_m / (2\pi C_{GS})$, according to

$$L_S = \frac{R_S - R_G - R_{LS} - R_{LG}}{2\pi f_T} \quad (9.54)$$

with the net result that the overall impedance match depends mostly on the cutoff frequency of the device rather than on transconductance as is the case of the common gate topology. Because of this, the optimum noise figure may be attained with additional flexibility for device size, power consumption, and transconductance. The resulting noise figure for the common source topology¹⁵ is

$$F \approx 1 + \frac{R_G + R_{LS} + R_{LG}}{R_S} + \gamma \cdot g_{d0} R_S \left(\frac{\omega_0}{\omega_T} \right)^2 \quad (9.55)$$

where g_{d0} is the channel conductance.

As a narrowband solution, the common source topology is favorable, with the potential for significantly lower noise figures than the common-gate topology due to the $(\omega_0/\omega_T)^2$ factor in the Eq. 9.56 noise figure expression. The narrowband voltage gain of the LNA relies on a large load impedance that may be resonated with the parasitic transistor and load capacitance. In general, the voltage gain of the common source with the matching inductors in place can be roughly expressed as

$$\frac{V_2}{V_1} \approx \frac{-g_m R_L}{1 + g_m R_{LS} - \omega^2 C_{GS}(L_G + L_S) + j\omega[g_m L_S + C_{GS}(R_S + R_G + R_{LG} + R_{LS})]} \quad (9.56)$$

where R_S and R_L are the source and load resistances, respectively. For an optimum power match as described above, this reduces to

$$\left| \frac{V_2}{V_1} \right| \approx \frac{g_m R_L \sqrt{C_{GS}(L_G + L_S)}}{\sqrt{1 + g_m R_{LS} [g_m L_S + C_{GS}(R_S + R_G + R_{LG} + R_{LS})]}} \quad (9.57)$$

9.4.2 CMOS LNAs for Highly Integrated Transceivers

The first design example is in a 0.6- μm CMOS process LNA suitable for GPS application at 1.5 GHz¹⁵. It is a cascode configuration with inductively degenerated source, L_S , as shown in Figure 9.23.

The circuit uses a combination of package bondwire and a 4-nH on-chip spiral inductor for 50 Ω , the gate inductance, L_G , which resonates with L_S and C_{GS} to impedance match to the source impedance, R_S . The off-chip matching network shown as the series transmission line, T_m , and shunt capacitance, C_m , facilitated the input resonance to occur at 1.5 GHz and results in an input voltage standing-wave ratio (VSWR) of 1.38. The on-chip spiral load inductance of 7 nH, L_D , was made to resonate with the drain junction capacitance of M2 and the input gate capacitance of M3. The output of the second gain/buffer stage is an open load that is connected to an external bias tee for the dc supply and drives the spectrum analyzer directly. The measured gain of 22 dB was achieved for a power dissipation of 30 mW from a 1.5-V supply. Half of that power dissipation was due to the 15.2 mW consumed by the second stage, and is not necessarily reflective of the achievable 15 mW of an

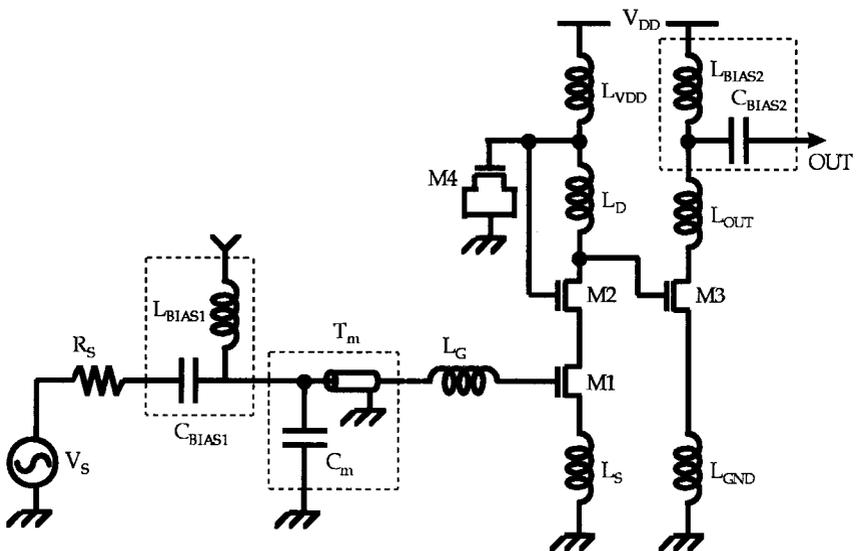


Figure 9.23 Low-noise amplifier in 0.6- μm CMOS using a cascode configuration.

integrated design that did not require an off-chip drive of the measurement equipment. The input referred IIP₃ was -9.3 dBm and reflects the emphasis of this design on low power and a low noise figure. The noise figure itself was measured as 3.5 dB, and may be expressed for this circuit configuration as

$$F = 1 + \frac{R_G}{R_S} + \gamma \cdot g_{d0} R_S \left(\frac{\omega}{\omega_T} \right)^2 \tag{9.58}$$

This expression indicates that device sizing may be optimized for both power and noise by minimizing the device width. The output conductance, *g_{d0}*, may be reduced by using smaller devices and, as long as the cutoff frequency, *f_T*, of the device is not degraded too much, the noise factor will be minimum. This means that small devices may be used to achieve both low power and low noise as long as the gate resistance is also minimized by using many parallel fingers. The impact of further reducing gate lengths from the 0.6-μm technology to 0.3-μm should result in noise figures of just over 1 dB for this design.

The second design example¹⁶ is a single-stage, differential-mode amplifier that builds on the previous single-ended device. The circuit schematic is shown in Figure 9.24.

The application for this differential topology is for wireless LAN at 1.9 GHz, with constraints for low-cost CMOS technology focusing on high levels of system integration and low power for mobile laptop PC use. The measured results for this

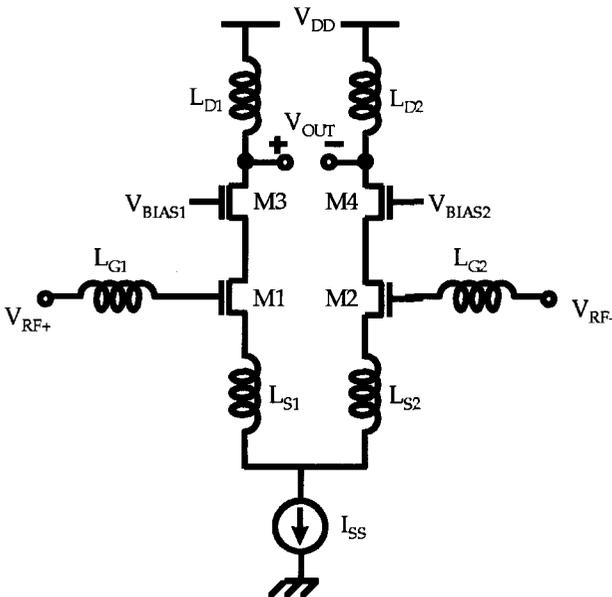


Figure 9.24 Schematic of a single-stage differential low-noise amplifier with a cascode configuration.

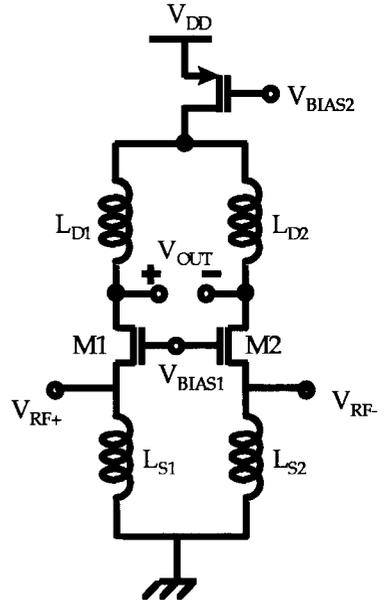


Figure 9.25 Differential low-noise amplifier design utilizing a single-stage common gate configuration.

LNA are a gain of 22 dB, a noise figure of less than 5 dB, and a power dissipation of 41 mW from a 3-V supply. As a differential topology, suppression of common-mode substrate coupling is a push toward more integrated solutions. This LNA was demonstrated in an entire integrated front-end receiver including the mixer and ADC and has an input-referred IIP_3 of -7 dBm for the entire receiver, an excellent linearity performance. An alternative differential topology is demonstrated in a third design example¹⁷, as shown in Figure 9.25.

The RF inputs are also differential here, but feed the upper FETs of the previous cascode design directly at their sources, and not through the gates of common-source devices. The load inductors, L_{D1} and L_{D2} , are extremely large 50-nH coils, which were fabricated using a special technique to remove the substrate material from underneath the coils and thereby reduce the significant parasitic capacitance of such large spirals. This technique facilitates the use of extremely large inductance that resonate to produce extremely large load impedance and improve the LNA gain response. The gain of the LNA peaks around 1 GHz at 20 dB. The noise match is fairly coincident with the input power match as reflected by the -16-dB return loss of the LNA at 1 GHz, which roughly corresponds to a 52.5Ω input impedance. The double-sideband noise figure of the LNA was measured to be about 3.2 dB, close to the theoretical limit of 2.9 dB of a single FET with a common gate input impedance of 70Ω . Similar to the differential cascode LNA described above, this was demonstrated in a complete receiver with mixer, and showed an even better IIP_3 of +8 dBm for the entire receiver, with the entire front-end receiver drawing only 9 mA from a 3-V supply.

The last design example is a CMOS-specific approach that incorporates the reuse of current to produce more transconductance at a given total drain.¹⁸ As shown in

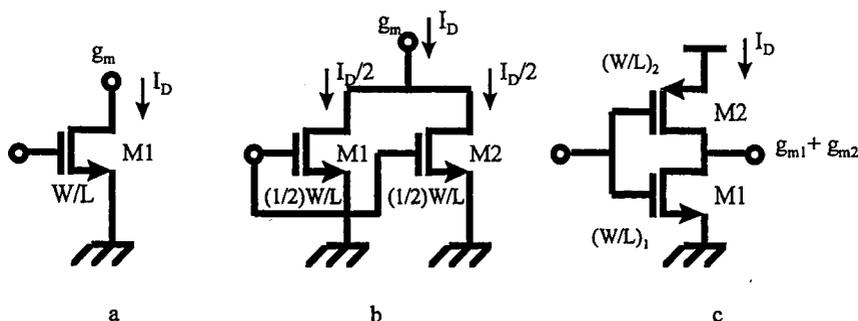


Figure 9.26 Comparison of the transconductance and current for (a) single NMOS, (b) scaled parallel NMOS devices, and (c) series connection of NMOS and PMOS to provide current reuse and to enhance transconductance.

Figure 9.26, for a given current (I_D), the transconductance for a single NMOS and parallel NMOS devices are the same, but by stacking a PMOS and NMOS in series, their transconductances are added when seen at the output node between them. This additional transconductance is a critical aspect of achieving a low noise-figure based on the expression for the noise factor of the common source NMOS topology as described above

$$F \approx 1 + \frac{8\omega^2 C_{GS}^2 R_S}{3g_{m1}} \quad (9.59)$$

Here, the transconductance must be maximized to reduce the noise figure, which often results in larger drain current and power dissipation for the LNA. By using the inverter topology of Figure 9.26, the transconductance is nearly doubled for the same drain current.

An advantage here is that the power consumption for the LNA may be reduced and the noise figure may be made lower even for that lower current level, so that both power consumption and the noise figure may be improved over standard NFET designs. The current reuse concept, as implemented in a full two-stage LNA shown in Figure 9.27, shows that for a two-stage design with a forward gain of 15.6 dB that the power dissipation can be as low as 20 mW for operation at 900 MHz.

The noise figure is an extremely low 2.2 dB into a 50- Ω -source match. Termination at the source would further reduce the noise figure to 1.9 dB. The linearity is excellent also, with an input-referred IIP₃ of -3.2 dBm. The use of the PMOS and NMOS devices demonstrates the usefulness of high-speed complementary technologies in the RF signal path and that the CMOS approaches lend versatility and alternative solutions to standard design architectures that can produce significant performance enhancements. Following is a summary of the state-of-the-art CMOS LNA performance outlined in Table 9.2.

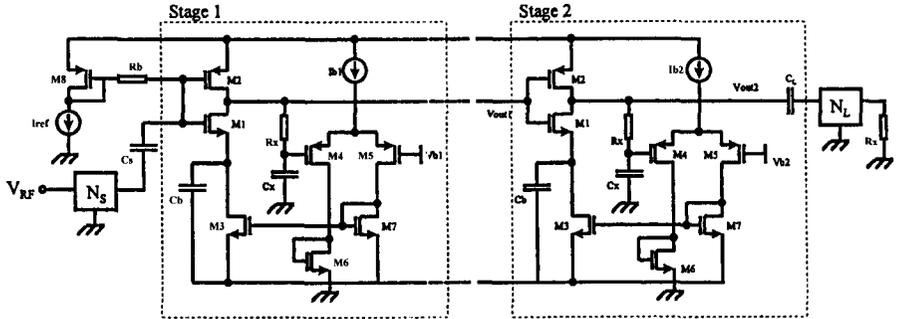


Figure 9.27 Full two-stage LNA design implemented using the concept of current reuse toward a fully integrated low-power implementation. (After Karanicolas, Ref. 17.)

TABLE 9.2 Comparison of Si-Based Monolithic LNAs

f_0 (Ghz)	Power (mW)	Noise Figure (dB)	Gain (dB)	IIP3 (dBm)	Area (mm ²)	Si-Based Process Technology	Reference
1.90	20	2.1	20	-2	15	0.6- μ m DPTM CMOS	16
0.9	20	2.2	5.6	-3.2	1.2	0.5- μ m DPTM CMOS	17
1.0	13.2	N/A	22	N/A	N/A	1.0- μ m DPTM CMOS	18

9.5 MIXERS AND FREQUENCY TRANSLATION

The frequency at which radio transmissions are made critically depends on many factors, such as propagation loss, available bandwidth, transmission power, cost of components, and required antenna size, to name just a few. Often, the frequencies of propagation are required to be high for available bandwidth and to reduce the antenna dimensions. The high-frequency carrier signals are modulated with a low-frequency signal that contains either voice and/or data information. To receive the information buried within these modulated sidebands, we must “downconvert” the signal by translating its frequency to a value low enough to permit the use of low-cost and low-power circuits. On the transmit side, the data information is used to modulate a signal or “upconvert” to the carrier signal at the propagation frequency. The component that performs these frequency translations is called a “mixer” and is fundamental to all types of communication systems. They rely on modulation and demodulation of high-frequency signals as discussed in the following section.

9.5.1 Mixer Fundamentals

The mixer is shown schematically in Figure 9.28 and illustrates the process of down conversion, showing how the RF input is heterodyned by the local oscillator, LO,

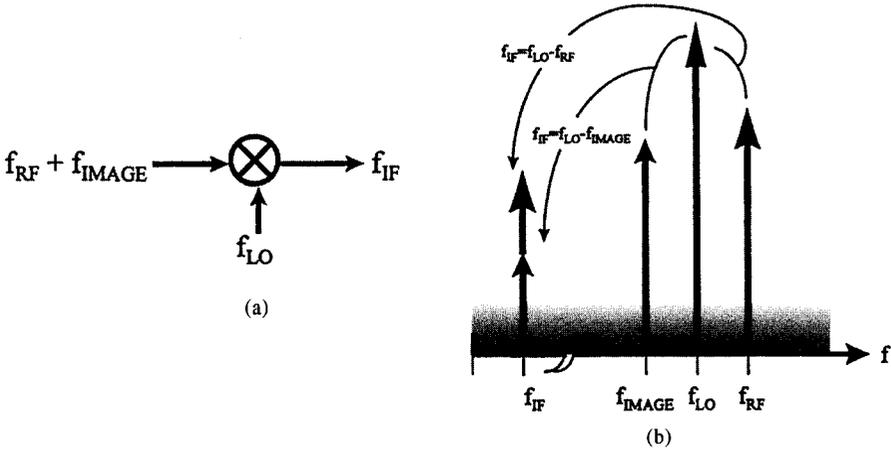


Figure 9.28 (a) A standard mixer that takes two inputs and produces f_{IF} at their difference frequency; (b) detail demonstrating the image frequency downconversion masking the desired frequency downconverted from f_f .

and an intermediate frequency, IF, results at the difference between the two frequencies. Similarly for the upconversion mixer, the input ω_{IF} multiplied by the local oscillator, ω_{LO} , generates ω_{RF} at the difference frequency. The basis for the frequency translation is that the multiplication of two sinusoids in the time domain is equivalent to the frequency shift in the frequency domain according to

$$\sin(\omega_1 t) \cdot \sin(\omega_2 t) = \frac{1}{2} \cos[(\omega_1 - \omega_2)t] - \frac{1}{2} \cos[(\omega_1 + \omega_2)t] \tag{9.60}$$

Note that the product of two signals at two different frequencies generates signals at both the sum and difference frequencies. This time-domain multiplication can be implemented in many ways using various circuit approaches. The first is simply the general phenomena that any nonlinear element, or circuit, can be considered as a nonlinear polynomial expansion, which includes higher-order products, and can be demonstrated by the following example of a nonlinear transconductance element that exhibits a nonlinear current (here expressed in a polynomial expansion with coefficients a_n) as a function of input voltage:

$$i = \sum_{n=0}^{\infty} a_n v^n \tag{9.61}$$

For a two-frequency input voltage with amplitudes A , frequency ω , and phase ϕ , of the form

$$v = A_{RF} \sin(\omega_{RF} t + \phi_1) + A_{LO} \sin(\omega_{LO} t + \phi_2) \tag{9.62}$$

the resulting current may be expressed as

$$i = \sum_{n=0}^{\infty} \sum_{p=0}^n C_p^n a_n V_{RF}^p V_{LO}^{n-p} \sin^p(\omega_{RF}t + \phi_1) \sin^{n-p}(\omega_{LO}t + \phi_2) \quad (9.63)$$

and includes the polynomials of the RF and LO fundamental terms, but also many cross-product terms, which are proportional to $a_n V_{RF}^p V_{LO}^q$ at frequencies $p\omega_{RF} \pm q\omega_{LO}$, where p and q are integers so that $p \geq q$ and $p + q = n$, the order of the nonlinearity. The mixing process is fairly inefficient with these mixing products of sum and difference frequencies becoming lower in amplitude with each increasing order of the multiplication, as is represented in the multiplication coefficients C_p . As a result, typically the second-order sum, $f_{RF} + f_{LO}$, or the difference, $f_{RF} - f_{LO}$, is of most interest. The complete expression of conversion loss and/or gain in translating between these frequencies is contained in the Manley–Rowe relations:¹⁹

$$\sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} \frac{mP_{m,n}}{mf_1 + nf_2} = 0 \quad (9.64)$$

$$\sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} \frac{nP_{m,n}}{mf_1 + nf_2} = 0 \quad (9.65)$$

where f_1 and f_2 are the frequencies of the input signals and $P_{m,n}$ is the resultant average real power at the mixer output at frequency $|mf_1 + nf_2|$. A further detail that is critical for the use of these mixers in actual systems concerns how they translate power at frequencies that are not the intended carrier. As an example in Figure 9.28b, the RF and LO frequencies will downconvert the intended signal power to an intermediate frequency $f_{IF} = f_{LO} - f_{RF}$, but will also downconvert unintended signal power from what is called the “image” frequency from f_{IMAGE} to $f_{IF} = f_{LO} - f_{IMAGE}$. This unintended power at the image frequency will directly overlap the desired signal, and will mask it with either an interference signal or simply the noise present at f_{IMAGE} , reducing the signal-to-noise ratio (SNR) resulting at the IF frequency.

To convert the RF signal down to a large output signal at the IF frequency based on device nonlinearity, two conditions must be met. The second-order nonlinear coefficient of the device, a_n , must be large, and the LO power must be large because the IF output power is directly proportional to it. The output IF signal will be maximized when the LO power is saturated, or equivalently clipped at its maximum and minimum values in a square wave. This mixer operation with the LO saturated only weakly depends on the amplitude of the LO for further increases in LO drive. This “square wave” LO can be considered simply as an ideal switch, and the output IF simply as the result of its multiplication with the RF signal as shown in Figure 9.29a. The conversion loss (or gain) may be expressed as the ratio of power at the IF frequency into a load, Z_L , to the available power from the source, Z_S , at the RF input frequency. In this first example, the expansion of the LO signal in the time

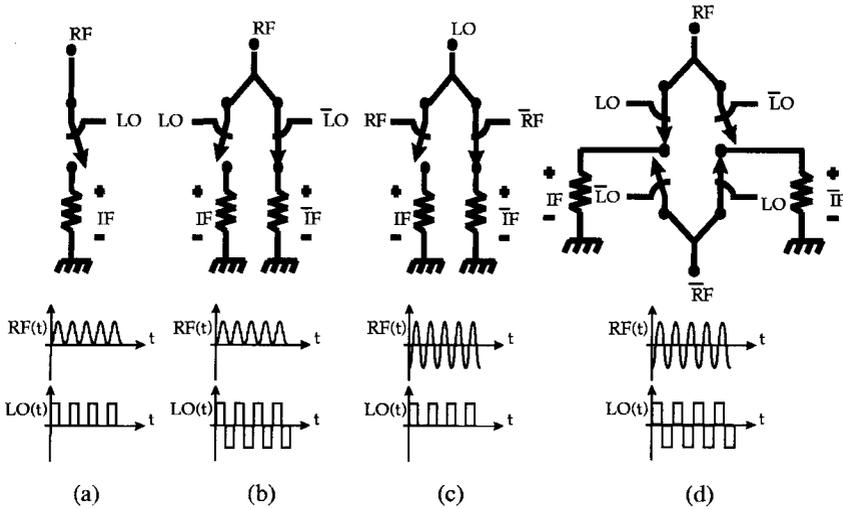


Figure 9.29 Mixer representations: (a) single-ended mixer, (b) LO-balanced, (c) RF-balanced, (d) double-balanced.

domain for a 50% duty cycle and ideal waveforms results in a conversion loss of $-20 \log(1/\pi) = -9.9$ dB. This is because the LO signal switches between 0 and 1. Therefore, its dc constant average value is positive and nonzero and, in multiplication with the RF signal, translates into the feedthrough of the RF signal to the IF output. In the second case in Figure 9.29b, two single-ended mixers have been combined so that the switching generates a zero dc average value to avoid this RF feedthrough by reversing the LO from positive to negative each half-cycle. This configuration is referred to as a *balanced mixer*, and its RF signal energy is completely converted to the IF frequency. This more efficient mixing improves the conversion loss by a factor of 2 to become $-20 \log(2/\pi) = -3.9$ dB, which is the minimum conversion loss achievable by a passive mixer.

The third example, in Figure 9.29c, demonstrates an RF-balanced-mixer topology that provides for zero dc content in the RF signal switching, thereby reducing LO feedthrough to the IF port. The final example, in Figure 9.29d, illustrates the case when both the RF and LO switching is symmetrical among four branches, or double-balanced. In this case the RF feedthrough and LO feedthrough to the IF port are minimized, as well as harmonics of the RF and LO which are also cancelled at the IF port. The additional advantage of the double-balanced configuration is its ability to reject the two-tone second-order products at the IF output. A summary of these various configurations, as well as their respective performance advantages, are detailed in Table 9.3¹⁹

The double-balanced configuration is the focus of most of the implementations seen for commercial application because of its excellent signal isolation between the ports and low spurious content at the output. This double-balanced-switching configuration, shown in Figure 9.29d can be viewed simply as a switch matrix that “steers” current to either one branch or another based on the LO switch settings. The

TABLE 9.3 Performance Comparison of the Various Mixer Configurations and Balancing

Performance Characteristic	Mixer Configuration			
	Single-Ended	LO-Balanced	RF-Balanced	Double-Balanced
LO/RF isolation	Poor	Good	Poor	Good
LO/IF isolation	Poor	Poor	Good	Good
RF/IF isolation	Poor	Good	Poor	Good
LO harmonic rejection	None	Even	All	All
RF Harmonics rejection	None	All	Even	All
Single-tone spur rejection	None	$p_{rf}^+ q_{LO}$ with q even	$p_{rf}^+ q_{LO}$ with p even	$p_{rf}^+ q_{LO}$ with p or q even
Two-tone 2nd-order intermod. rejection	No	No	Yes	Yes

CMOS implementation of these topologies is accomplished by simply substituting appropriate MOSFET devices for the switches, as in the case of the double-balanced FET ring shown in Figure 9.30.

An important point is that passive and upper quad of the active mixer topologies are actually quite the same. In fact, Figure 9.30 demonstrates that the FET ring topology, M1–M4 in Figure 9.30a, and the four upper FETs of the Gilbert cell active mixer topology, M1–M4 in Figure 9.30b (to be discussed in detail later), are actually the same circuit. The active Gilbert cell schematic simply includes two additional

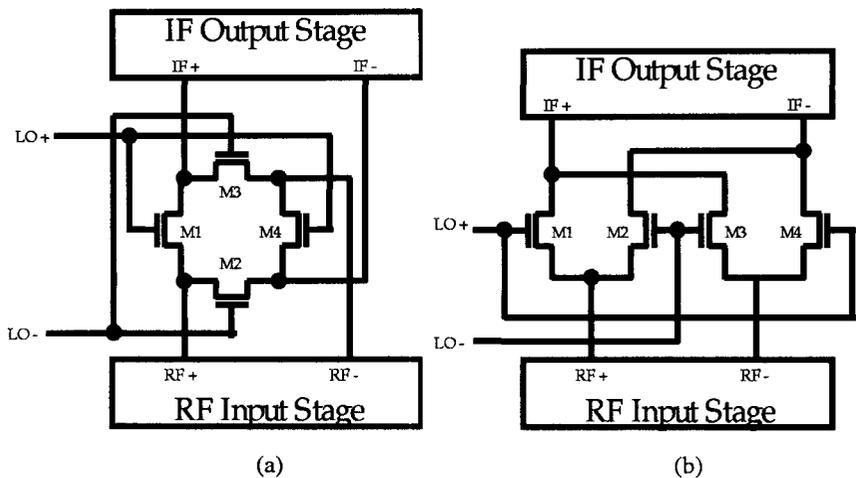


Figure 9.30 General topology of differential mixers. The (a) “FET-ring” topology and (b) upper quad of the “Gilbert cell” topology are the same circuit, simply redrawn.

FETs in a differential pair making up the RF input stage at the bottom of the figure. The fundamental difference between active and passive mixers really has to do with this RF input stage and whether some form of amplification is cascaded in front of it. This amplification preceding the RF input has a significant effect in reducing the mixer noise figure, but may limit the linear dynamic range if the RF input FETs are easily driven into compression or saturation.

9.5.2 CMOS Mixers for Highly Integrated Transceivers

The first design example demonstrates the capability of CMOS in passive FET ring topologies for high-frequency downconverters.²⁰ The FET ring as discussed in the previous section may be driven with the RF on the drain of the MOSFETs, or with the RF on the gates. The use of the gate drive for RF is a critical design aspect that focuses on linearity and power efficiency for the passive ring topology. The circuit schematic is shown in Figure 9.31.

The MOSFETs are biased strongly in the linear regime with a large $V_{GS} - V_t$ to assure a small R_{ON} for a small input MOSFET that maintains good frequency response and large input bandwidth. The linearity of the MOSFET ring does not depend on the voltage-to-current conversion characteristic of the transistors. Linearity is simply a function of the speed of the FETs as pass transistors and any spurious signals generated during their switching. The specific dc bias must maintain the MOSFETs in the triode region, meaning that the minimum dc bias on the RF ports must be at least V_t above the largest possible source voltage to keep their channels strongly inverted and out of cutoff, and the drainsource voltage must be kept lower than the smallest $V_{GS} - V_t$ to keep the MOSFETs from saturating. The voltage-range limitations of the design are defined according to the following

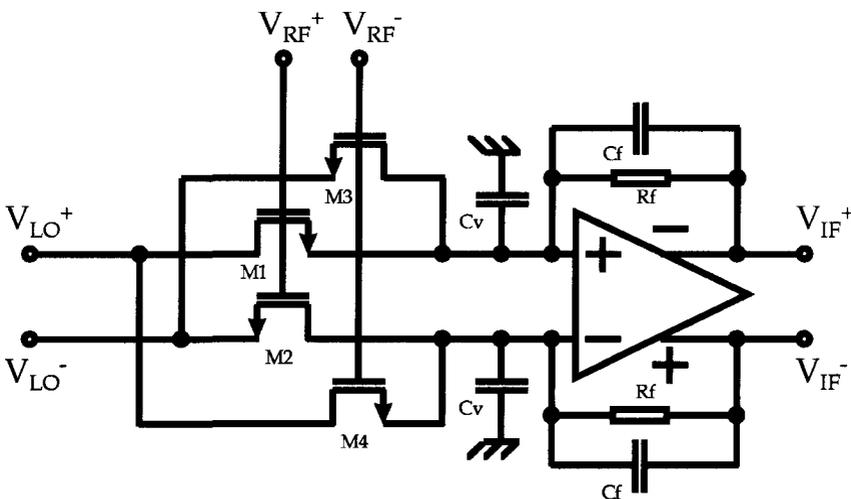


Figure 9.31 A gate-driven 1.5-GHz highly linear CMOS downconversion mixer. (After Crols and Steyaert, Ref. 20.)

relationships for maximum voltage swing and describe the fundamental power limitations for maximum input power on the RF and LO ports:

$$\text{Maximum LO signal (differential } V_{p-p}) = V_{\text{LO,AC}}^{\text{MAX}} \leq 4 \cdot V_{\text{LO,DC}} \quad (9.66)$$

$$\text{Maximum LO signal (dBm)} = 10 \times \log_{10} \left(\left(\frac{1000}{2.50 \Omega} \right) \cdot \left(\frac{V_{\text{LO,AC}}^{\text{MAX}}}{2} \right)^2 \right) \quad (9.67)$$

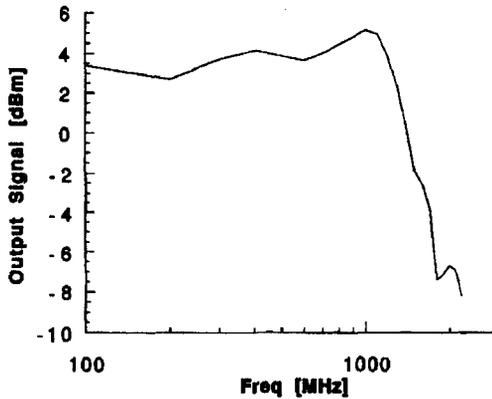
$$\text{Maximum RF signal (differential } V_{p-p}) = V_{\text{RF,AC}}^{\text{MAX}} \leq 4 \cdot (V_{\text{RF,DC}} - V_{\text{LO,DC}} - V_T) - V_{\text{LO,AC}} \quad (9.68)$$

$$\text{Maximum RF signal (dBm)} = 10 \times \log_{10} \left(\left(\frac{1000}{2.50 \Omega} \right) \cdot \left(\frac{V_{\text{RF,AC}}^{\text{MAX}}}{2} \right)^2 \right) \quad (9.69)$$

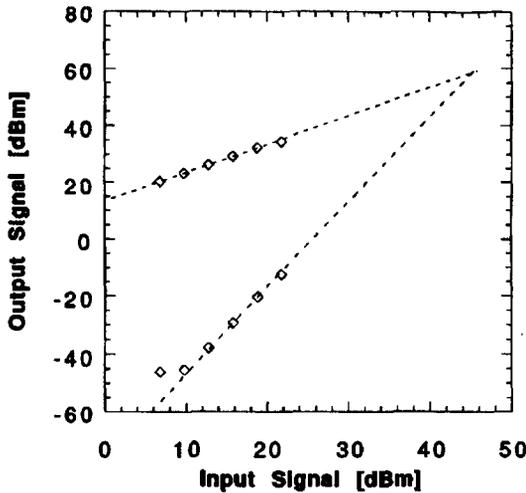
This particular implementation used an LO dc level of $V_{\text{LO,DC}} = 1.15$ V, and an RF voltage of $V_{\text{RF,DC}} = 3.85$ V, so that with an LO ac level of $2.5 V_{p-p}$ (which is below the limit of $4.6 V_{p-p}$) the maximum RF signal level of $5.5 V_{p-p}$ (18.7 dBm) could be attained. The cross-coupled, double-balanced structure of this FET ring assures that all quadratic components in the transconductance of the FET characteristics cancel each other, as well as canceling the common-mode dc bias effects, bulk substrate effects, and the nonlinear dependence of g_{DS} on the drain-to-source voltage V_{DS} . This particular implementation required special attention to the IF termination using large capacitive loading of the op amp virtual-ground inputs, shown in Figure 9.31. To terminate the high-frequency signals passed by the cross-coupled NMOSs, these large capacitors are used to filter them out and prevent them reaching the op amp input. The op amp inputs still provide adequate virtual grounds at low frequency and with the higher-frequency filtering in place, its low-frequency performance may be optimized independent of the higher-frequency constraints of the FET ring. For the op amp following the mixer stage, a relationship describing the effect of gain mismatch, $\Delta\beta$, in the op amp produces the following quadratic distortion in the output signals:

$$V_{\text{OUT}}^+ - V_{\text{OUT}}^- = \beta \cdot R_f \cdot (V_{\text{RF}}^+ - V_{\text{RF}}^-) \cdot (V_{\text{LO}}^+ - V_{\text{LO}}^-) + \Delta\beta \cdot R_f \cdot (V_{\text{LO}}^+ - V_{\text{LO}}^-)^2 \quad (9.70)$$

which further illustrates the design criteria of using the RF to drive the gates. The second-order term in the LO produces a signal at twice its frequency, which is easily filtered out and a dc component plus small contribution due to phase noise at a very low frequency. If the LO were driving the gates, the second-order term above would be a function of $(V_{\text{RF}}^+ - V_{\text{RF}}^-)^2$ and produce an undesirable low-frequency signal over the entire bandwidth of the RF signal (~ 100 MHz). This may be an important aspect for direct conversion topologies that convert the RF signal directly to a zero IF frequency about dc and cannot afford the overlay of this mismatch onto the RF bandwidth. Another advantage of driving the gates with the RF is power



(a)



(b)

Figure 9.32 (a) The input bandwidth of the CMOS downconversion mixer; (b) the measured IP₃ of the CMOS mixer.

consumption, which for this design is minimal and measured to be 1.3 mW from a 5-V power supply. Even at these low static power consumption levels, the mixer exhibits extremely high linearity as shown in Figure 9.32 with a measured IIP₃ of 45.2 dBm, and a third-order IMD₃ of -46.4 dB at the calculated maximum RF input power of 22 dBm.

The conversion gain of the mixer/op amp chain is 18 dB for a 12-dBm differential LO signal. The noise figure of this mixer, 24 dB, is much higher than desired for

many applications, because of the insertion loss of the ring topology as implemented here. However, the intermodulation-free dynamic range; over which the mixer operates from noise floor to acceptable IMD_3 levels, is 59.6 dB. Since the preceding gain stages may often mitigate the impact of mixer conversion loss and noise figure, this design affords some wide dynamic range advantage.

A second design approach using the passive ring topology addresses the noise figure issue of the first design by sacrificing some dynamic range. This design provides conversion gain instead of loss and minimizes the power consumption compared to other active mixer cells by implementing a current reuse approach. Similar to the LNA example discussed previously, the mixer implementation embeds a passive FET ring topology made up of both NMOS and PMOS FETs between the NMOS and PMOS of a standard CMOS inverter.¹⁸ In this arrangement, the additional transconductance of the lower and upper NMOS and PMOS respectively contribute to amplification of the RF input signal, reducing the noise figure of the embedded mixer block, and providing positive conversion power gain for the mixer. Figure 9.33 shows that the input matching for the RF is similar to that of the LNA examples discussed above, using the inductors L_G , L_{S1} , and L_{S2} to resonate with the input capacitance of the outer NMOS and PMOS to impedance-transform the input to $50\ \Omega$.

The embedded mixer cell design is shown in Figure 9.34, and illustrates how the switching of the LO voltage “steers” the current either through the left or right NMOS/PMOS pair as a function of LO phase.

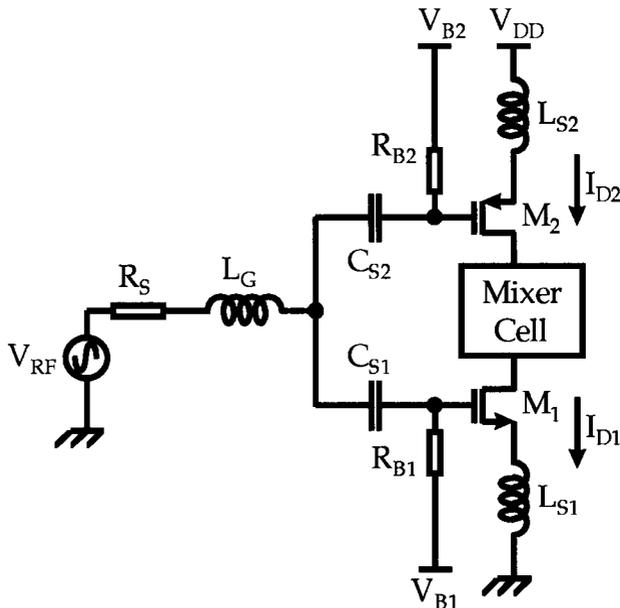


Figure 9.33 Mixer schematic employing current reuse on an embedded mixer cell for application at 900 MHz.

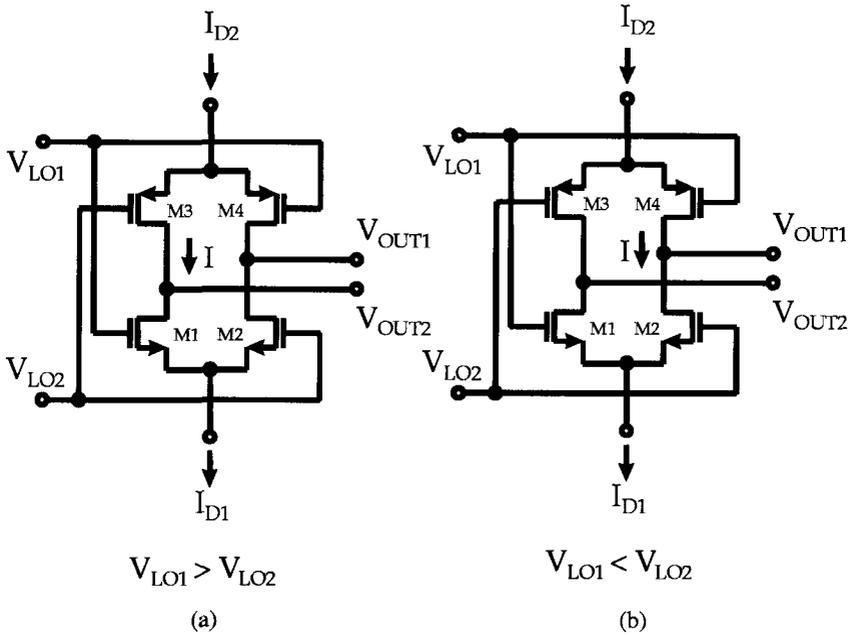


Figure 9.34 Detail for the embedded mixer cell of Figure 9.33 showing current steering during the two extremes of balanced LO switching: (a) $V_{LO1} > V_{LO2}$; (b) $V_{LO1} < V_{LO2}$.

This current steering chops the RF signal down to the IF outputs located between those transistor pairs. The FET ring topology discussed here uses the LO signal to drive the gates of the ring and will have a slightly lower intrinsic conversion loss for the ring at a slightly higher power consumption, measured here to be 7 mW from a 2.7-V supply. This value is as small as it is because of the efficient re-use of the current to reduce the required power consumption. The experimental results are impressive, with high-side downconversion using 0 dBm LO power at 1 GHz to translate the 900-MHz RF signal down to a 100-MHz IF. The mixer provides a double-sideband noise figure referred to $50\ \Omega$ of only 6.7 dB from a 2.7-V supply. This double-sideband noise figure may be converted to a single-sideband noise figure by simply adding 3 dB as long as $f_{LO} - f_{IF}$ and $f_{LO} + f_{IF}$ exhibit the same conversion gain. That is the case here and the measured conversion gain is 8.8 dB. The linearity is not as good as the previous FET ring example. However, it has a significantly lower IIP3 of -4.1 dBm and an input 1 dB compression level (maximum RF input signal power) of -16.1 dBm. This design example is specifically useful with the complementary logic capabilities of CMOS technology and cannot be implemented without fairly high-speed PMOS devices, which can be made with today's deep-submicrometer CMOS processes. These first two design examples of passive FET ring topologies illustrate the tradeoff between linearity, noise figure, and conversion loss/gain. The first is more appropriate for zero IF second-stage, downconversion that has an ultralinear characteristic and wide

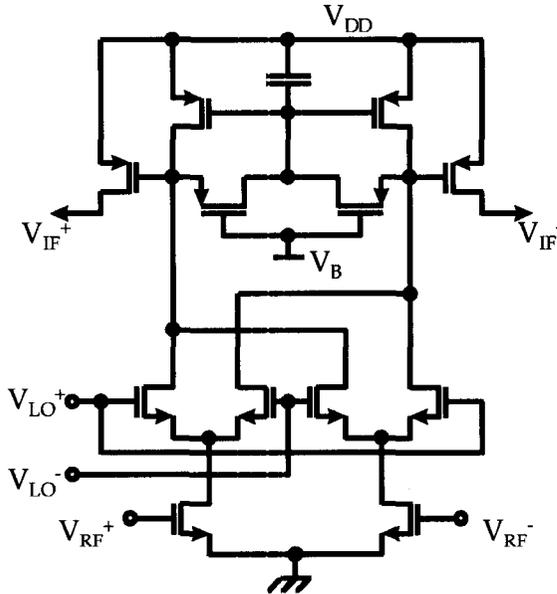


Figure 9.35 Gilbert cell topology for a fully balanced mixer implementing active loading with tunable voltage control.

dynamic range. The second is appropriate for first-stage, nonzero IF exhibiting a low noise figure and reasonable conversion gain. Both demonstrate the ability of CMOS to perform at extremely low power dissipation levels and lead us to our third design example of an active Gilbert cell mixer topology.

The Gilbert cell topology, is shown in Figure 9.35, represents a double-balanced mixer made up of a quad-FET ring, which is gate-driven by the differential LO, and the lower differential pair, which is driven by the RF. In this way, the currents flowing in the differential pair are switched by the LO to flow into one or the other PFET active loads.

The RF differential pair consists of source-coupled NMOS FETs in saturation and serves as a linear voltage-to-current converter producing a linear differential-output current as a function of its small-signal input voltage:

$$I_{\text{OUT,RF}} = \mu C_{\text{OX}} \frac{W}{L} (V_{\text{GS}} - V_{\text{T}}) V_{\text{IN,RF}} \quad (9.71)$$

It is this relationship that sets the overall linearity of the multiplication process for the mixer. The noise of the mixer is set by the LO switching of the upper quad. It contributes negligible noise during the regions of limited amplitude between transitions when the LO state is fixed to direct current through one or the other load. However, an extremely large burst of noise is generated as the quad amplifies its input signals during the zero-crossing transitions between commutations. During these time periods, when both loads are driven with current from the LO quad, the

noise figure of the mixer is degraded and it is critical to minimize this time and maximize the speed at which the zero crossings are executed. Large LO signals and steep slopes for the LO quad output current, zero crossings will reduce the mixer noise figure. Another noise contributor is the low-frequency noise of the loads. In this case they are designed to be extremely large PMOS devices, each with a total gate width of $960\ \mu\text{m}$, so that the flicker noise components of the current sources may be minimized. For direct conversion where the IF is at very low frequencies, the $1/f$ noise magnitudes become critical and may mask the resultant mixer output if both the design and the device technology are not carefully controlled. This design was fabricated in a standard $1\text{-}\mu\text{m}$ CMOS process and demonstrates excellent performance even for these relatively long gate lengths. The conversion gain measured at $900\ \text{MHz}$ was $-3\ \text{dB}$ and was compromised somewhat to achieve the excellent linearity of an input-referred IIP3 of $+30\ \text{dBm}$. The LO level is $1\ V_{p-p}$ and translates into approximately $+4\ \text{dBm}$ into a $50\text{-}\Omega$ load, and results in a double-sideband noise figure of $15\ \text{dB}$ at $\text{IF} = 10\ \text{MHz}$ and $20\ \text{dB}$ at $\text{IF} = 160\ \text{kHz}$. The additional 5-dB noise at the lower IF is due to the flicker noise of the PMOS loads and the LO quad during zero-crossing.

The implementation of this quadrature-cancellation technique is extended in the following design example (Fig. 9.36). By implementing a second pair of mixers that are cascaded to combine the I and Q outputs of the four input paths, one can cancel all the unwanted image frequency sidebands and coherently sum the desired downconverted RF.¹⁶

This approach is particularly appropriate for highly integrated CMOS-based designs in that it is fully differential for all signals to avoid substrate coupling and common-mode interference and does not need any off-chip components, namely, the image reject filter. The quadrature combination of the signals provides a measured

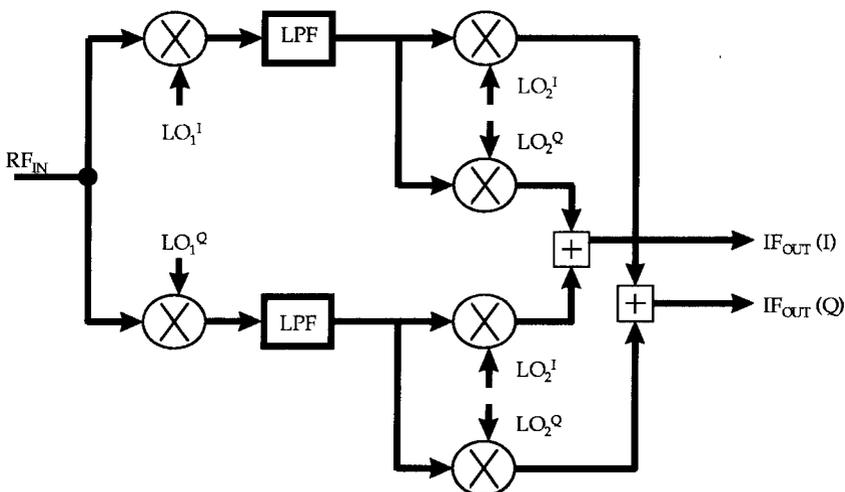


Figure 9.36 Extended two-stage Hartley quadrature mixer architecture. (After Rudell et al., Ref. 16.)

45 dB of image rejection. The performance is excellent with both the first and second stage RF to IF mixers consuming 17 mW from a 3.3-V supply with a variable-gain adjustment of 0–10 dB provided by a tunable MOS diode control of the mixer loading.

The mixer as part of an entire integrated CMOS receiver has a sensitivity of –90 dBm, an input-referred IP3 of –7 dBm, and a receiver gain that is adjustable from 26 to 78 dB. The total receiver image rejection is 55 dB, and downconverts a 1.9-GHz signal to on-chip switch capacitor filtering and an integrated 10-bit DACs (digital/analog converters). Table 9.4 summarizes the state-of-the-art CMOS mixers discussed in this section.

9.6 CMOS VOLTAGE-CONTROLLED OSCILLATORS

The process of generating a modulated carrier for transmission in communication systems requires a spectrally pure source for the carrier. Similarly, in receiving an incoming signal, a very clean signal must be used to downconvert and demodulate it. These signal sources constitute an entire class of circuits called *oscillators*, which most often are embedded in phase-locked-loop (PLL) structures that stabilize and control the output frequency through feedback and comparison to an ultraclean reference signal at a lower frequency. These synthesizer architectures are fundamentally based on the oscillator itself and require a facility for tuning using a control voltage, V_{CONTROL} , that allows some finite adjustment in output frequency, ω_{out} . The following sections will discuss the fundamental properties of CMOS voltage-controlled oscillators (VCOs), and provide some design examples from the state of the art.

9.6.1 Voltage-Controlled Oscillator Fundamentals

In its simplest form the VCO is a voltage-to-frequency converter as shown in Figure 9.37 according to

$$\omega_{\text{OUT}} = \omega_0 + K_{\text{VCO}} \cdot V_{\text{CONTROL}} \quad (9.72)$$

where K_{VCO} is in units of hertz per volt and expresses the relationship between output frequency and input voltage. A key measure of the VCO is the linearity of this relationship for application in PLLs.

The fundamental block signal diagram of an oscillator is as shown in Figure 9.38,²¹ where $K(A)$ represents the gain of an active device in the loop and is dependent on the oscillation output amplitude A , $H(\omega)$ represents a frequency selective or resonant element, and n represents the noise inherent in the loop. From this general diagram, all basic oscillator topologies may be described and the two shown in Figure 9.38*a,b* are negative conductance and negative impedance oscillators showing the detail of how the transistors and resonators relate to this general schematic.

TABLE 9.4 Comparison of Si-Based Monolithic Mixers

f_0 (Ghz)	Design	Power (mW)/ Supply (V)	DSB NF (dB)	LO Power (dBm)	Gain (dB)/ LO Power (dBm)	IIP3 (dBm)	Area (mm ²)	Si-Based Process Technology	Reference
1.5	FET Ring	1.3/5	32	20/12	20/12	45.2	1	1.2- μ m CMOS	20
0.9	I_{ds} reuse	7/2.7	6.7	8.8/0	8.8/0	-4.1	1.2	0.5- μ m CMOS	17
0.9	Gilbert	15/3	15	-3/4	-3/4	+30	NA	1.0- μ m CMOS	18

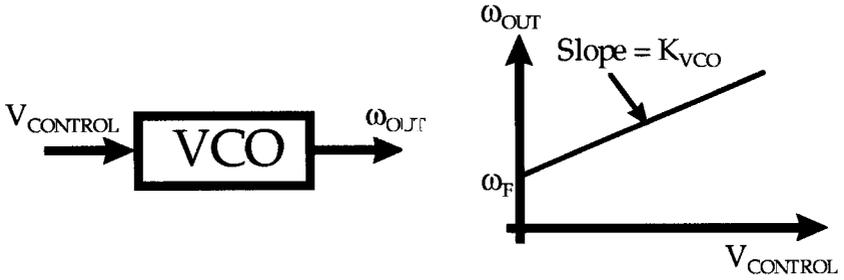


Figure 9.37 The general form of a voltage-controlled oscillator (VCO) and its associated tuning characteristic.

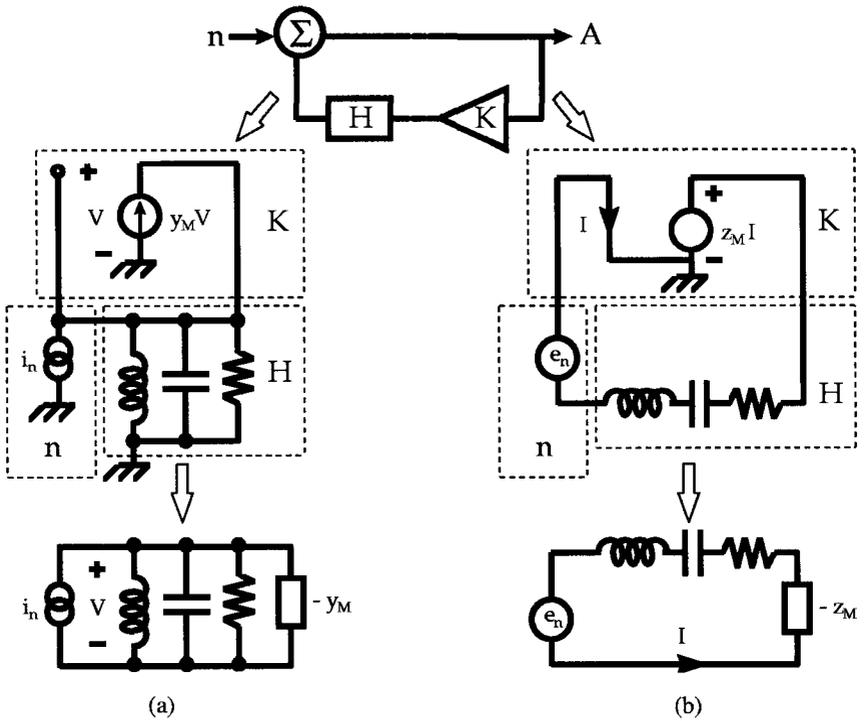


Figure 9.38 Fundamental block diagram of any oscillator, and decomposition into its representation of (a) negative conductance and (b) negative impedance schematics.

The expression describing this oscillator is

$$A = \frac{n}{1 - KH} \tag{9.73}$$

where $(1 - KH)$ is the inverted transfer function of the oscillator and relates the additive noise of the circuit to the final amplitude at its output according to

$$n = DA \tag{9.74}$$

where

$$D = 1 - KH = \frac{Z_{\text{DEVICE}} - Z_{\text{LOAD}}}{R} = \frac{Y_{\text{DEVICE}} - Y_{\text{LOAD}}}{G} = 1 - \Gamma_{\text{DEVICE}} \Gamma_{\text{LOAD}} \quad (9.75)$$

where Z_{DEVICE} , Z_{LOAD} , Y_{DEVICE} , Y_{LOAD} , Γ_{DEVICE} , and Γ_{LOAD} are the impedance, admittance, and reflection coefficient looking into the device and load, respectively. This inverted transfer function, D , plays a key role in all analysis of the oscillator and provides a broad description of its behavior, the most fundamental of which is the condition for oscillation that establishes the frequency at which the oscillator will operate described by the Barkhausen criteria as

$$D(A_0, f_0) = 0 \quad (9.76)$$

where A_0 is the steady-state amplitude of oscillation at the output and f_0 is the operating frequency. This translates to the requirement that the loop gain of the oscillator, $K(A_0) \cdot H(\omega)$, be exactly equal to 1 in operation, and that the total phase shift of the loop be zero or, equivalently, an integer multiple of 2π . The fundamental basis for this can be seen in the general schematic of Figure 9.38 and expression (9.71), where finite injected noise is divided by the inverted transfer function, D , to produce an output amplitude sinusoid at a given frequency set by $H(\omega)$. When D goes to zero, large amplification of the noise at that very specific frequency occurs, producing a large output amplitude. The limit to this process occurs because the loop gain, $K(A_0) \cdot H(\omega)$, reduces to a steady-state value of 1 where the device gain exactly compensates the losses. At that point the loop gain is just enough to maintain the oscillation at this stable amplitude and frequency. The mechanism of this limiting depends on the circuit topology, but is often a result of device nonlinearity that results in odd-order harmonics. These harmonics combine to create a component at the fundamental frequency that cancels part of the input and lowers the loop gain to satisfy the Barkhausen criteria. A difficult task in the design of oscillators is the calculation of the final output power. The frequency itself is easily determined by small-signal analysis and the resulting phase condition that it must be zero around the loop. However, the steady-state oscillator output and phenomena during startup of the oscillation are strongly a function of the active device's nonlinearities, and require an accurate nonlinear model.

Important characteristics of oscillators for application in communication systems are that they have a wide tuning range, the loop gain and voltage remains fairly constant, the input–output characteristic is relatively linear, and the frequency is spectrally pure and exhibits minimum noise modulation. The latter design consideration requires significant attention. Noise in oscillators consists of both amplitude fluctuation and phase fluctuation as described by and shown in Figure 9.39.

$$A = (A_0 + \Delta a) \cdot \sin(\omega_0 t + \varphi + \Delta \varphi) \quad (9.77)$$

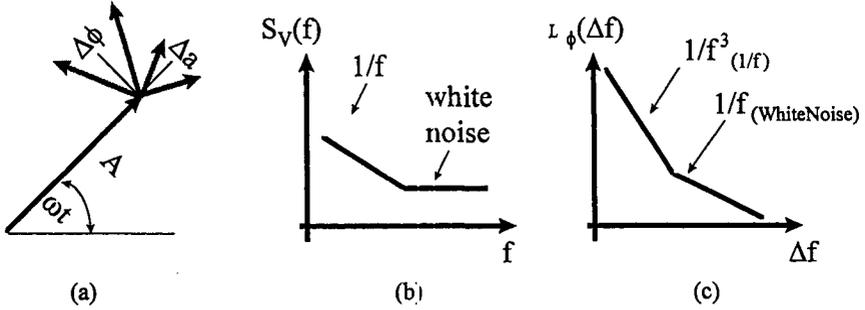


Figure 9.39 The graphical representation of oscillator noise as (a) the AM (Δa) and PM ($\Delta \phi$) fluctuations on signal A, (b) the baseband noise $S_V(f)$ as a function of frequency, and (c) its conversion to sideband phase-noise.

These amplitude-modulated (AM) and phase-modulated (PM) signals may be considered as orthogonal components, which induce random fluctuations in what is called *near-carrier noise* of the oscillation. This noise causes a spectral broadening of the carrier from a pure single-frequency tone to a tone with significant sidebands or noise “skirts,” which consist of both AM and PM components, as shown in Figure 9.39. The Δa and Δf may consist of additive and/or multiplicative components. *Additive noise* is noise at the same frequency as the carrier that simply adds to its spectrum, while *multiplicative noise* is frequency translated from another frequency due to device nonlinearity to mix the carrier and noise so that noise skirts are formed as a combination of the two. The actual noise components that make up these skirts may consist of either white noise, whose power spectral density is uniform over all frequencies, or flicker noise, which is exhibited only at extremely low frequencies with a power spectral density increase as the frequency decreases. Overall, it exhibits a slope/frequency curve of $1/f$, as shown in Figure 9.39, along with the associated phase noise that results.

For a pure $1/f$ flicker noise characteristic, the noise power integrated over frequency is a constant, and the $1/f$ characteristic extends from extremely low frequencies (measurements have verified $1/f$ down to 10^{-3} Hz), up to where it meets the white noise at the corner frequency, f_c . Because of this very large amplitude noise at frequencies that, when upconverted to the carrier are largest near the oscillation frequency, the largest component of phase-noise sidebands is due to $1/f$ noise. This is especially true for CMOS-based VCOs because CMOS devices exhibit extremely high $1/f$ noise levels compared to other technologies, and have corner frequencies, typically at 1 MHz and above. The basic representation for the input-referred, $1/f$ gate-voltage noise spectral density, $e_n^2(1/f)$, in CMOS FETs operated in saturation is expressed by

$$e_n^2(1/f) = \frac{k_f}{C_{OX} W_{EFF} L_{EFF} f^{ef}} \tag{9.78}$$

where $K_f(V^2F)$ is a parameter whose value depends on the technology used and takes on values typically on the order of $10^{-24}V^2F$, C_{OX} is the oxide capacitance density in $F/\mu m^2$, L_{EFF} and W_{EFF} are the effective gate length and width, respectively, in μm^2 , f is the frequency, and ef is the power of the frequency that describes the noise, typically equal to 1. Generally speaking, the input-referred voltage noise is somewhat independent of bias conditions in both gate voltage and drain current. The noise is inversely proportional to the active gate area because the surface-state interaction suspected to be the cause of the noise is averaged out over a larger area and effectively reduced. The inverse dependence on gate capacitance per unit area, C_{OX} , can be physically understood from the fact that the surface-state charge density, Q_{SS} , contributes a fluctuation to the threshold voltage that is inversely proportional to C_{OX} . As we approach higher-performance processes that have shorter gate lengths and thinner gate oxides, this situation fundamentally remains the same significant problem for near-carrier phase noise in oscillators.

9.6.2 CMOS VCOs for Highly Integrated Transceivers

The first design example is a ring oscillator based on cascading of an odd number of inverting stages whose output is cycled back to its input. The output is an inverted (and delayed) version of the input, and when it arrives back at the input node, it switches the input to an opposite state, initiating the entire cascade to toggle their states until the output changes state once again. The entire delay of the chain is equal to the sum of all the separate stage delays. Because the inverting signals require two loops around to complete a period for the oscillation, the total delay of the ring oscillator may be expressed as

$$T_{OUT} = 2NT_D \tag{9.79}$$

where T_D represents the stage delay for the case where all N stages are the same, as shown in Figure 9.40.

The general schematic may be broken down more specifically to a linearized model as in Figure 9.41.²³

Where each stage exhibits an inverting transconductance, $-G_M$, output load resistance, R , load capacitance, C , and may be modeled as including an output noise current, I_n . A critical aspect of ring-oscillator design is the number of stages and how the power consumption, speed, and phase noise result. By calculating the open-loop transfer function and evaluating the additive noise, the power dissipation versus

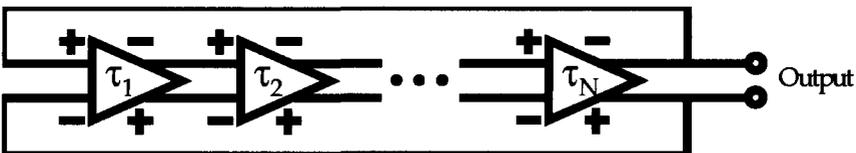


Figure 9.40 General schematic of a differential ring oscillator VCO.

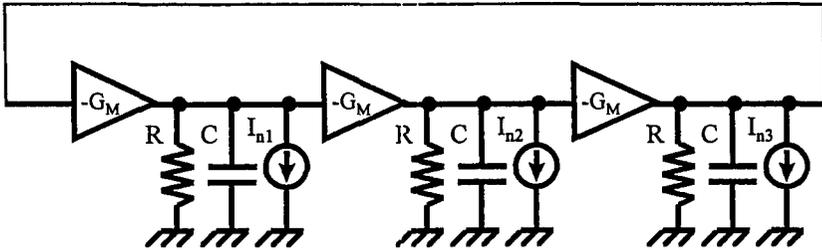


Figure 9.41 General linearized model for a three-stage ring oscillator.

TABLE 9.5 Comparison of Three- and Four-Stage Ring-Oscillator Design

Performance Characteristic	Ring Oscillators	
	3-Stage	4-Stage
Minimum required dc gain	2	$\sqrt{2}$
Noise shaping function	$\frac{R^2}{27} \left(\frac{\omega_0}{\Delta\omega}\right)^2$	$\frac{R^2}{16} \left(\frac{\omega_0}{\Delta\omega}\right)^2$
Open loop— Q	$\frac{3\sqrt{3}}{4} \approx 1.3$	$\sqrt{2} \approx 1.4$
Total additive noise	$8kT \frac{R}{9} \left(\frac{\omega_0}{\Delta\omega}\right)^2$	$8kT \frac{R(1 + \sqrt{2})}{12} \left(\frac{\omega_0}{\Delta\omega}\right)^2$
Power dissipation (mW)	1.8	3.6

Source: Razavi, Ref. 23.

phase noise may be evaluated as summarized in Table 9.5 for three-stage and four-stage designs.²³

For the four-stage design to operate at the same frequency as the three-stage design, its load resistance must be roughly 60% of the three-stage value. For this load resistance value, the four-stage design exhibits roughly the same additive thermal noise and open-loop Q , yet the power dissipation is a factor of 2 higher. The advantage of the four-stage design is that phases of 0° , 90° , 180° , and 270° are extremely useful in “quadrature” circuits that are prevalent in communication systems. This power dissipation is critical for low-power VCOs, and may be elaborated on by noting that for the three-stage design when $G_M R \sim 2$, the total additive voltage noise at the output may be expressed as

$$|V_{nTOT}|^2 = 8kT \cdot \frac{2}{9G_M} \left(\frac{\omega_0}{\Delta\omega}\right)^2 \tag{9.80}$$

To minimize this phase noise, the G_M must be increased by increasing the device size and bias current. For a constant supply voltage, the power dissipation will increase by the same amount that the phase noise is decreased. The tradeoff between power consumption and phase noise in all oscillators is significant, but most

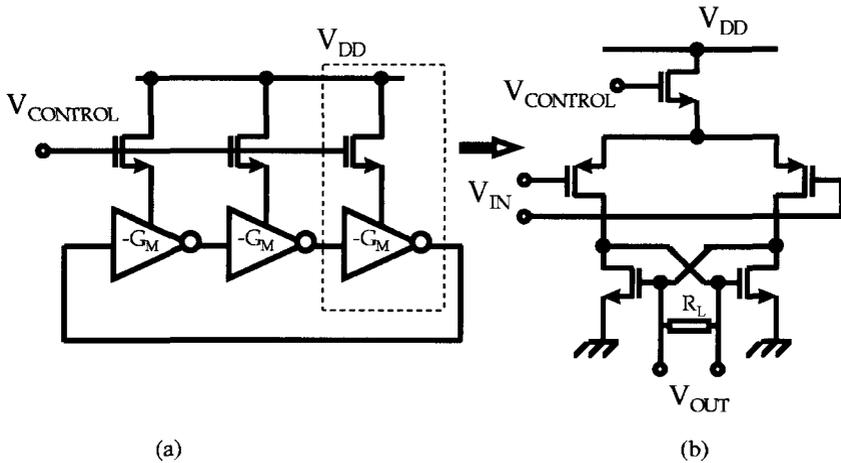


Figure 9.42 Ring oscillator topology with (a) three-stages utilizing power supply noise suppression and (b) the individual relaxation oscillator making up each stage of the ring.

restrictive for ring oscillators where low values of open-loop Q and relatively high phase noise force higher power consumption to achieve performance.

A design aspect that is important to prevent additional degradation in noise performance is the rejection of noise from the power supply. An individual inverter stage²⁴ is shown in Figure 9.42a along with a bias circuit that isolates bias supply noise from the oscillator.

A second type of VCO topology shown with this power supply noise rejection approach is shown in Figure 9.42b. This “relaxation” oscillator topology relies on the switched charging and discharging of a loading capacitance. Combined with the device currents and speed, the capacitance value determines the oscillation period. Because the reactive element, C , is not frequency-selective, this oscillator topology relies on the time delay of charging time and fall time to provide feedback and as such is susceptible to jitter in the discrete times at which switching thresholds are reached.

The final topology, shown in Figure 9.43, is the most prevalent of the high-performance CMOS oscillators and is based on a narrowband resonant load consisting of an LC parallel tank. The LC oscillators exhibit lower phase noise at a given power level than the previous two topologies, but typically suffer from a smaller tuning range due to the narrowband resonant load.

The inductors in this circuit shown in Figure 9.43 are made to resonate with the varactor diode and device capacitance to produce a large impedance at the oscillation frequency. The differential-mode, equivalent circuit for this topology is shown in Figure 9.44.

Where g_D is the device output conductance, R_l is the series resistance of the inductor, and L_1 is the inductance value. We define $C_{G1} = C_{GS1} + C_{D2} + C_{J2}$, and $C_{G2} = C_{GS2} + C_{D1} + C_{J1}$, where C_{GSn} is the gate-source capacitance of device n , C_{Dn} is the reverse-bias diode junction tunable capacitance of diode n , and C_{Jn} is the

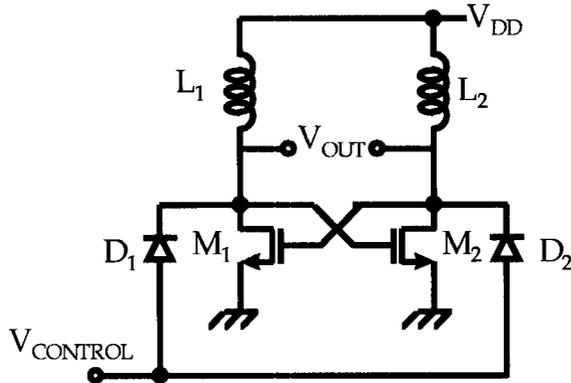


Figure 9.43 LC voltage-controlled oscillator utilizing a pair of resonant loads and diode tuning.

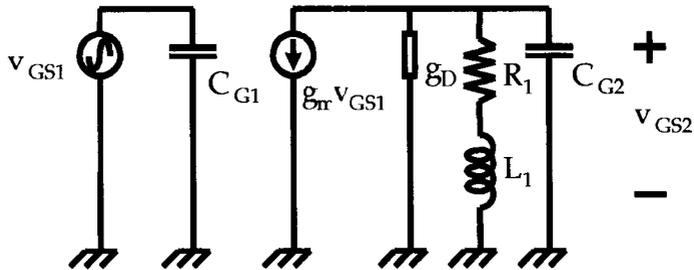


Figure 9.44 Differential mode equivalent circuit for the LC VCO of Figure 9.41.

output drain junction capacitance of device n . The resonant frequency is defined at the resonance of the LC tank circuit according to

$$f_0 = \frac{1}{\sqrt{L_1 C_{G2}}} \cdot \sqrt{1 - \frac{R_l^2 C_{G2}}{L_1}} \quad (9.81)$$

The loop gain for this circuit may be derived as

$$A_{\text{LOOP}} = G_M \frac{(R_l + sL)}{1 + sR_l C_{G2} + s^2 L C_{G2}} \quad (9.82)$$

To satisfy the Barkhausen criteria for oscillation, the loop gain at resonance must be greater than one and then must self-limit at unity gain. Therefore, to oscillate, the minimum transconductance of the transistor must be greater than the load conductance at resonance:

$$g_m > \frac{1}{r_n} = \frac{R_l}{R_l^2 + \omega_0^2 L^2} \approx \frac{R_l}{\omega_0^2 L^2} \approx \frac{1}{Q \cdot \omega_0 L} \quad (9.83)$$

This is readily achievable for CMOS-based oscillators in this configuration, and actually allows this topology to be much more efficient up to high frequencies than the Pierce, Colpitts, and Clapp oscillators for the same transistor size. An important note is that as the inductor losses get larger and R_i increases, the inductor real impedance at resonance can compensate for this effect and lower the requirement for device gain to start the oscillation. The challenge here is to use large inductors with small resistance to relax this constraint. This design issue is discussed in more detail in the section to follow on integrated inductors. As a design example of the LC oscillator, the schematic in Figure 9.45 exploits the use of inductor design to achieve lower phase noise.²⁵

The oscillation frequency is 1.8 GHz and is achieved in a 0.7- μm CMOS process. The power supply voltage is only 1.5 V and the VCO consumes only 4 mA for a mere 6 mW in dissipated power. The resulting phase noise shown in Figure 9.46 (and defined as the phase noise power to carrier ratio in decibels in a 1-Hz bandwidth) is only -116 dBc/Hz at a 200-kHz offset and represents the state of the art for fully integrated CMOS VCOs. The circuit also exhibits a 14% tuning range of over 250 MHz and an output power of -20 dBm.

To fully integrate transceiver functions, one requirement is the generation of quadrature outputs from the oscillator, or signals separated by 90° in phase, primarily to feed quadrature upconversion or downconversion mixers as discussed previously. One method is the close coupling of two separate oscillators shown in Figure 9.47a.²⁶

The direct coupling of oscillator B outputs to oscillator A inputs produces a current through the load $i_A = g_m(v_B + v_A)$. Similarly, the outputs of oscillator A cross-coupled to the inputs of oscillator B result in a current of $i_B = g_m(v_B - v_A)$. The vector diagram of these circuits is shown in Figure 9.47b and reveals that for

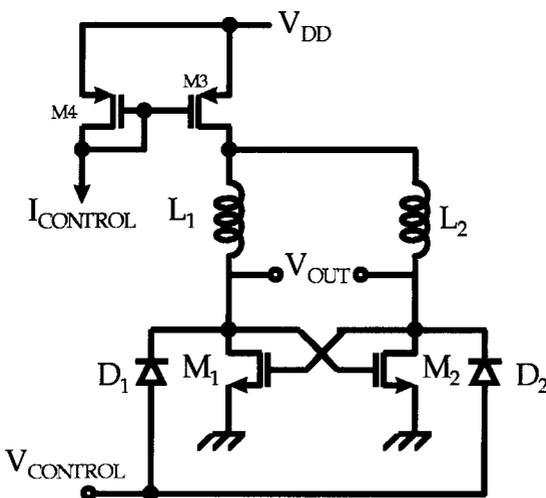


Figure 9.45 VCO topology utilizing LC resonant loads power supply noise isolation, and current-controlled tuning.

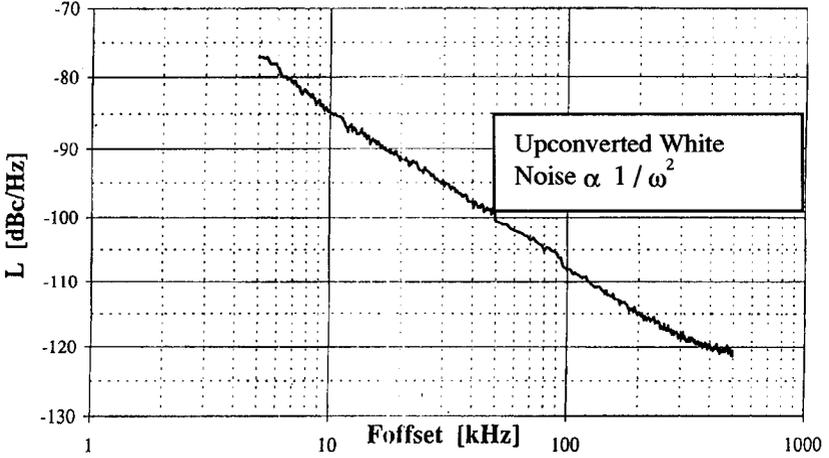


Figure 9.46 Measured phase noise of the enhanced LC-tank CMOS VCO of Figure 9.45.

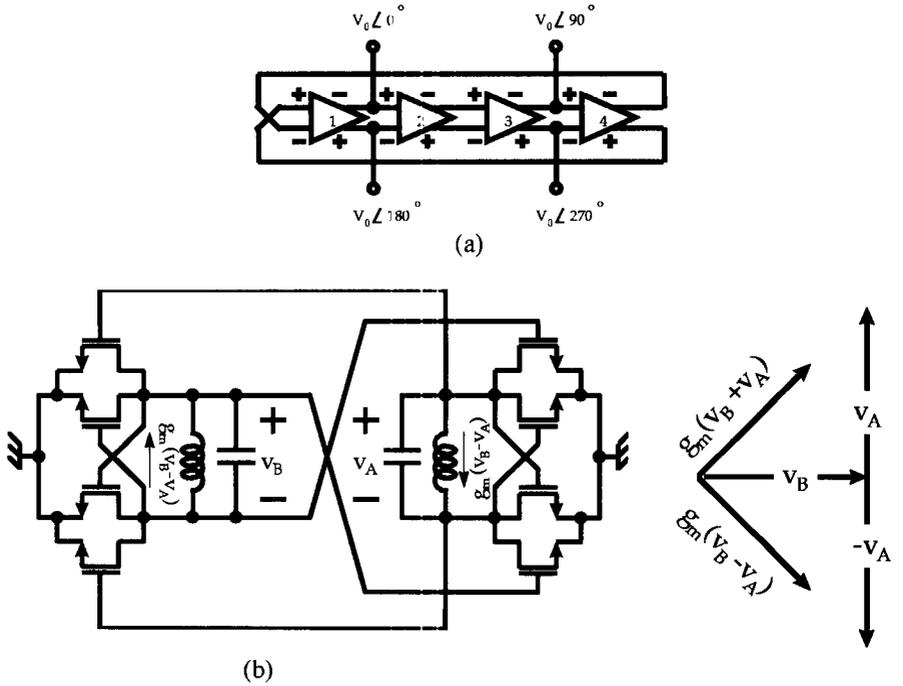


Figure 9.47 The use of coupled oscillators to generate quadrature phase outputs: (a) general schematic of four-stage polyphase VCO and (b) detail of oscillator showing the requirement for quadrature outputs.

equal amplitudes in these symmetrical oscillators ($v_A = v_B$), the resulting phase relationship between the two differential outputs will be 90° . An example of the use of this approach is shown in the circuit of Figure 9.47b at an oscillation frequency of 850 MHz.^{27,28} The resulting phase noise characteristic is $L(f) \cong -110$ dBc/Hz at a 200-kHz offset for a 1- μm CMOS process device. The integrated inductors are each 50 nH spirals forward in the process' second-level metal. Spiral postprocessing was used to remove the substrate selectively from underneath the inductor to lower the parasitic capacitance of such a large coil. This enables a much larger self-resonant frequency and usable frequency range for the inductor. The device consumes 30 mW from a 3-V supply for an output power of -25 dBm. The tuning range is achieved via the upper PMOS current source for a total of 13% of the 900-MHz oscillation frequency, or about 120 MHz. The importance of this device lies in the impressive results for fully integrated solutions with less than state-of-the-art technologies and holds significant promise for smaller gate lengths in future implementations. Table 9.6 summarizes progress made recently toward lower power and lower phase noise in fully integrated VCOs. The bipolar and BiCMOS examples are given as a reference to the best CMOS results achieved to date.

An example of the full implementation of these VCO designs and concepts into full synthesizers is described next. In practical communication transceivers, the basic VCO must be very accurately controlled with a phase and frequency feedback loop to lock it to a specific frequency. This feedback is implemented in a "phase-locked loop" (PLL), and is shown in general form in Figure 9.48a along with its functional representation in Figure 9.48b.

The output frequency f_{OUT} is locked to an integer function of its inputs and adjusted according to the mixing and divide ratios of the feedback loop according to

$$\omega_{\text{OUT}} = \omega_M + N \cdot \omega_{\text{REF}} \quad (9.84)$$

The classic tradeoffs in the loop are the frequency resolution, output phase noise, switching and settling times, and power dissipation. The phase-noise of the VCO dominates outside the PLL bandwidth but within the bandwidth it is attenuated by an amount equal to the division ratio. The disadvantages of a large loop bandwidth are high levels of spurious tones that are caused by modulation of the VCO with the harmonic content output of the phase detector.

Figure 9.49a shows an example of a complementary clock-driven architecture for the CMOS divider circuits of a phase-locked loop. This particular design uses complementary clocks that allow only the faster NMOS devices to be used in the signal path, thus avoiding the slower PMOS circuits required if only a single clock were used. Further detail of this NFET-based D-Q flip-flop-based divider circuit is provided in Figure 9.49b. These components form the basis for the divider-based architecture of Figure 9.48a making up a CMOS highly integrated synthesizer.³⁹ The implementation demonstrated is in a partially scaled, 0.1- μm CMOS technology for a divide-by-2 and PLL circuit.³⁹ The measured speed of the proposed frequency divide-by-2 circuits is shown in Figure 9.50, demonstrating clean division up to 13 GHz for a power consumption of 28 mW from a 2.6 V supply.

TABLE 9.6 Comparison of Si-Based Monolithic VCOs

f_0 (GHz)	Power (mW)	Phase Noise (dBc/Hz)	P_{OUT} (dBm)	Tuning (MHz)	Area (mm ²)	Si-Based Process Technology	Reference
0.90	10	-101 dBc/Hz at 100 kHz	-3	N/A	N/A	25-GHz bipolar	29
2.20	43.2	-106 dBc/Hz at 200 kHz	-3	264 MHz/12%	0.96	15-GHz bipolar	30
1.50	28	-109 dBc/Hz at 200 kHz	-6.6	150 MHz/10%	0.50	0.8- μ m BiCMOS	31
1.80	70	-88 dBc/Hz at 100 kHz	-25	200 MHz/11%	0.2	10-GHz BiCMOS	32
2.00	3	-74 dBc/Hz at 100 kHz	-25	N/A	N/A	20-GHz BiCMOS	33
2.40	54	-92 dBc/Hz at 1000 kHz	-13.5	N/A	0.50	12-GHz BiCMOS	34
4.10	14	-98 dBc/Hz at 200 kHz	-5	328 MHz/8%	0.60	0.5- μ m BiCMOS/5 Metal: CMOS only in oscillator	35
0.85	30	-84 dBc/Hz at 200 kHz	-25	120 MHz/14%	5.00	1- μ m CMOS/quadrature etched substrate	36
1.80	7.6	-92 dBc/Hz at 200 kHz	-25	126 MHz/7%	N/A	0.6- μ m CMOS/ quadrature	37
1.80	24	-115 dBc/Hz at 200 kHz	-35	80 MHz/4.5%	5.18	0.7- μ m CMOS/bondwire inductor	38

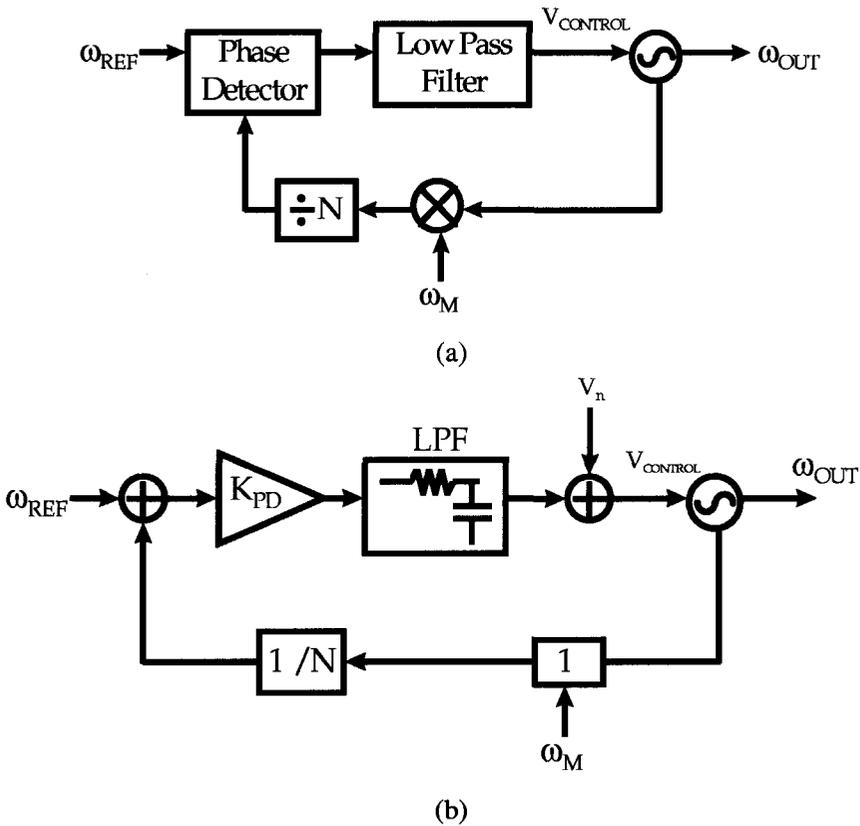


Figure 9.48 Architecture of a typical phase-locked loop showing (a) general schematic and (b) system block detail including noise contributions.

An alternative CMOS-based PLL architecture for higher-speed output frequencies is shown in Figure 9.51a. The PLL is driven by an input buffer from a 50-Ω input impedance that balances symmetrically with the frequency, power level, and impedance presented by the VCO output. The mixer takes these two inputs and downconverts to a low-frequency signal that is filtered, current amplified and used to drive the current-controlled oscillator (VCO). The current controlled oscillator is single-ended to allow headroom within the constraint of the <3-V supply voltage that would not be available with a comparable differential design, and although sensitivity to substrate and supply noise is an issue, its current control biasing is intended to mitigate these issues as shown in Figure 9.51b.

This maximum speed in the PLL is limited by the maximum frequency of the current-controlled oscillator because the increasing control voltage forces more and more triode region operation of the PMOS devices killing the gain of each stage. The phase noise is approximately -100 dBc/Hz at a 40-kHz offset for operation at

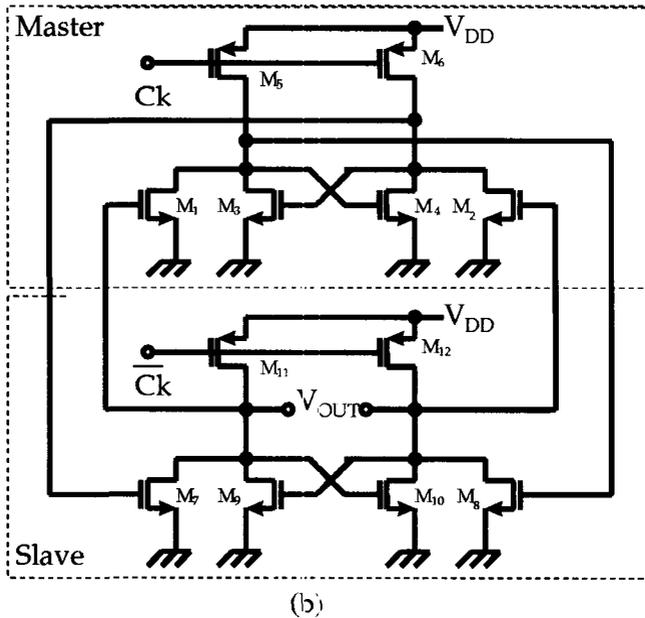
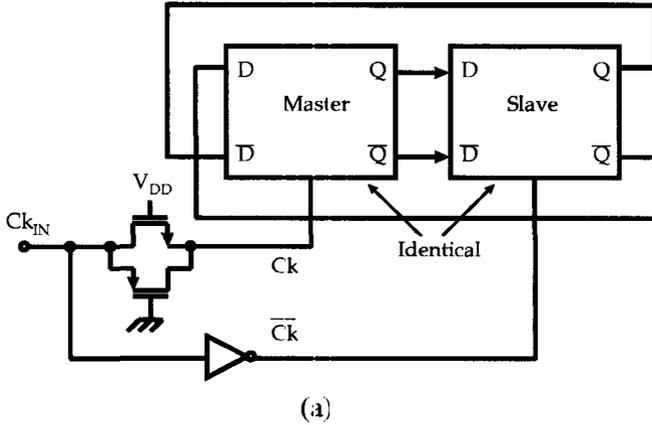


Figure 9.49 (a) Master–slave divider architecture with complementary clocks and (b) the detailed schematic for the divider circuit.

3 GHz and is shown in Figure 9.52. The active area of the PLL is a mere $60 \times 100 \mu\text{m}$, excluding the low-pass-filter loop capacitor, and can achieve an output frequency of 3.3 GHz with a capture range of $\pm 300 \text{ MHz}$. The extremely high-frequency operation of these circuits is attributable to excellent design and the enhanced capability of deep-submicrometer CMOS technologies. As gate lengths approach $0.18 \mu\text{m}$ in the standard process, the use of CMOS in RF transceiver systems will become widespread.

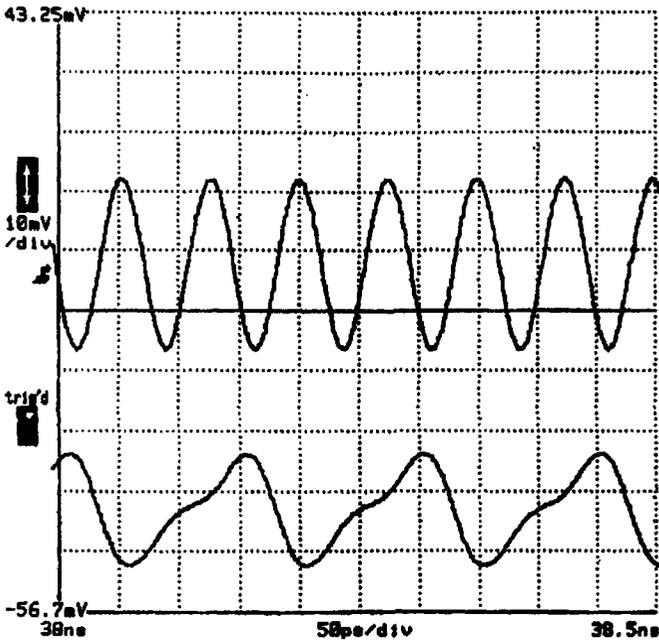


Figure 9.50 Measured speed of the divide-by-2 circuit of Figure 9.47b.

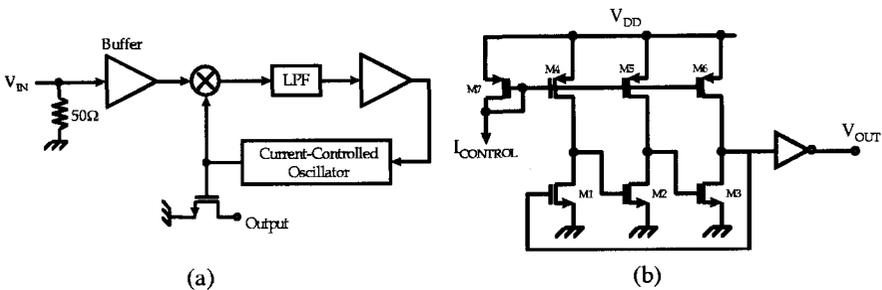


Figure 9.51 (a) Phase locked-loop architecture and (b) detailed schematic for the current-controlled oscillator. (After Razavi et al., Ref. 39.)

9.7 CMOS POWER AMPLIFIERS

Power amplifiers are a significant challenge in integrated transceiver design, especially for CMOS technologies. The requirements for linearity and efficiency are very demanding because telecommunication equipment requires roughly 1 W output power at 1–2 GHz. The class of operation falls into linear (A, AB) and nonlinear (B, C, E, F) groups, and each describes different amounts of waveform “clipping” in the device current that causes the output waveform to contain increased harmonic content.

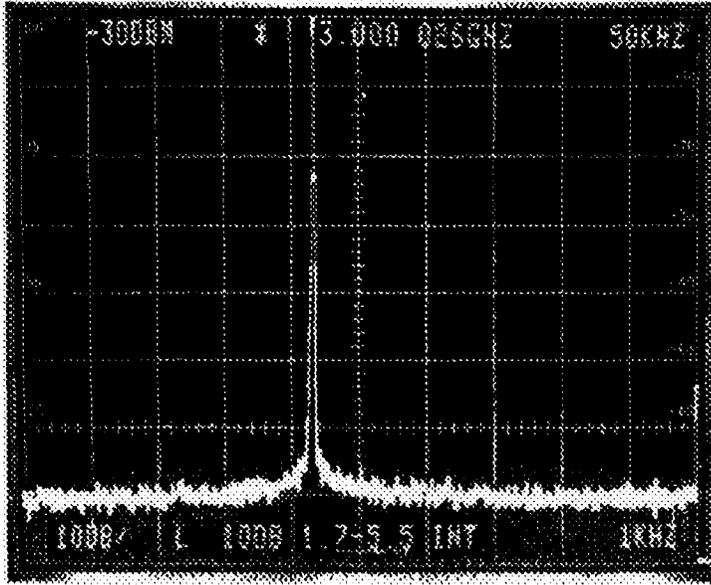


Figure 9.52 Measured spectrum of the phase-locked loop of Figure 9.50 at 3 GHz.

9.7.1 Power Amplifier Fundamentals

All power amplifiers can be seen as taking some RF signal and converting dc power into RF amplification that produces significantly more output power than is present at the input. The power-added efficiency, therefore, is a critical figure of merit that gauges the “conversion” efficiency of dc power into RF amplification and is defined as:

$$\eta_{\text{PAE}} = \frac{P_{\text{RF}}^{\text{OUT}} - P_{\text{RF}}^{\text{IN}}}{P_{\text{DC}}} \quad (9.85)$$

where $P_{\text{RF}}^{\text{OUT}}$ is the RF output power, $P_{\text{RF}}^{\text{IN}}$ is the RF input power, and P_{DC} is the dc power dissipated by the device. The power-added efficiency implicitly indicates whether the amplifier has reasonable gain, in that the numerator of the expression is the difference between the output and input power, and will have a small value for devices with little or no gain. The basic single-ended power amplifier schematic shown in Figure 9.53 consists of a large active device, an inductive load, a reactive matching network, and the load resistor that is the intended target for the amplified signal. The inductive load is applied as a large-choke inductance that allows the output voltage to swing $\pm V_{\text{DD}}$ about the drain dc voltage (V_{DD}) so that the transistor output swings from ground to twice the supply voltage ($2V_{\text{DD}}$). Therefore, the output power is increased by a factor of 2 and the power-added efficiency is increased from a theoretical maximum of 25 to 50% for the class A amplifier we will discuss.

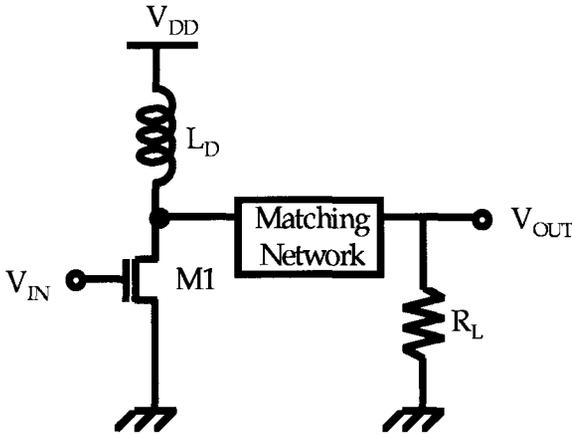


Figure 9.53 Schematic of a single-ended, common source, power amplifier.

The class A stage is the most linear of the power amplifier classes, and is biased to prevent any clipping of the signal and its resulting distortion. The bias of the transistor is shown in Figure 9.54a along with its resulting waveforms under peak operation conditions. The curve exhibits a significant current and voltage quiescent value during operation (V_Q, I_Q) with a large dc current consumption.

For the case of such a high quiescent power dissipation, the efficiency is relatively low and is easily calculated from the RF and dc values displayed for peak operation as shown in Figure 9.54b. Assuming the RF gain to be least 20 dB, the difference

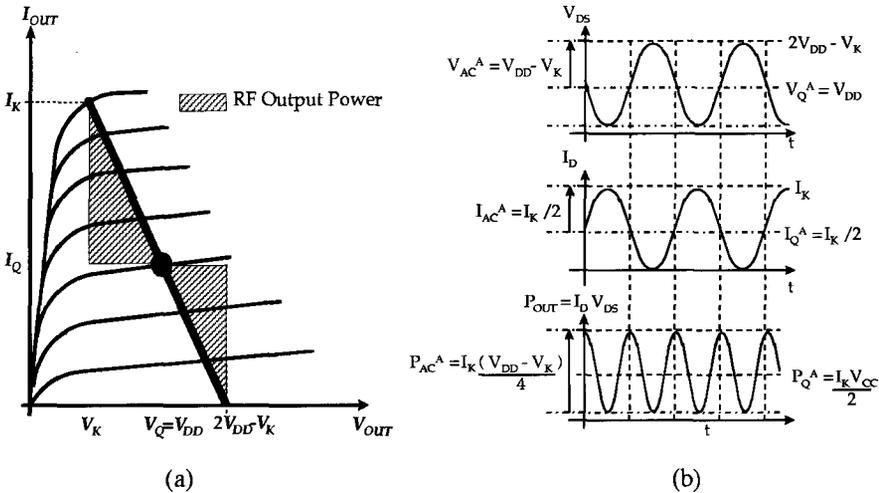


Figure 9.54 Class A operation of a power amplifier showing (a) the interaction of load line and dc device characteristics determining η_{PAE} , P_{OUT} , and (b) detailed current and voltage waveforms for class A operation.

between the output power and input power can be calculated as

$$P_{\text{RF}}^{\text{OUT}} - P_{\text{RF}}^{\text{IN}} = \left(1 - \frac{1}{100}\right) \cdot P_{\text{RF}}^{\text{OUT}} \approx P_{\text{RF}}^{\text{OUT}} = \frac{I_K(V_{DD} - V_K)}{2} \quad (9.86)$$

and the resulting power added efficiency may be expressed as

$$\eta_{\text{PAE,A}} = \frac{1}{2} \cdot \frac{(V_{DD} - V_K)}{V_{DD}} \quad (9.87)$$

This result indicates the importance of minimizing the knee voltage at which peak currents occur, and below which the MOSFETs enter the triode region. Significant ON resistance is one contributor to finite slopes of the $I_D - V_{DS}$ characteristic in this triode region, causing higher V_K values. Although the ultimate theoretical limit for efficiency in class A is 50%, for a 3-V supply, every 150 mV added to V_K will induce another 5% decrease in efficiency, and typically efficiencies of 40–45% are achievable. Further loss in efficiency can be due to matching-network losses and inductances to ground that degrades the transistor.

The result for the RF output power can be seen geometrically, as one-half the shaded triangle in the $I-V$ characteristic of Figure 9.54a. This is the integral of the current swing about the quiescent dc operating point taken over the output voltage sweep, and is simply the area between the load line and the dc operating current between the extremes of the voltage swing. This total shaded area is divided by 2 as the average power is the product of rms voltage and current. One aspect that is made clear about class A is that for linear operation, before any clipping occurs, as much RF power is output at voltages above the dc operating point as below it, and the dc operating point is, therefore, forced to consist of a large constant current.

The class B amplifier modifies this approach by biasing at a zero dc drain current and pumping output power only at voltages below that quiescent dc operating voltage. Class B is defined such that for exactly one half of the cycle, the output drain current is “clipped” and the transistor turns off during that interval. As shown in Figure 9.55, the current swings from the peak “knee” current at I_K down to zero where the device is cutoff, while the voltage swings from V_K up to $2V_{DD} - V_K$.

The time interval over which current flows is termed the *conduction angle*, and is 180° for class B (360° for class A). The theoretical maximum efficiency in class B is higher than in class A, and can be derived from the RF output power and dc power dissipation as shown in Figure 9.55b

$$\eta_{\text{PAE,B}} = \frac{P_{\text{RF}}^{\text{OUT}} - P_{\text{RF}}^{\text{IN}}}{P_{\text{DC}}} \approx \frac{P_{\text{RF}}^{\text{OUT}}}{P_{\text{DC}}} = \frac{\pi}{4} \cdot \frac{(V_{DD} - V_K)}{V_{DD}} \quad (9.88)$$

The theoretical maximum efficiency is $(\pi/4) \approx 78.5\%$, but is reduced by the constraint of the knee voltage V_K . Typical efficiencies achieved in class B are 65–70%. To further compare class A and class B, the ratio of maximum load power, $P_{L,\text{MAX}}$ to the maximum device dissipation, illustrates the required capability of the

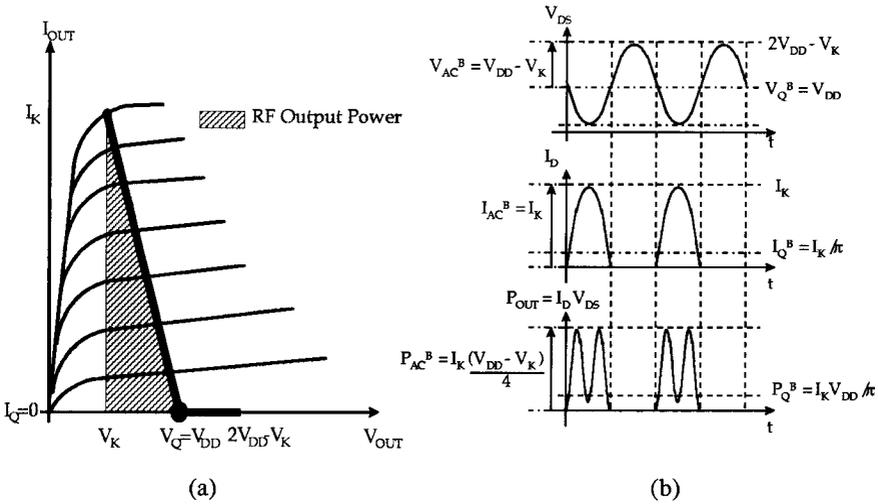


Figure 9.55 Class B operation of a power amplifier showing (a) the interaction of load line and dc device characteristics determining η_{PAE} , P_{OUT} , and (b) detailed current and voltage waveforms for class B.

device to dissipate power for a desired RF output power:

$$\left. \frac{P_{L,MAX}}{P_{DISS}} \right|_{\text{class A}} = \frac{1}{2} \cdot \left(\frac{V_{DD} - V_K}{V_{DD}} \right) \tag{9.89}$$

$$\left. \frac{P_{L,MAX}}{P_{DISS}} \right|_{\text{class B}} = \frac{\pi}{4} \cdot \left(\frac{V_{DD} - V_K}{V_{DD}} \right) \tag{9.90}$$

For a 1-W power amplifier with a 150-mV knee voltage to run from a 3-V supply and class A operation device must be able to dissipate a total of 2.1 W, while the class B amplifier requires a total dissipation of only 1.34 W. Generally, class B amplifiers dissipate almost 50% less power than class A, and the main disadvantage is the worse linearity of class B. In between these extremes lies a region where these tradeoffs can be balanced between classes A and B, called class AB, where the drain current is clipped in cutoff but for less than half the entire cycle. This compromise is an attractive one for linear amplifiers that require greater linearity than operation in full class B, and higher efficiency than achievable in class A.

Class C is the last class of operation and occurs for the drain current clipped and equal to zero for more than half the entire cycle. The theoretical limit for efficiency is 100%, but as the short bursts of drain current narrow and efficiency approaches this maximum, the output power decreases toward zero. This is shown in Figure 9.56, which summarizes all the power amplifier classes of operation discussed in terms of efficiency and maximum output power as a function of the conduction angle⁴⁰ and illustrates that class C efficiencies typically come at the price of very little output power.

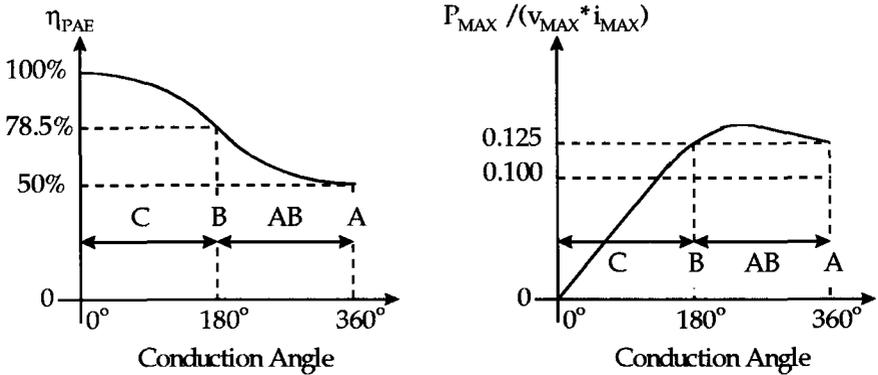


Figure 9.56 The (a) power-added efficiency, η_{PAE} , and (b) output power, P_{OUT} , for power amplifiers as a function of conduction angle. Continuous improvement in η_{PAE} is seen as one progresses from class A to 100% efficient class C, but there is also a corresponding decrease in P_{OUT} to zero.

9.7.2 CMOS Power Amplifiers for Highly Integrated Transceivers

Having discussed the issues endemic to all power amplifiers, we see that the specific challenges for CMOS in power amplifier applications lie in the tradeoff between f_T and breakdown voltage. This has been addressed best in the lateral diffusion MOS (LDMOS) processes. Although they are not specifically integrable with standard digital CMOS for ultra-large-scale integration, they illustrate the potential of the MOS structure in PA applications and illuminate some of the design tradeoff required for CMOS if it is to excel in this application. LDMOS basically involves the lateral diffusion of p^- on the source side to underneath the gate to increase the transconductance, and the lateral diffusion from under the gate, out toward the drain of a drift region for reduced ON resistance and higher breakdown voltage. For the example⁴¹ discussed here, the gate length of 0.6 μm was patterned over a 400-Å-gate oxide, with planarization steps to maximize field-oxide thickness and minimize capacitance. Operating from a single power supply, the devices were tested using a large-signal test called “load pull,” which tunes the impedance at the load under large-signal conditions to evaluate the ultimate power amplification and efficiency of the device under matched conditions. Figure 9.57 shows the results at 850 MHz for a 12-dB gain and a 75 percent power-added efficiency with a bias from a 3.4-V supply.

This 22-mm-wide device demonstrated a peak output power of almost 29 dBm, for approximately 36 mW/mm. This makes it clear that modified CMOS processes have tremendous capabilities to address the PA markets with low-cost and high-performance solutions with more integrated solutions soon to come.

9.8 SUMMARY AND FUTURE TRENDS

As the CMOS technology continues to evolve for achieving higher f_T and f_{max} and obtaining high- Q reactive components (capacitors and inductors), it will be possible

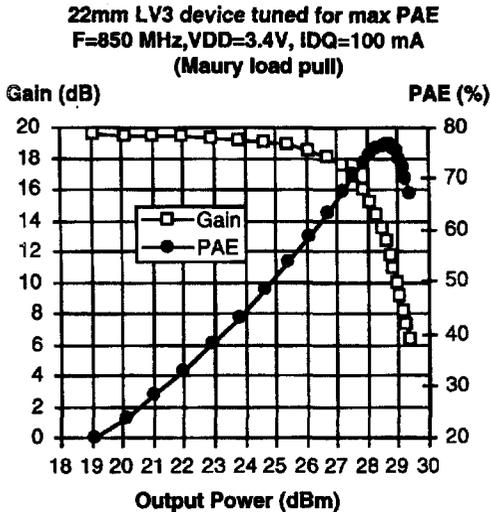


Figure 9.57 A $0.6\ \mu\text{m}$ LDMOS power amplifier operating at 850 MHz with a 12 dB gain and a 75% power-added efficiency.

in the near future to realize mixed-signal systems, such as a radio-on-a-chip, based on a mainstream, submicrometer CMOS technology. Future CMOS technology is anticipated to offer not only very high-speed digital functions but also high-performance analog and RF functions in a monolithic form. In this chapter, we have reviewed and summarized critical issues in developing such CMOS technology and its state of the art for multifunction mixed-signal applications.

ACKNOWLEDGMENT

The authors are grateful to many colleagues at the Rockwell Science Center for numerous discussions of the issues presented here. Among others, we especially thank Dr. Wei-Heng Chang for contributing the section of digital CMOS logic circuits and Dr. M. C. Vincent Ho for his work in CMOS power amplifier and summary of the state of the art of the CMOS technology. We also acknowledge the support of Dr. Jon Rode and Dr. Derek Cheung throughout this work.

REFERENCES

1. Y. Mii, S. Rishton, Y. Taur, D. Kern, T. Lii, K. Lee, K. A. Jenkins, D. Quinlan, T. Brown Jr., D. Danner, F. Sewell, and M. Polcari, "Experimental High Performance Sub- $0.1\ \mu\text{m}$ Channel MOSFETs," *IEEE Electron Device Lett.* **15**(1), 28 (1994).
2. P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3 ed., Wiley, New York, 1993, p. 74.

3. J. M. Rabaey, in C. G. Sodini, ed., *Digital Integrated Circuits*, Prentice-Hall Electronics and VLSI Series, 1996.
4. D. J. Allstot and W. C. Black, "A Substrate Referenced Data Conversion Architecture," *IEEE Trans. Circ. Syst.* **28**, 1212 (1991).
5. B. J. Hostica and W. Brockherde, "The Art of Analog Circuit Design in a Digital VLSI World," *Proc. IEEE Int. Symp. Circuits and Systems*, 1990, p. 1347.
6. M. Maleki, S. Kiaei, and D. Allstot, "Synthesis Techniques for CMOS Folded Source-Coupled Logic Circuits," *IEEE J. Solid-State Circ.* **27**(6) (1992).
7. M. Mizuno et al., "A GHz MOS Adaptive Pipeline Technique Using MOS Current-Mode Logic," *IEEE J. Solid State Circ.* **31**(5) (1996).
8. S. Kiaei, S. H. Chee, and D. J. Allstot, "CMOS Logic Circuits for Mixed Mode VLSI," *Proc. IEEE Int. Symp. Circuits and Systems*, 1990, p. 1608.
9. S. Maskai and S. Kiaei, *Synthesis of Complex Source-Coupled Logic Devices*, Tech. Report, Electrical and Computer Engineering Dept., Oregon State Univ., 1991.
10. D. J. Allstot, S.-H. Chee, S. Kiaei, and M. Shrivastawa, "Folded Source-Coupled Logic vs. CMOS static Logic for Low Noise Mixed-Signal Ics," *IEEE Trans. Circ. Syst.* **40**(9) (1993).
11. L. C. M. G. Pfenning, W. G. J. Mol, J. J. J. Bastiaens, and J. M. F. Van Dijk, "Differential Split-Level CMOS Logic for Subnanosecond Speeds," *IEEE J. Solid-State Circ.* **SC-20**(5), 1050 (1985).
12. K. M. Chu and D. L. Pulfrey, "A Comparison of CMOS Circuit Techniques: dcVSL Versus Conventional Logic," *IEEE J. Solid State Circ.* **SC-22**(4) (1987).
13. S. A. Maas, *Nonlinear Microwave Circuits*, Artech House, 1988, p. 160.
14. B. Razavi and A. A. Abidi, "Integrated Circuit Design for Wireless Transceivers," short course, UCLA, Sept. 8–10, Los Angeles, 1997.
15. D. K. Shaeffer and T. H. Lee, "A 1.5 V 1.5 GHz CMOS Low Noise Amplifier," *IEEE Symp. VLSI Circuits Digest of Technical Papers*, 1996, p. 32.
16. J. C. Rudell et al., "A 1.9 GHz Wideband IF Double Conversion CMOS Integrated Receiver for Cordless Telephone Applications," *ISSCC Digest Tech. Papers*, 1997, p. 304.
17. A. N. Karanicolas, "A 2.7 V 900 MHz CMOS LNA and Mixer," *IEEE J. Solid-State Circ.* **31**(12), 1939 (1996).
18. A. Rofougaran, J. Y. C. Chang, M. Rofougaran, and A. A. Abidi, "A 1 GHz CMOS rf Front-End IC for a Direct-Conversion Wireless Receiver," *IEEE J. Solid-State Circ.* **31**(7), 880 (1996).
19. R. Goyal, *High-Frequency Analog Integrated Circuit Design*, Wiley, New York, 1995, p. 269.
20. J. Crols and M. S. J. Steyaert, "A 1.5 GHz CMOS Highly Linear CMOS Downconversion Mixer," *IEEE J. Solid-State Circ.* **30**(7), 736 (1995).
21. A. Riddle, *Oscillator Noise: Theory and Characterization*, Ph.D. dissertation, Dept. Electrical Engineering, North Carolina State Univ., 1986, pp. 110–138.
22. D. B. Leeson, "A Simple Model of Feedback Oscillator Noise Spectrum," *Proc. IEEE*, 329 (Feb. 1966).
23. B. Razavi, "A Study of Phase-noise in CMOS Oscillators," *IEEE J. Solid-State Circ.* **31**(3), 331 (1996).

24. Z.-X. Zhang, H. Du, and M. S. Lee, "A 360 MHz, 1.5 mW at 1.35 V CMOS PLL with 1 V Peak-to-peak Power-supply Noise Tolerance," *ISSCC Digest Tech. Papers*, 1996, p. 134.
25. J. Craninckx and M. S. J. Steyaert, "A 1.8 GHz CMOS Low-Phase-Noise Voltage-Controlled Oscillator with Prescaler," *IEEE J. Solid-State Circ.* **30**(12), 1474 (1995).
26. A. W. Buchwald and K. W. Martin, "High-Speed Voltage Controlled Oscillator with Quadrature Outputs," *Electron. Lett.* **27**(4), 309 (1991).
27. A. A. Abidi, "Low-Power Radio-Frequency IC's for Portable Communications," *Proc. IEEE*, **83**(4), 544 (1995).
28. A. A. Abidi, "Direct-Conversion Radio Transceivers For Digital Communications," *IEEE J. Solid-State Circ.* **30**, 1399 (1995).
29. A. Ali and J. Tham, "A 900 MHz Frequency Synthesizer with Integrated LC Voltage-Controlled Oscillator," *ISSCC Digest Tech. Papers*, 1996, p. 392.
30. B. Jansen, K. Negus, and D. Lee, "Silicon Bipolar VCO Family for 1.1 GHz to 2.2 GHz with Fully Integrated Tank and Tuning Circuits," *ISSCC Digest Tech. Papers*, 1997, p. 392.
31. L. Dauphinee, M. Copeland, and P. Schvan, "A Balanced 1.5 GHz Voltage Controlled Oscillator with an Integrated LC Resonator," *ISSCC Digest Tech. Papers*, 1997, p. 390.
32. N. Nguyen and R. G. Meyer, "A 1.8 GHz Monolithic LC Voltage-Controlled Oscillator," *IEEE J. Solid-State Circ.* **27**(3), 444 (1992).
33. T. Aytur and B. Razavi, "A 2 GHz 6 mW BiCMOS Frequency Synthesizer," *ISSCC Digest Tech. Papers*, 1995, p. 264.
34. M. Soyeur, K. A. Jenkins, N. Burghartz, and M. D. Hulvey, "A 3 V 4 GHz nMOS Voltage-Controlled Oscillator with Integrated Resonator," *IEEE J. Solid-State Circ.* **31**(12), 2042 (1996).
35. M. Soyeur, K. A. Jenkins, J. N. Burghartz, and M. D. Hulvey, "A 3 V 4 GHz nMOS Voltage-Controlled Oscillator with Integrated Resonator," *ISSCC Digest Tech. Papers*, 1996, p. 394.
36. A. Rofougaran, J. Rael, M. Rofougaran, and A. A. Abidi, "A 900 MHz CMOS LC-Oscillator with Quadrature Outputs," *ISSCC Digest Tech. Papers*, 1996, p. 390.
37. B. Razavi, "A 1.8 GHz CMOS Voltage Controlled Oscillator," *ISSCC Digest Tech. Papers*, 1997, p. 388.
38. J. Craninckx and M. Steyaert, "A CMOS 1.8 GHz Low-Phase-Noise Voltage-Controlled Oscillator with Prescaler," *ISSCC Digest Tech. Papers*, 1995, p. 266.
39. B. Razavi, K. F. Lee, and R. H. Yan, "Design of High-Speed, Low-Power Frequency Dividers and Phase-Locked Loops in Deep Submicron CMOS," *IEEE J. Solid-State Circ.* **30**(2), 101 (1995).
40. H. L. Krauss, *Solid-State Radio Engineering*, Wiley, New York, 1980, p. 348.
41. G. Ma, W. Burger, X. Ren, J. Gibson, and M. Shields, "High Efficiency Submicron Gate LDMOS Power FET for Low Voltage Wireless Communications," *IEEE MTT-S Digest*, 1997, p. 1303.

PROBLEMS

- 9.1** A CMOS technology is developed over time to reduced gate lengths, each version implemented in a circuit containing 20,000 transistors. At any given

time, half of the transistors are switching at the full clock rate, and the other half are turned off. Assume that the direct-path conduction time over which both the n- and p-type transistors conduct is equal to 0.2 ns, and that each device has width of 5 μm . Use the information in the table below for each technology version to calculate the static, dynamic, and total power dissipation for the circuit given each of the following clock rates: 1 MHz, 10 MHz, 100 MHz, and 1 GHz.

L_G (μm)	T_{ox} (Å)	C_L (fF)	I_{peak} ($\mu\text{A}/\mu\text{m}$)	I_{off} ($\mu\text{A}/\mu\text{m}$)	V_{ss} (V)
0.8	150	37	230	0.2	5
0.6	120	38	250	2	5
0.5	90	38	260	2	3.3
0.35	65	34	270	2	3.3
0.25	45	30	280	2	2.5

- 9.2 A single-stage low-noise amplifier operates at 900 MHz over a bandwidth of 20 MHz and exhibits $\text{SNR}_{\text{INPUT}} = 2.5$, $\text{SNR}_{\text{OUTPUT}} = 1.4$, $\text{SNR}_{\text{MIN}} = 3$ db, gain = 20 dB, $\text{IMD}_3 = -20$ dBm at $P_{\text{IN}} = -20$ dBm and $P_{1\text{dB}} = -5$ dBm. For this amplifier calculate (a) noise figure, (b) IIP_3 , (c) SFDR, and (d) BDR.
- 9.3 The design of low-noise amplifiers requires treatment of both power and noise matching. Given that the single transistor common-source LNA circuit topology in Figure 9P.1a is designed for extremely low-power operation at 900 MHz:
- (a) Noting the simplified equivalent circuit for the LNA in Figure 9P.1b, derive the input impedance Z_{IN} .

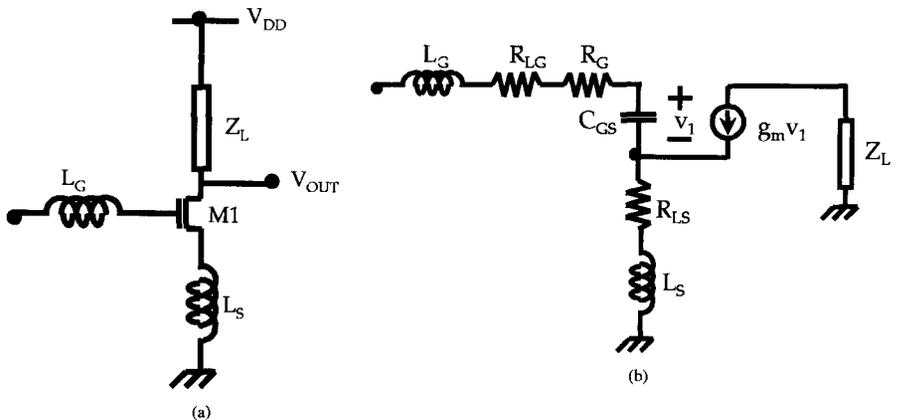


Figure 9P.1 (a) Single-transistor common source LNA topology and (b) simplified small-signal equivalent circuit for LNA topology in problem 9.3.

- (b) Derive the voltage gain, V_2/V_1 .
- (c) Given the following expression for drain current, with the cutoff frequency for the device shown:

$$I_D = \frac{k_n W}{2L} (V_{GS} - V_T)^2 (1 + \lambda V_{DS})$$

$$\omega_T \approx \frac{g_m}{C_{GS}}$$

and a 0.5- μm CMOS technology with $L_G = 0.5 \mu\text{m}$, $V_{DD} = 3.3 \text{ V}$, $C_{OX} = 3.83 \text{ fF}/\mu\text{m}^2$, $k_n = 120 \cdot 10^{-6} \text{ A}/\text{V}^2$, $\lambda = 0.05$, and $V_T = 0.8 \text{ V}$. Assuming a 50- Ω source impedance and a 400- Ω load impedance:

- (i) Given a dc power dissipation of 10 mW, gate voltage overdrive of roughly 175 mV = $V_{GS} - V_T$, inductor resistances of 1 Ω each, gate resistance of 3 Ω /square, and cutoff frequency of 5 GHz at this bias, what should the source inductance, L_S , be for optimum power match to a device with layout geometry of 50 fingers?
- (ii) What should the gate inductance, L_G , be for optimum power match?
- (iii) Given $\gamma = 5$, what is the resulting noise figure for this optimum power matching condition? What is the resulting voltage gain?

9.4 Often a transistor requires “external” matching with off-chip components for optimum source matching to reduce losses due to on-chip inductor resistance. Given that the optimum source match for noise for the transistor at 900 MHz is determined experimentally to be a conjugate match to $Z_{IN} = 20 \Omega - j \cdot 10 \Omega$, design the necessary matching networks to transform the transistor impedance up to a 50- Ω real impedance. (Refer to Fig. 9P.2.)

9.5 The noise characteristic of mixers is often represented in terms of effective-noise temperature, which relates to the noise figure according to

$$F = \frac{G_T(T_n + T_S)}{G_T T_S} \Big|_{T_S = T_0 = 290\text{K}} = 1 + \frac{T_n}{T_0}$$

where T_n accounts for the additive noise of the mixer, T_S is the noise temperature of the source, and G_T is the conversion gain (or loss) from the input power at the RF port to the output power at the IF port. From this

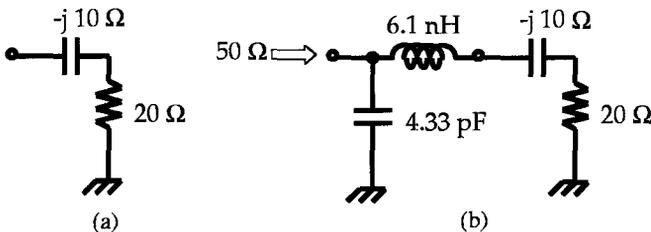


Figure 9P.2 (a) Equivalent circuit for transistor to be matched; (b) complete equivalent circuit for matched transistor following impedance transform up to 50 Ω .

expression, derive the condition that the noise figure of a passive mixer is equal to its conversion loss.

- 9.6** For the mixer circuit of Figure 9.31, what are the minimum dc voltages on the LO and RF terminals that will guarantee avoiding saturating the devices, given a maximum allowable LO input-signal power of 10 dBm, and a maximum allowable RF input signal power of 15 dBm in a 0.5- μm CMOS technology with threshold voltage $V_T = 0.6\text{ V}$?
- 9.7** Derive the expressions for power-added efficiency in class C operation from the ac voltage and current waveforms. Assume an ideal amplifier without distortion, an available supply voltage V_{DD} , the knee voltage V_K , the knee current I_K , and the conduction angle θ (the amount in radians of the time current flows during the entire 2π cycle of the periodic current waveform).

High-Speed or Low-Voltage, Low-Power Operations

I. C. CHEN

Worldwide Semiconductor Manufacturing Corporation
Hsinchu, Taiwan, ROC

W. LIU

Texas Instruments
Dallas, TX

10.1 INTRODUCTION

Since the late 1970s, we have witnessed the incredible growth, momentum, pace, and unbelievable achievements of the integrated circuit (IC) industry. Technology feature size has shrunk from greater than 1–2 μm to the current 0.25 μm or smaller. The memory [e.g., dynamic random access memory (DRAM)] density has increased from a few thousand bits to the current 64 million bits per chip. The clock frequency of the central processing unit (CPU or microprocessor) has increased from less than 1 MHz to the current 400 MHz for low-end personal computer applications and to 500–700 MHz for high-end workstation applications. The worldwide IC market has also grown 5–10 times to the current about 150 billion U.S. dollars per year (for 1997). The major end products that propelled such performance improvement, density increase, and market growth have been the computer (personal, workstation, etc.), consumer-related, as well as the recently emerging personal communication and/or Internet-related products. For the computer-related products to continue market growth, it is necessary to continue to increase the speed performance and functionality, as well as to reduce the power consumption and price per function of integrated circuits. At the same time, the personal communication and other handheld products demand higher performance, including higher operating frequencies and lower noise, longer battery lifetime, lower power consumption, and eventually integration of the whole system onto a single chip.

This chapter details the device-level considerations to achieve higher performance and lower power for digital IC technologies, and a higher operating frequency and lower noise figure for radiofrequency (RF) IC technologies.

10.2 HIGH-SPEED CONSIDERATIONS FOR DIGITAL APPLICATIONS

10.2.1 Performance Figure of Merit (FOM) for Digital CMOS

To illustrate the CMOS device design considerations for high-speed digital logic applications, we'll introduce the performance figure of merit (FOM) for digital logic circuits. Since the speed of digital logic circuits is usually determined by the pullup and pulldown time of each individual gate, let's use a simple inverter chain to introduce the performance FOM. The propagation delay (or delay per stage, τ_{delay}) of any inverter stage is an average of the n-MOS and p-MOS delay times,

$$\tau_{\text{delay}} = \frac{1}{2}(\tau_n + \tau_p + \tau_{\text{gate}}) \propto \frac{1}{\text{FOM}} \quad (10.1)$$

where τ_n and τ_p are sum of the pulldown and pullup time (in seconds) of the n-MOS and p-MOS, respectively, and τ_{gate} is related to the RC time constant for the signal to propagate the n-MOS and p-MOS gate electrodes. The τ_n is proportional to the total charge on the output node divided by the drive current of nMOS^{1,2}

$$\tau_n \propto \frac{C_{\text{total}} \cdot V_{DD}}{I_{\text{drive}}^n} \equiv \frac{1}{\text{FOM}_n} \quad (10.2)$$

and similarly τ_p can be expressed as

$$\tau_p \propto \frac{C_{\text{total}} \cdot V_{DD}}{I_{\text{drive}}^p} \equiv \frac{1}{\text{FOM}_p} \quad (10.3)$$

where V_{DD} is the power supply voltage, I_{drive}^n and I_{drive}^p (in $\mu\text{A}/\mu\text{m}$) are the drive current (measured at $|V_G| = |V_D| = V_{DD}$) per unit device width for the n-MOS and p-MOS, respectively, and C_{total} is the total capacitance at the output node and can be expressed as

$$C_{\text{total}} = \text{FO} \cdot (C_{\text{gate}}^n + C_{\text{gate}}^p) + (C_j^n + C_j^p) + C_{\text{interconnect}} \quad (10.4)$$

where FO is the fanout of the inverter, C_{gate}^n (or C_{gate}^p) is the gate capacitance for the n-MOS (or p-MOS) of the next stage, C_j^n (or C_j^p) is the junction capacitance for n-MOS (or p-MOS) of the output node of the current stage, and $C_{\text{interconnect}}$ is the interconnect capacitance between the current and the next stage.

As mentioned above, the τ_{gate} in Eq. 10.1 is related to the RC time constant required for the digital signal to propagate across the gate electrode, and can be

expressed as³

$$\tau_{gate} \propto (b_n \cdot R_{sh}^n + b_p \cdot R_{sh}^p) \cdot C'_{ox} \cdot W_n^2 \tag{10.5}$$

where R_{sh}^n (or R_{sh}^p) is the sheet resistance of the n-MOS (or p-MOS) poly gate, C'_{ox} is the gate capacitance per unit area, W_n is the width of the n-MOS transistor, and b_n (or b_p) is a geometric factor associated with the n-MOS (or p-MOS) and is discussed in more detail in Section 10.2.9.

Now, substituting Eqs. 10.2, 10.3, and 10.5 into Eq. 10.1, we have

$$FOM = \frac{2}{(1/FOM_n) + (1/FOM_p) + a \cdot (b_n \cdot R_{sh}^n + b_p \cdot R_{sh}^p) \cdot C'_{ox} \cdot W_n^2} \tag{10.6}$$

where a is an empirical constant. For a silicided poly technology with reasonably low gate sheet resistance ($R_{sh}^n, R_{sh}^p < 8 \Omega/\text{square}$) and relatively small W/L such that the τ_{gate} term can be neglected relative to the $1/FOM_n + 1/FOM_p$ term, the FOM formula can be simplified as

$$FOM \approx \frac{2}{\frac{1}{FOM_n} + \frac{1}{FOM_p}} = \frac{2}{C_{total} \cdot V_{DD} \cdot [(1/I_{drive}^n) + (1/I_{drive}^p)]} \tag{10.7}$$

We will use this FOM for most of the discussions except when the gate RC term is large and cannot be neglected. The validity of the FOM formula, Eq. 10.1 and Eq. 10.7, has been verified experimentally.^{1,2,4} Figure 10.1a shows such an example,

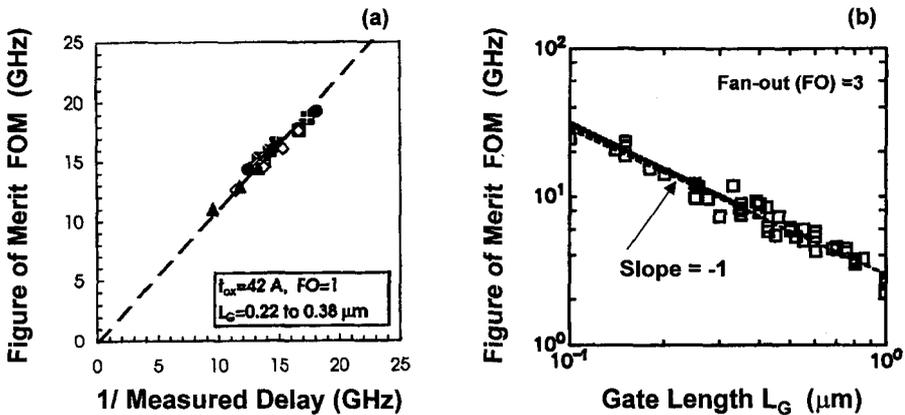


Figure 10.1 (a) Calculated CMOS figure of merit (FOM) (see Eq. 10.7, with measured I_{drive} , C_j , t_{ox} , etc.) plotted against measured $FO = 1$ inverter delays for some 0.25- μm CMOS wafers. The linear relationship between the two supports the validity of using the FOM formula as a speed indicator for digital CMOS. (After Nandakumar et al., Ref. 4.) (b) The FOM trend of published CMOS works (for L_G between 1 – 0.1 μm) versus gate length. The fact that the slope of the best-fit line is almost equal to -1 means that the CMOS performance is roughly inversely proportional to L_G . (After Rodder et al., Ref. 5.)

where the inverse of measured inverter delays ($1/t_{\text{delay}}$) are plotted against the corresponding FOMs that are calculated using measured parameters (e.g., I_{drive} , C'_{ox} , C_j). The excellent straight-line fit to the data points indicates a good linear correlation between FOM and $1/t_{\text{delay}}$, as shown in Eq. 10.1.

Using the FOM calculation, we can plot the calculated FOMs of reported CMOS technologies versus the corresponding gate lengths (between 0.1 and 1.0 μm), as shown in Figure 10.1b, where the data points have an excellent correlation with L_G with a slope of -1 , meaning that historically CMOS performance is inversely proportional to L_G . How to maintain this historically linear performance trend from the current 0.25 μm technology nodes toward 0.1 μm is definitely one of the major challenges for the device design and process integration engineers.

10.2.2 Drive Current of Deep-Submicrometer CMOS Transistor

It is obvious from Eq. 10.6 that to maximize the CMOS speed or performance FOM, the drive currents (I_{drive}) need to be maximized. To illustrate the fundamental parameters affecting I_{drive} , let's use the simple velocity saturation model⁶

$$I_{\text{drive}} \approx W \cdot C'_{\text{inv}} \cdot (V_{GS} - V_T - V_{D\text{sat}}) \cdot v_{\text{sat}} \quad (10.8)$$

where W is the device width, C'_{inv} is the gate capacitance per unit area measured in the inversion region, V_{GS} is the gate-to-source bias, V_T is the transistor threshold voltage, v_{sat} is the saturation velocity of electrons ($\approx 8 - 10 \times 10^6$ cm/s) or holes ($\approx 6 - 8 \times 10^6$ cm/s) in the inversion layer, and $V_{D\text{sat}}$ is the channel potential of the point where carriers are traveling at the saturation velocity v_{sat} . It can be shown that $V_{D\text{sat}}$ can be expressed as⁶

$$V_{D\text{sat}} = \frac{(V_{GS} - V_T) \cdot \xi_c \cdot L_{\text{eff}}}{(V_{GS} - V_T) + \xi_c \cdot L_{\text{eff}}} \quad (10.9)$$

where L_{eff} is the effective channel length, and ξ_c ($\approx 4 \times 10^4$ V/cm for electrons) is the critical electrical field at which the carrier velocity reaches v_{sat} . ξ_c can be approximated as

$$\xi_c = \frac{v_{\text{sat}}}{2\mu_{\text{eff}}} \quad (10.10)$$

where μ_{eff} is the effective carrier mobility in the inversion layer.

Note that the I_{drive} expression (Eq. 10.8) is valid for the ideal case where there is *no* parasitic resistance in series with the source or drain regions. For a practical device where there are parasitic resistances R_S and R_D connected in series with the source and drain terminal, respectively, then the drive current I_{drive}^\dagger

becomes

$$\begin{aligned}
 I_{drive}^\dagger &= \frac{W \cdot C'_{inv}}{(1 + W \cdot C'_{inv} \cdot R_S \cdot v_{sat})} \cdot (V_{GS} - V_T - V_{Dsat}) \cdot v_{sat} \\
 &= \frac{I_{drive}}{(1 + W \cdot C'_{inv} \cdot R_S \cdot v_{sat})}
 \end{aligned}
 \tag{10.11}$$

It is apparent that the higher the R_S , the lower the resultant I_{drive}^\dagger .

10.2.3 Device Design Considerations for High-Performance Digital CMOS

As we can see from Eqs. 10.2–10.6, the performance figure of merit (FOM) of a given CMOS technology is a function of the following factors: drive currents (I_{drive}^n and I_{drive}^p), gate capacitance (C_{gate}^n and C_{gate}^p , which are a function of the gate oxide thickness t_{ox}), junction capacitance (C_j), interconnect capacitance ($C_{interconnect}$), gate sheet resistance (R_{sh}^n and R_{sh}^p), and layout (which affects the geometric factors b_n and b_p).

In addition to these design variables, there are a number of parameters usually predetermined by the technology, system, or customer requirements, including gate length L_G , power supply voltage V_{DD} , and maximum OFF-state leakage current $I_{OFF}(max)$ (measured at $V_G = 0$ and $V_D = V_{DD}$). Moreover, as device dimensions scale, a number of other device design issues are becoming progressively more serious. Some examples of these issues are (1) direct tunneling leakage through the thin gate oxide, (2) boron penetration from the p^+ gate through the thin gate oxide, (3) diode leakage, (4) gate-induced drain leakage (GIDL), (5) random dopant fluctuation, and (6) device reliability [gate oxide integrity (GOI), channel hot-carrier (CHC), electrostatic discharge (ESD), latchup, etc.]. These design considerations are summarized in Table 10.1.

TABLE 10.1 Design Parameters For Digital CMOS

Input Parameters	$L_G, V_{DD}, I_{off}(max)$
Drive current I_{drive}	C_{inv}, L_{eff}, V_T , gate length control, short-channel effect (X_j , dopant profile), $R_{SD}, \mu_{eff}, v_{sat}$, width reduction (ΔW)
Loading capacitances	C_{gate}, C_{GD} (overlap), C_j (junction), $C_{interconnect}$
Resistance and layout	Gate R_{sh} , active-region $R_{sh}, W/L_G$
Other issues	Boron penetration Direct tunneling leakage Diode leakage Gate-induced drain leakage (GIDL) Random dopant fluctuation Device reliability

In the following sections, we discuss the effects of many of these factors and issues individually.

10.2.4 Gate Length, Power Supply Voltage, and Maximum Off-state Leakage Current

Gate length (L_G), power supply voltage (V_{DD}), and maximum OFF-state leakage current [$I_{OFF}(\max)$] are usually predetermined by the system and/or technology requirements.

Gate length is usually determined by the lithography (and etch) capability of a given technology generation. Therefore it should be fairly understandable that L_G cannot be shrunk *unconditionally* to gain more speed.

The power supply voltage is usually determined by the following considerations: system compatibility, power consumption, device reliability, and performance. It is generally considered desirable to keep V_{DD} of a given chip the same as the V_{DD} used by other chips on the same system board. The voltage compatibility can be solved by having two different V_{DD} values for the input/output devices and for the rest of the core circuit. Power and performance tradeoff is another major consideration for determining V_{DD} . Since the active power of a high-performance circuit is proportional to $C \cdot V^2 \cdot f$, the lower the V_{DD} the lower the power consumption. On the other hand, from the I_{drive} (Eq. 10.8) and the FOM formula (Eq. 10.6), it is clear that once the power supply voltage V_{DD} is lowered, the speed is likely to decrease. Therefore, a judicious choice of V_{DD} is important such that a reasonable tradeoff between speed and power consumption can be obtained.

Maximum OFF-state leakage current, $I_{OFF}(\max)$, is usually the parameter that determines the standby power of an integrated circuit. For high-performance circuits used in desktop applications where leakage current under standby condition is not a major concern, the $I_{OFF}(\max)$ can be relatively high (e.g., 1 nA/ μm). On the other hand, for handheld applications where battery lifetime is of primary importance, the $I_{OFF}(\max)$ needs to be relatively low (e.g., 10 pA/ μm or lower) to conserve battery power during standby. Once the $I_{OFF}(\max)$ is determined, the threshold voltages of the CMOS transistors are pretty much determined, since the range of subthreshold swing for the bulk CMOS is almost the same (80–90 mV/decade) for properly designed transistors.

10.2.5 Effective Channel Length L_{eff} and Threshold Voltage V_T

L_{eff} and V_T are of primary importance to determine drive current, as shown in Eqs. 10.8 and 10.9. However, for given gate length and maximum off-state leakage $I_{OFF}(\max)$, V_T and L_{eff} are actually not independently adjustable parameters for device designers. As mentioned above, V_T is mostly determined by $I_{OFF}(\max)$ for bulk CMOS whose subthreshold swing is usually around 75–90 mV/decade. Also, as is shown in the next section, gate length control and V_T rolloff also impact the nominal V_T .

L_{eff} is largely determined by the tradeoff among short-channel effect, R_{SD} , hot-carrier reliability, and performance (i.e., drive current). Drive current will increase with decreasing L_{eff} , since a shorter L_{eff} will result in smaller $V_{D\text{sat}}$ and thus larger I_{drive} (see Eqs. 10.8 and 10.9). However, for a given target L_G , shorter L_{eff} will aggravate the short-channel effect (e.g., increase the V_T rolloff) and thus degrade the performance. Therefore L_{eff} is usually a certain fraction (e.g., $\frac{2}{3}$ to $\frac{3}{4}$) of L_G with limited freedom for adjustment.

10.2.6 Gate Length Critical Dimension (CD) Control, $L_G(\text{min})$, and $L_G(\text{nom})$

Gate length CD control, defined as the maximum (or 3σ) poly-gate length variation for a given nominal design gate length, is of vital importance to the CMOS performance. Before we talk about the impact of L_G CD control on performance, let's define the terms of minimum gate length [$L_G(\text{min})$], nominal gate length [$L_G(\text{nom})$], and the V_T and drive currents corresponding to these gate lengths. Figure 10.2a shows a typical V_T versus L_G (or V_T rolloff) for an nMOS transistor design. The L_G corresponding to $I_{\text{OFF}} = I_{\text{OFF}}(\text{max})$ is defined as $L_G(\text{min})$. For a given L_G CD control, the L_G equal to [$L_G(\text{min}) + L_G$ CD control] is defined as $L_G(\text{nom})$, the nominally manufacturable gate length at which all devices should have I_{OFF} smaller than $I_{\text{OFF}}(\text{max})$. The drive currents corresponding to $L_G(\text{min})$ and $L_G(\text{nom})$ are defined as $I_D(\text{strong})$ and $I_D(\text{nom})$, respectively. It is the $I_D(\text{nom})$, rather than the $I_D(\text{strong})$, that should be used in the FOM formula (Eqs. 10.1–10.6) to estimate the performance of a given technology.

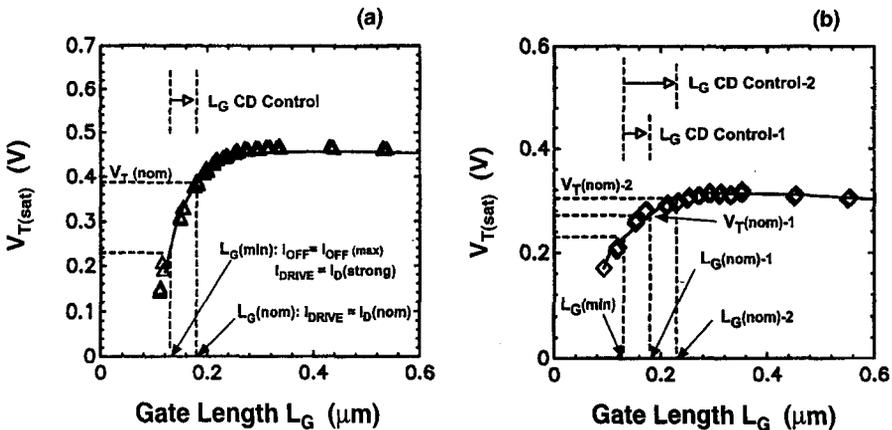


Figure 10.2 (a) An illustration of definitions of $L_G(\text{min})$ [the L_G at which $I_{\text{off}} = I_{\text{off}}(\text{max})$] and $L_G(\text{nom})$ [= $L_G(\text{min}) + L_G$ CD control]. The drive current at $L_G(\text{nom})$, $I_D(\text{nom})$, is used in the FOM equation. (b) An illustration of the impact of the L_G CD control on the CMOS performance. For the larger L_G CD control (case 2), the corresponding $L_G(\text{nom})-2$ is longer and $V_T(\text{nom})-2$ is higher, and thus the $I_D(\text{nom})$ will be lower. Therefore, L_G CD control is of critical importance to achieve high performance.

The importance of L_G CD control on the CMOS performance can be further illustrated in Figure 10.2*b*, which shows the V_T versus L_G plot for a given transistor design but with two different L_G CD control capabilities. It is apparent that case 2 will have lower $I_D(\text{nom})$ because the $V_T(\text{nom})$ is higher and $L_G(\text{nom})$ is longer compared to those of case 1. Usually it is considered reasonably aggressive for the L_G CD control to be 10% of the target $L_G(\text{nom})$. How to meet this goal has been and will continue to be a major challenge for process engineers.

10.2.7 Gate Oxide Thickness t_{ox}

Gate oxide thickness (t_{ox}) is one of the most fundamental device parameters. The choice of t_{ox} affects the performance (I_{drive} , channel concentration, carrier mobility, and gate capacitance, etc.), reliability (gate oxide integrity, channel hot carrier, etc.), and power consumption (V_{DD} and gate capacitance) of a CMOS technology. Although the dependencies on t_{ox} are quite involved, how to determine t_{ox} is becoming fairly straightforward. Figure 10.3*a* shows the trend of t_{ox} of logic technologies as a function of L_G (from 1 to 0.1 μm), where the t_{ox} values of the SIA (Semiconductor Industry Association) *Roadmap* for sub-0.25- μm technologies are also included. Figure 10.3*b* shows the corresponding V_{DD}/t_{ox} versus L_G plot for the technologies shown in Figure 10.2. It is quite clear that the V_{DD}/t_{ox} ratio gradually increases for shorter L_G ; for sub-0.25- μm technologies the V_{DD}/t_{ox} ratio is between 4 to 5 MV/cm, which is also consistent with the SIA *Roadmap*. As mentioned in Section 10.2.5, V_{DD} is usually not an adjustable parameter for device designers, therefore t_{ox} is roughly equal to $V_{DD}/(5 \text{ MV/cm})$ for sub-0.25- μm technologies.

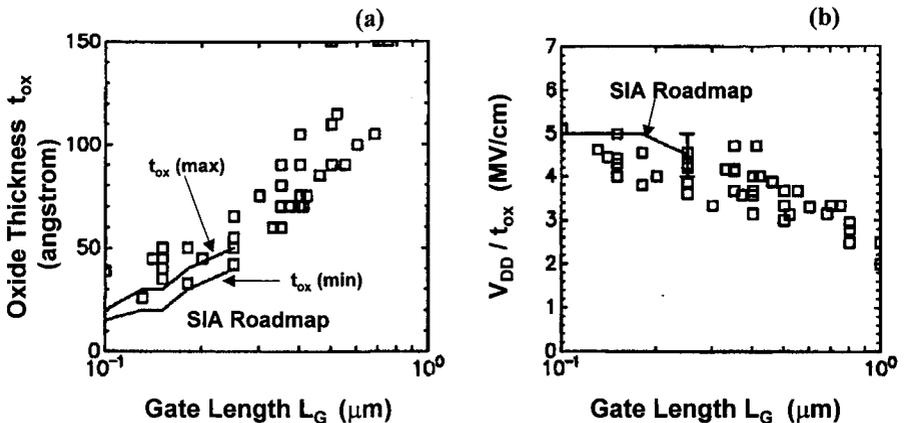


Figure 10.3 (a) A historical trend of t_{ox} as a function of gate length. For sub-0.25- μm technologies, the t_{ox} are projected to be thinner than 40 Å (including data from 1997 SIA *Roadmap*). (b) A trend of V_{DD}/t_{ox} for the same technologies shown in Figure 10.3*a*. The V_{DD}/t_{ox} trends upward from about 3.5 MV/cm for 0.5- μm technologies toward 5 MV/cm for sub-0.2- μm technologies (including data from 1997 SIA *Roadmap*).

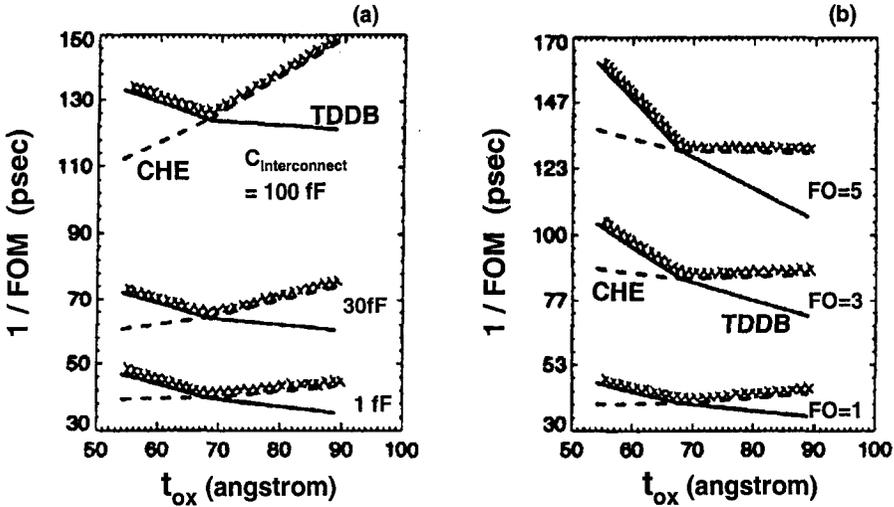


Figure 10.4 Using the allowable V_{DD} , the projected inverter delay ($1/FOM$) versus t_{ox} plots: (a) for various interconnect capacitances; (b) for various fanouts. (After Rodder et al., Ref. 2.)

Although the determination of t_{ox} is becoming straightforward and defined in the SIA Roadmap, the validity of the t_{ox} so determined is not as simple and worth mentioning.^{2,7} Figure 10.4a,b show the inverter delays as a function of t_{ox} with varying interconnect capacitance ($C_{interconnect}$) and varying fanout ($FO = 1-5$), respectively. For driving a long metal line with large $C_{interconnect}$, the higher the I_{drive} , the better. Thus, the minimum delay always occurs at around $t_{ox} = 6-7$ nm, where the allowable V_{DD} and I_{drive} are the highest. On the other hand, for driving logic gates with various fanouts (FOs), minimum delay occurs at $t_{ox} = 6-7$ nm for smaller FO ($= 1-3$); while for larger FO (of ≥ 5) the minimum delay becomes relatively insensitive to t_{ox} and gradually shifts toward thicker t_{ox} (8-9 nm). This is because for large FO, loading capacitance is dominated by gate capacitance which increases linearly with $1/t_{ox}$. However, I_{drive} does not increase linearly with $1/t_{ox}$ even at a fixed $V_G - V_T$, due to the higher vertical effective electrical field and thus lower surface mobility, which in turn increases the ξ_c (see Eq. 10.10) and thus lowers the I_{drive} (Eq. 10.8). Therefore, under high fanout situations, the optimum t_{ox} tends to be a little thicker than what is optimal for driving large $C_{interconnect}$. Nonetheless, the choice of t_{ox} for logic technologies has been determined by the scenario of driving large $C_{interconnect}$, simply because such a t_{ox} is optimum for driving long lines and near optimum for driving large fanouts.

10.2.8 Short-Channel Effects (SCE)

The short-channel effect is one of the most important challenges for a device designer to overcome for each new technology node. As CMOS technology and gate length continue to scale, the device design for a previous generation likely cannot be

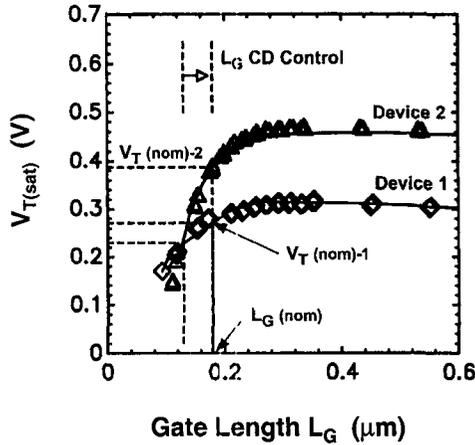


Figure 10.5 An illustration of the importance of minimizing short-channel effects (SCE represented by the V_T rolloff shown) to obtain high-performance. Both device designs support the same L_G (min). The V_T -rolloff (or SCE) of the case 1 device is smaller, the V_T (nom) is lower, and I_{drive} is higher.

used without modification. Among the many manifestations of the short-channel effects [e.g., V_T rolloff, drain-induced barrier lowering (DIBL), subthreshold swing degradation], V_T rolloff is chosen here to study the performance implication due to short-channel effects. Figure 10.5 shows the V_T rolloff for two different transistor designs: cases 1 and 2. Both designs support the same L_G (min), but case 2 has a much larger V_T rolloff than that of case 1 because of short-channel effects. For a given L_G CD control, case 1 has a lower V_T (nom), and, thus, a likely higher I_D (nom) than case 2. Also, the performance variation for L_G around L_G (nom) is smaller for case 1 than case 2. Therefore, it is obvious that one should reduce short-channel effects to achieve high-performance operation.

Short-channel effects are a function of junction depth, width of S/D depletion regions, which in turn is a function of substrate doping concentration and bias, and oxide thickness. Specifically, the shortest channel device without significant short-channel effects, defined as L_{min} , can be expressed as⁸

$$L_{min} = 0.4 [x_j \cdot t_{ox} \cdot (W_S + W_D)^2]^{1/3} \tag{10.12}$$

where x_j is the S/D junction depth, t_{ox} is the the gate oxide thickness, and $W_S + W_D$ is the sum of the source and drain depletion width.

Among the four fundamental parameters (t_{ox} , x_j , N_{sub} , and V_{DD}) affecting the short-channel effects, V_{DD} and t_{ox} are determined by other considerations as mentioned in the previous sections. Therefore, channel/substrate dopant concentration (N_{sub}) and junction depth (x_j) are the only adjustable parameters to improve short-channel effects.

SCE—Channel Dopant N_{sub} Engineering

There are generally three ways to engineer channel doping (1) conventional near-uniform channel doping, (2) retrograde channel doping, and (3) pocket-implanted channel doping.

The conventional doping scheme includes the use of V_T adjust and punchthrough-stopping implant before the gate oxide growth. The implant species are usually boron (B) or BF_2 for n-MOS, and phosphorus (P) for p-MOS. Because of the relatively large diffusivities for B and P, the channel doping profiles after all the thermal processes are relatively more uniform compared to the other two channel doping schemes.

Retrograde channel doping means that the dopant concentration is low at the surface and the concentration increases toward the substrate. There are several implementations for such a dopant profile, such as ground plane,⁹ supersteep retrograde (SSR),¹⁰ or delta doping.¹¹ The purposes of such a dopant profile are to avoid carrier mobility degradation due to increased coulombic scattering resulting from high dopant concentration and to maintain good immunity to short-channel effects and subsurface punchthrough by increasing the subsurface concentration. The retrograde dopant profile can be implemented by either ion implantation of heavy ions with low diffusivity (e.g., indium for n-MOS and arsenic or antimony for p-MOS)^{4,10} or low-temperature Si epitaxy of an undoped layer on a heavily doped region.¹¹ Figure 10.6a compares a typical retrograde (SSR) dopant profile with a conventional near-uniform one.

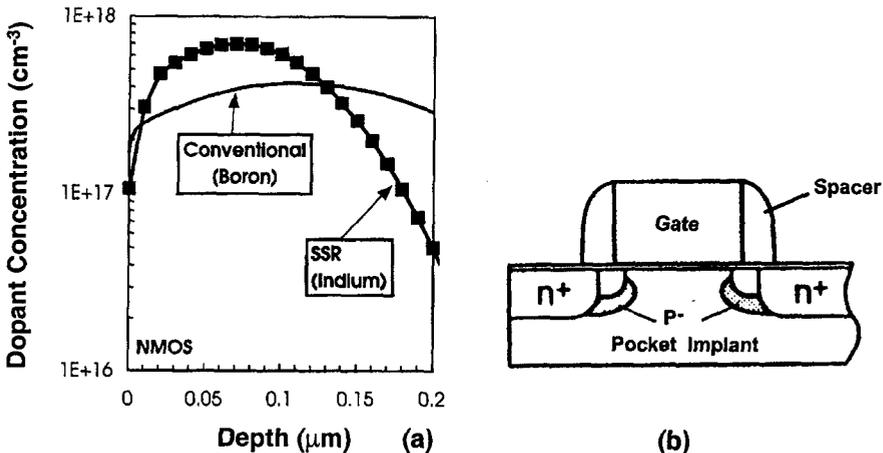


Figure 10.6 (a) Examples of channel dopant profiles: a conventional (near uniform case), and a supersteep retrograde (SSR) case. Also shown is an optional surface counter-doping profile to achieve lower V_T . (After Nandakumar et al., Ref. 4.) (b) A schematic diagram of a pocket (or halo)-implanted MOSFET. Since the relatively high dopant concentration is close to the S/D regions only, the pocket-implanted device can maintain low long-channel V_T and achieve minimal short-channel V_T rolloff. (After Hori, Ref. 13.)

The pocket or halo implant is done after the poly gate is defined, such that the dopant profile looks like pocket or halo.¹² Figure 10.6b shows the cross sectional view of a pocket implanted n-MOS transistor. In order to increase its effectiveness, the pocket implant can be done at a tilt angle with rotation so that the pocket dopant can be placed in front of the S/D extensions.¹³ The pocket implant has been used prevalently for sub-0.25- μm devices^{5,14–20} because of its excellent resistance to V_T rolloff at short L_{eff} when the pocket dopants from the source and drain merge and effectively increase the channel concentration.

Recent SCE Modeling For uniformly-doped and delta-doped (or retrograde-doped) devices, the short-channel effect, including V_T rolloff and drain-induced barrier lowering, can be modeled as²¹

$$V_T = V_{T0} - 1.8\sqrt{V_D + 0.8}e^{-(L_{\text{eff}}/2l)} \quad (10.13)$$

where V_{T0} is the long-channel threshold voltage, V_D is the drain bias, and l is the characteristic length of V_T rolloff and can be approximated as²²

$$l \approx (t_{\text{ox}} \cdot x_{bg} \cdot x_j)^{1/3} \quad (10.14)$$

where x_j is S/D junction depth and x_{bg} is the back-gate thickness. For bulk CMOS, x_{bg} is the depletion width underneath the channel.

For a given acceptable V_T rolloff, namely, $V_{T0} - V_T \equiv \Delta V_T$, the L_{eff} is the shortest acceptable channel length, L_{min} . Thus L_{min} can be expressed as

$$L_{\text{min}} = 2l \cdot \ln \left[\frac{1.8\sqrt{V_D + 0.8}}{\Delta V_T} \right] \approx 7l \quad (10.15)$$

This equation assumes $\Delta V_T = 0.1 \text{ V}$, and $V_D = 2 \text{ V}$. After substituting the x_{bg} with the depletion width at $V_G = V_T$, it can be shown that²²

$$L_{\text{min}} \approx 7 \left[\frac{3n(2\phi_g + V_{SB})}{V_T - (V_{FB} + 2\phi_B)} \right]^{1/3} \cdot t_{\text{ox}}^{2/3} \cdot x_j^{1/3} \quad (10.16)$$

where ϕ_B is the potential difference between the Fermi level and the midgap of Si, V_{FB} is the flat-band voltage, V_{SB} is the source to substrate bias, and n is a unit-less constant. For *ideal* uniformly doped device, $n = 2$; while for *ideal* delta-doped device, $n = 1$. Therefore, the delta-doped device has shorter L_{min} than the uniformly doped device because x_{bg} is smaller for the delta-doped case. For regular dopant profiles, it is natural that $1 < n < 2$.

Figure 10.7a shows the V_T versus L_{min} according to Eq. 10.16 for uniformly doped ($n = 2$, the solid curve) and delta-doped ($n = 1$, the dotted curve) devices. Also plotted in the figure are experimental data points published in the literature (with $t_{\text{ox}} = 4 \text{ nm}$, $x_j \approx 40\text{--}70 \text{ nm}$). The fact that almost all the data points lie between

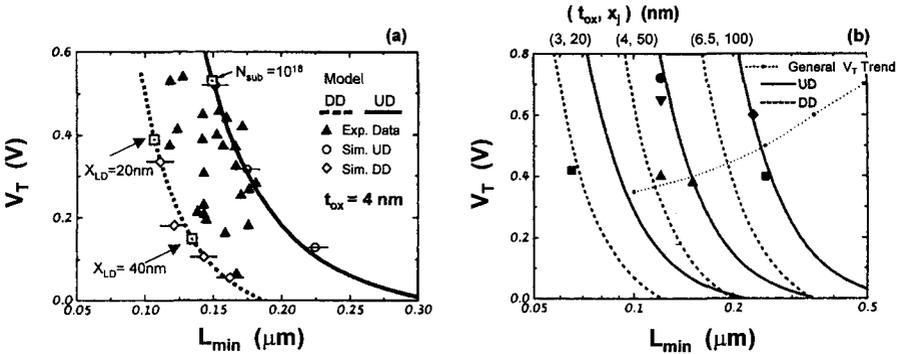


Figure 10.7 (a) The relationship of long-channel V_T and minimum channel length L_{min} for MOSFET with uniformly doped (UD) and delta-doped (DD) channel profiles ($t_{ox} = 4 \text{ nm}$; x_j 40–70 nm). (After Wann et al., Ref. 22.) (b) A plot similar to Figure 10.7a showing the importance of scaling both t_{ox} and x_j to achieve smaller L_{min} . (After Wann et al., Ref. 22.)

the bounds of the delta-doped and uniformly doped cases supports the validity of the model (Eq. 10.16). Figure 10.7b shows a similar plot of V_T versus L_{min} but with several different pairs of t_{ox} and x_j . It is clear that scaling t_{ox} and x_j are important to continue to scale L_{min} . For a given t_{ox} and x_j , the delta-doped device can support a shorter L_{min} than the uniformly doped device. In general, the advantage of using delta doping for channel engineering is about one generation of L_{min} for a given t_{ox} and x_j .²² Figure 10.8a is an example of a V_T rolloff comparison for a SSR (retrograde

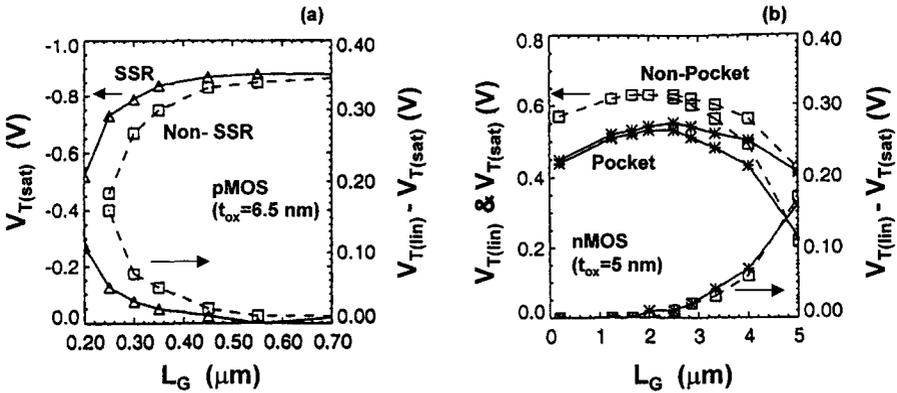


Figure 10.8 (a) A comparison of V_T rolloffs for a supersteep retrograde (SSR) and a conventional (non-SSR) channel profile. For the same long-channel V_T , the SSR case apparently has a shorter L_{min} than the non-SSR case. (After Rodder et al., Ref. 5.) (b) A comparison of V_T rolloff for a pocket-implanted and a conventional (non-pocket-implanted) device. Both devices support a L_G (min) of $0.2 \mu\text{m}$, while the pocket device has a lower V_T (nom) at L_G (nom) of $0.25 \mu\text{m}$. Both $V_{T(lin)}$ and $V_{T(sat)}$ are shown for each device. (After Rodder et al., Ref. 5.)

doped) and a non-SSR device. For the same long-channel V_T , the L_{\min} for the SSR device is apparently shorter than the non-SSR device.

For pocket-implanted devices, a different V_T rolloff expression (from Eq.10.13) should be used as ²²

$$\Delta V_T = V_{T0} - V_T = -C_1 e^{-(L_{\text{eff}}/l)} + (k-1)C_2 e^{-(L_{\text{eff}}/2l)} \quad (10.17)$$

where C_1 and C_2 are parameters dependent on V_D , and k is dependent on pocket-implant conditions as

$$k = \sqrt{\frac{V_T \ln(N_P/N_{\text{SUB}})}{V_D}} \cdot \frac{L_P}{l} \quad (10.18)$$

where N_P and L_P are the concentration and length of the pocket implant. It is clear that with the help of the pocket-implant term [i.e., $(k-1)C_2 e^{-(L_{\text{eff}}/2l)}$], ΔV_T for decreasing L_{eff} will tend to increase slightly then decrease instead of decreasing monotonically for uniformly or retrograde doped devices. Therefore, the use of a pocket implant can further scale the minimum acceptable channel length L_{\min} for a given t_{ox} and x_j . Figure 10.8b shows such an example of comparing the V_T rolloff of a pocket-implanted and a nonpocket device. In this example, the L_{\min} for the two cases are about the same, while the long channel V_T for the pocket case is about 0.15 V lower than the non-pocket case, indicating better short-channel effect. In general, by using a pocket implant one can improve L_{\min} by two generations compared to a device with near-uniform channel doping with the same t_{ox} and x_j .²²

Although the retrograde doping or the pocket implant can improve the short-channel effect and L_{\min} compared to uniformly doped devices, the drive currents (measured at the same $V_G - V_T$) can degrade with these channel doping techniques.²³ The drive current degradation can be attributed to either the bulk charge effect due to reduced x_{bg} ^{22,23} and/or mobility degradation due to high concentration at the source side. Therefore, device designers must use good judgment to control the short-channel effect and to avoid excessive performance degradation.

SCE—Junction Depth x_j , and Source/Drain Resistance (R_{SD})

As mentioned above, junction depth (x_j) also plays an important role in controlling the short-channel effect. Besides, since the S/D extension implants are self-aligned to the gate edge, it is obvious that x_j needs to be scaled proportional to L_G such that L_{eff} can maintain a reasonable fraction of L_G . According to the SIA Roadmap, x_j is forecasted to follow L_G as

$$0.3 \cdot L_G \leq x_j \leq 0.5 \cdot L_G \quad (10.19)$$

A direct tradeoff against the shallow junction x_j is the sheet resistance, which is an important component of the total S/D parasitic resistance R_{SD} .^{24,25} As mentioned in

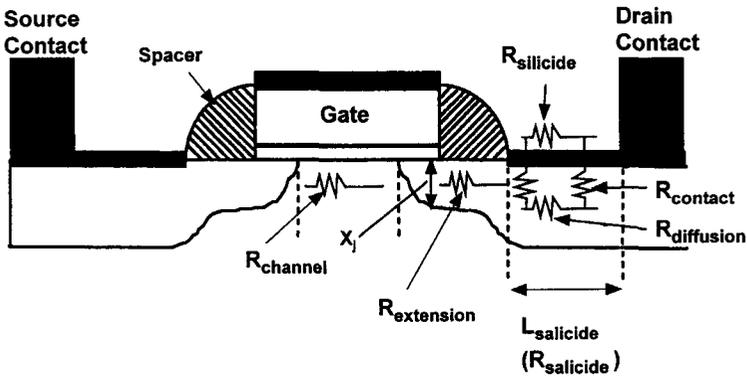


Figure 10.9 A schematic diagram of a salicided MOS transistor showing various resistance components along the current path.

Eq. 10.11, increasing R_{SD} , particularly the source resistance R_S , will result in a reduction of I_{drive} . Figure 10.9 shows the resistance components of a typical salicided MOSFET. For a symmetrical device, the total parasitic source–drain resistance R_{SD} can be expressed as

$$R_{SD} = 2 \cdot (R_{salicide} + R_{extension}) \tag{10.20}$$

where $R_{extension}$ is the sheet resistance in the source–drain extension region, and $R_{salicide}$ is the resistance in the source–drain salicide region, and can be expressed as²⁶

$$R_{salicide} = \frac{R_{silicide} \cdot R_{diffusion}}{R_{salicide} + R_{diffusion}} \cdot \frac{L_{salicide}}{W} + R_{contact} \tag{10.21}$$

where $R_{silicide}$ and $R_{diffusion}$ are the sheet resistances of the silicide and underlying diffusion region, respectively; $L_{salicide}$ is the length of the salicide region measured from the edge of contact hole toward the gate (see Fig. 10.9); W is the transistor width (thus $L_{salicide}/W$ is the number of squares for the salicide region); and $R_{contact}$ is the contact resistance between the silicide and the underlying diffusion region. As can be seen from Eq. 10.21, the resistance in the S/D extension region directly contributes to R_{SD} . Therefore, as we scale the x_j of the S/D extension regions, we need to weigh the benefit of reduced short-channel effects against the disadvantage of increased source–drain resistance.

SCE—Junction Depth x_j : Processing Techniques

The processing techniques to achieve shallow junctions can generally be categorized as ion implantation and diffusion.

Ion Implantation Ion implantation has been and will continue for some time to be the primary doping technique simply because it is easy to use; photoresist can be used as a mask to selectively introduce dopants. The major drawback or limitation of ion implantation is that a subsequent high-temperature ($>800^{\circ}\text{C}$) annealing is necessary to activate the implanted dopants as well as to anneal out the damages or defects created by the implant. Both the high-temperature annealing and the defect-induced transient-enhanced diffusion (TED)²⁷ deepen the junction depth.

Besides the conventional ion implantation (with energy $\geq 20\text{ keV}$) that has been used in production for over a decade, the following techniques have been explored to achieve shallow x_j : (1) low-energy ($< 5\text{ keV}$) implantation, (2) implantation of heavy ions, (3) preamorphization implant (PAI) before implanting the desired dopants, and (4) plasma doping.

Using a lower-energy implant to form shallower junctions is a natural trend for ion implantation. The lower energy implant can reduce the depth of the as-implanted profile, and reduce the amount of implant damage. Thus, the associated thermal budget for damage annealing and the effect of transient enhanced diffusion can be reduced. The feasibility of implanting B^+ ion at 200 eV^{28} to 1 keV^{25} had been demonstrated. Recently production-worthy low-energy ($< 1\text{ keV}$) ion implanters are becoming commercially available. It is expected that shallow junctions formed by proper conditions of low-energy implant and subsequent rapid thermal annealing should satisfy the shallow x_j requirements of 0.18 and $0.13\text{ }\mu\text{m}$ technology nodes.

Generally there are two reasons to implant heavy ions: (1) to form a retrograde channel profile and (2) to form shallow junctions. To achieve this latter purpose, heavy single-element (e.g., Sb for n^+ junction²⁹) or compound ions (e.g., BF_2 or $\text{B}_{10}\text{H}_{14}$ ³⁰) have both been explored. The primary reason for choosing such heavy ions is to reduce the effective implant energy.

The use of preamorphization implant (PAI) prior to the implantation of an intended dopant species is another way of forming shallow junctions.^{31,32} One primary motivation of using PAI is that the preamorphized layer can prevent the subsequent dopants from “channeling” through the Si lattice and thus can reduce or eliminate the dopant profile “tail.” Using Si or Ge as the PAI elements had been reported.³¹ In addition, the use of heavy elements with opposite conductivity type as PAI species (e.g., Sb for B p^+ junction, and In for As n^+ junction),³² can further sharpen the dopant profile by counterdoping.

Plasma doping is done by exposing a Si wafer to a dopant plasma (e.g., B_2H_6 for p-type dopant³³) and letting the energetic dopant ions in the plasma be “implanted” onto the wafer except without the ion species and energy selection mechanisms. The lack of ion-selection mechanism is a major drawback of the plasma doping, because contaminants such as metallic or carbon ions will be implanted into the wafer together with the desired dopants.

Diffusions As opposed to ion implantation, three diffusion schemes are proposed as alternate doping methods relative to the mainstream of ion implantation: (1) doped epitaxial or amorphous Si, (2) doped deposited oxide, and (3) gas-immersion laser doping. Because there is no physical ion bombardment on the Si substrate,

these doping schemes are free from the defect-induced, transient-enhanced diffusion (TED) and it is easier to achieve shallower junctions. This is a common advantage of these diffusion schemes. However, because of the high-temperature nature of the processes, a hard mask (e.g., an oxide or nitride layer) must be used, which explains why these diffusion schemes have seldom been used in manufacturing so far. Nonetheless, these diffusion schemes are backup approaches in case the limitation of ion implantation is reached.

Dopant diffusion from doped silicon is one of the techniques reported to form shallow junctions. The way it works is to deposit in situ-doped selective epitaxial Si³⁴ and then to use thermal anneals to drive out the dopants and form shallow junction. Besides using doped epitaxial silicon, two schemes using doped *amorphous* Si have been reported: (1) selective deposition of doped amorphous Si³⁵ and (2) nonselective deposition of amorphous Si.³⁶

Using in situ doped CVD oxide as sidewall spacers is similar in principle to the doped Si sidewall spacers mentioned above. Both phosphorus-doped silicate glass (PSG)³⁷ or boron-doped silicate glass (BSG)³⁸ were used as diffusion sources to form n⁺ and p⁺ shallow junctions, respectively.

Gas-immersion laser doping (GILD) immerses the wafer in a chamber filled with some desired dopant gas, then uses high-energy laser pulses to locally heat and melt the Si to incorporate the dopants into the Si and form shallow junctions.³⁹ Junction depth and dopant concentration can be controlled by the laser pulse duration and dopant gas pressure. The major attractiveness of this scheme is that the laser beam can potentially melt a certain specified region through the use of a mask, and eliminate the inconvenience of using hard masks. However, the feasibility of this doping scheme is still to be proven, mainly depending on the development of a cost-effective laser stepper system and masking scheme.

Figure 10.10 compares the sheet resistance with the junction depth for the ion implantation and diffusion schemes mentioned above. For any given x_j, the lower the

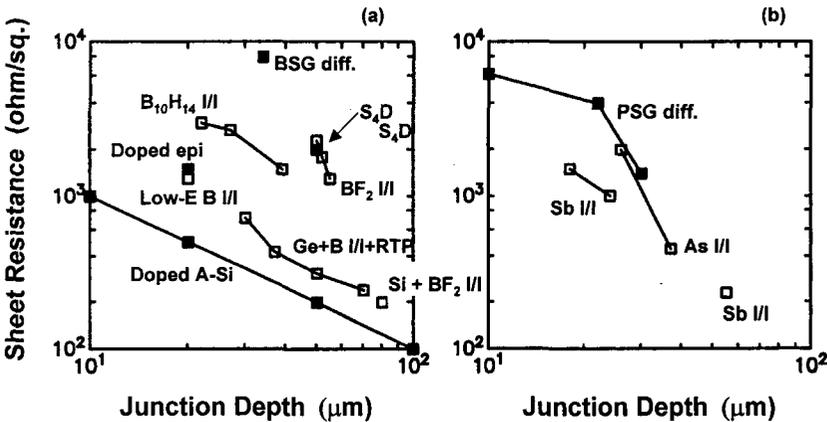


Figure 10.10 A summary of published sheet resistance versus junction depth for (a) p⁺ and (b) n⁺ shallow junctions. The open squares represent for ion-implant-related techniques; the filled squares, are diffusion-related techniques.

sheet resistance the better. For the p^+ junction (Fig. 10.10a), except for the doped α -Si case, the lower resistance cases are all related to ion implantation. The same applies for the n^+ junctions (Fig. 10.10b).

10.2.9 Gate/Active-Region Sheet Resistances

Gate Sheet Resistance

Gate resistance is one of the key parameters that determines the high frequency (or radiofrequency) and high-speed characteristics of a digital CMOS technology. Its importance for RF applications are described later in this chapter.

For digital applications, the performance degradation with increasing gate resistance results from the increasing delay in propagating a voltage signal across a resistive poly line. Consequently, there is a delay in switching across the entire width of the transistor. As briefly mentioned in Section 10.2.1, this effect is modeled as an additional delay (τ_{gate}) proportional to the RC product of a poly line with a linewidth equal to the gate length. The length of this line depends on the transistor width, as well as the layout geometry. As shown in Eq. 10.5, τ_{gate} can be modeled as³

$$\tau_{\text{gate}} \propto (b_n \cdot R_{\text{sh}}^n + b_p \cdot R_{\text{sh}}^p) \cdot C'_{\text{ox}} \cdot W_n^2 \quad (10.22)$$

where R_{sh}^n (or R_{sh}^p) is the sheet resistance of the n-MOS (or p-MOS) poly gate, C'_{ox} is the gate capacitance per unit area, W_n is the width of the n-MOS transistor, and b_n and b_p are geometric factors associated with the n-MOS and p-MOS.

The determination of the geometric factors b_n and b_p can best be illustrated by the examples shown in Figure 10.11. Let's define W_n and W_p as the n-MOS and p-MOS transistor width, respectively, while w_n and w_p are the width of each gate finger for n-MOS and p-MOS, respectively. Then the definition of b_n and b_p can be expressed as

$$b_n = \left(\frac{w_n}{W_n} \right)^2, \quad b_p = \left(\frac{w_p}{W_n} \right)^2 \quad (10.23)$$

Therefore, for the layout shown in Figure 10.11a, $b_n = 1$ and $b_p = 4$, while for the layout example shown in Figure 10.11b, $b_n = b_p = 1$.

This model of τ_{gate} has been verified extensively via simulations as well as experimental measurements. Figure 10.12a shows the ideal inverter delays (e.g., calculated 1/FOM *without* considering the gate resistance effect) plotted against the measured delays. Note that without considering the τ_{gate} term, the 1/FOM predictions do not correlate with measurements. Figure 10.12b shows the corresponding comparison with the gate resistance *included* in the FOM calculation. The good correlation between the predicted and measured delays supports the validity of the τ_{gate} model, Eq.10.22.

Based on this τ_{gate} model, one can construct the gate resistance requirements for various technology nodes. Assuming equal sheet resistance for n- and p-type gates (i.e., $R_{\text{sh}}^n = R_{\text{sh}}^p = R_{\text{sh}}$) and the single-gate layout shown in Figure 10.11a, the

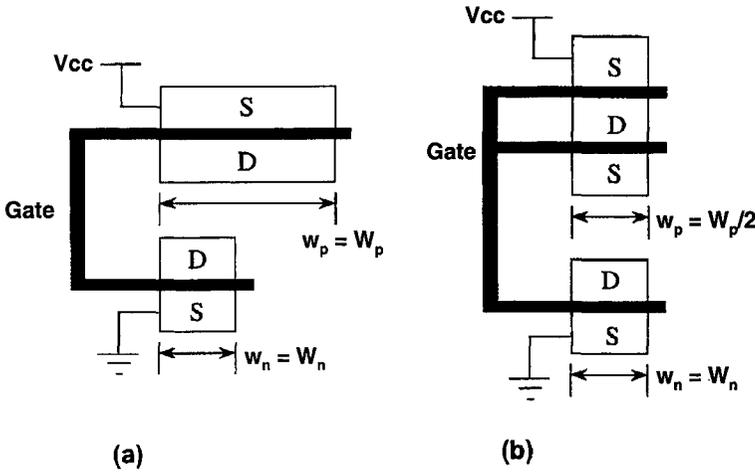


Figure 10.11 Schematic diagrams of two inverter layouts used to evaluate the gate resistance effect on speed performance: (a) a single-gate (or poly) layout; (b) a parallel-gate (for p-MOS) layout. (After Chatterjee et al., Ref. 3.)

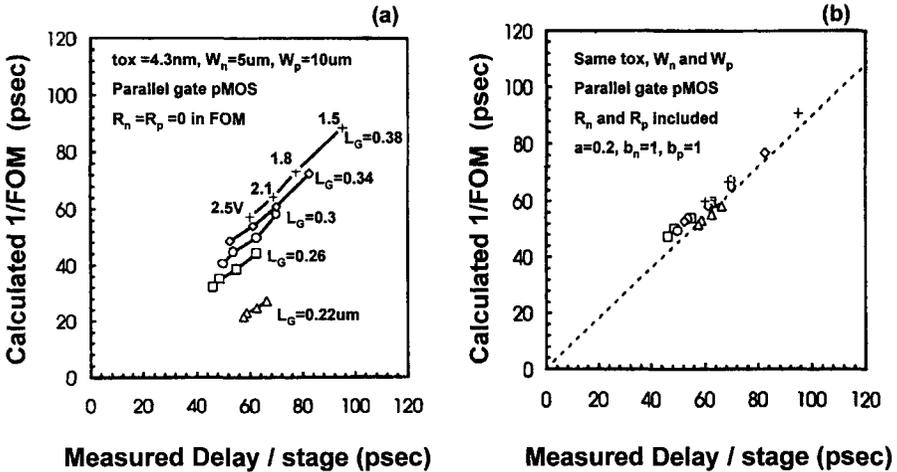


Figure 10.12 Calculated 1/FOM versus measured delay for some high gate resistance devices and inverter chains; (a) poor correlation between the two when the gate resistance effect is not considered in the FOM calculation, (b) good correlation when the gate resistance effect is taken into account. (After Chatterjee et al., Ref. 3.)

requirement of R_{sh} for a given performance degradation Δ can be derived from Eq.10.6 as

$$R_{sh} = \frac{\tau_{delay} \cdot \Delta}{a \cdot C'_{ox} \cdot (W_n^2 + W_p^2)} \tag{10.24}$$

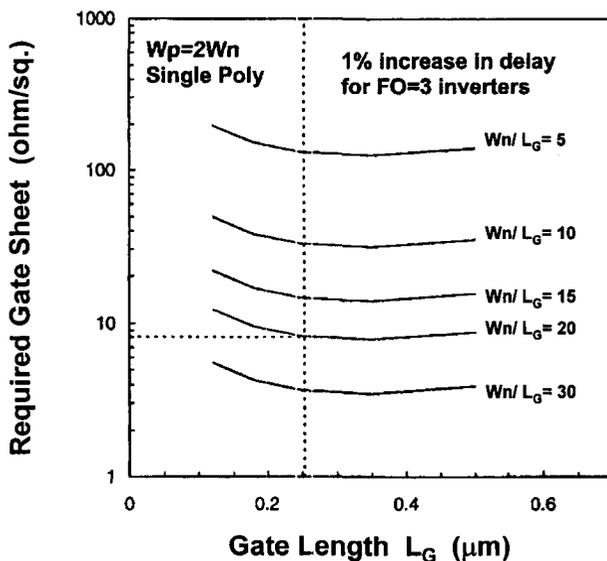


Figure 10.13 Required gate sheet resistance for an inverter delay RAM 1 as a function of gate length for several W/L_G ratios. (After Chatterjee et al., Ref. 3.)

where τ_{delay} is the desired ideal delay, and $(1 + \Delta) \cdot \tau_{\text{delay}}$ is the delay including the effect of the gate RC . For a given roadmap of τ_{delay} (shown in Fig. 10.1b) and t_{ox} , the R_{sh} requirements can be derived from this equation. For properly scaled technologies where transistor width scales proportional to L_G , the R_{sh} requirement can be fairly constant relative to L_G as shown in Figure 10.13. For example, if W_n/L_G is fixed at 20, then a R_{sh} of $\leq 8 \Omega/\text{square}$ is sufficiently low to ensure that the $\text{FO} = 3$ inverter delay will not increase by more than 1% of the ideal $R_{\text{sh}} = 0$ case.

Source-Drain Sheet Resistance and Requirements

MOSFETs with partially contacted source-drain regions are often used in ASIC standard-cell designs as well as other logic circuits to improve the layout packing density.⁴⁰ Figure 10.14a shows two commonly used source-drain contacting schemes. The fully contacted case (case A) has the least impact on transistor performance, but has the lowest packing density because unrelated metal 1 lines cannot pass over the gate area. The opposite is true for the diagonally contacted transistors (case B); the layout density is high but there is some performance impact.

The impact of the additional S/D resistance of the diagonal contacting scheme (case B) can be modeled by a distributed resistor and conductor ($R-G$) network.⁴⁰ Figure 10.14b shows the model results comparing to the experimental data. For the fully contacted case (case A), the drain current is proportional to the device width as expected. While for the diagonally contacted case (case B), drain current versus width deviates significantly from the linear relationship of the full-contact case and can be accurately reproduced by the distributed $R-G$ model.

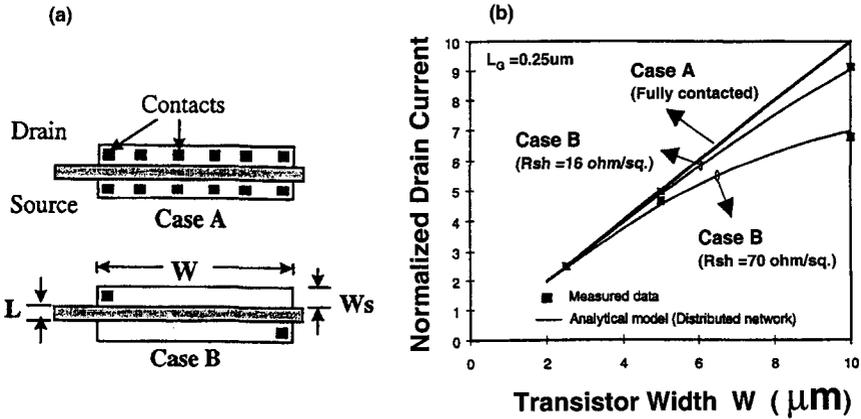


Figure 10.14 (a) Schematic layouts of transistors with two different S/D contact schemes. For layout case A (fully contacted case), the requirement on S/D region resistance is minimal; for layout case B (diagonally contact case), the requirement on S/D region sheet resistance is the highest. (After Mehrotra et al., Ref. 40.) (b) Normalized drain currents versus transistor width for the fully and diagonally contacted transistors (with two different S/D sheet resistances). The diagonally contacted transistors have lower drive currents due to higher series resistance. An analytical model can fit the data very well. (After Mehrotra et al., Ref. 40.)

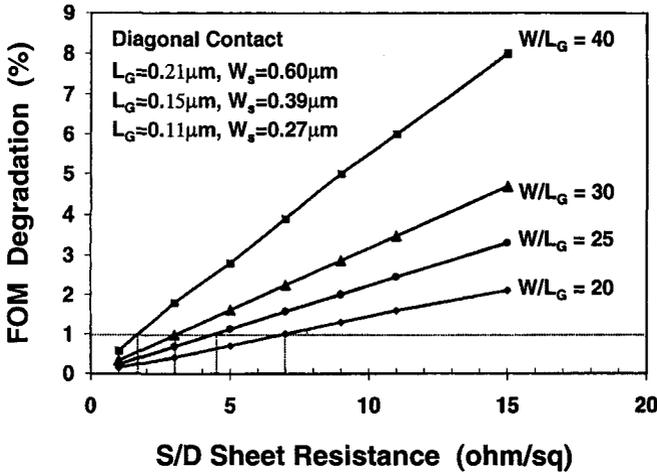


Figure 10.15 Performance FOM degradation versus the sheet resistance in the S/D region with W/L as parameters. For W/L of 20, the S/D sheet resistance needs to be lower than $7 \Omega/\text{sq}$ to prevent degradation of performance by 1%. (After Mehrotra et al., Ref. 40.)

With the help of the model, one can generate the requirements on $R_{sh}^{S/D}$. Figure 10.15 shows the percentage degradation of the FOM for diagonally contacted CMOS transistors as a function of $R_{sh}^{S/D}$ for technology nodes in the range of 0.2–0.1 μm . It turns out that since the general layout rules roughly scale with gate length L_G , the FOM degradation/ $R_{sh}^{S/D}$ relationships are fairly independent of the

technology node; for example, for a layout rule of $W/L_G = 20$, $R_{sh}^{S/D}$ needs to be $\leq 7 \Omega/\text{square}$ so that the impact on performance FOM is less than 1%.

Note that this result (i.e., $R_{sh}^{S/D} \leq 7 \Omega/\text{square}$) is almost the same as the requirement on gate sheet resistance ($R_{sh} \leq 8 \Omega/\text{square}$) discussed in the previous section. Therefore, $7 \Omega/\text{square}$ is used as the common requirement for both the gate sheet (R_{sh}) and S/D sheet resistance ($R_{sh}^{S/D}$) for any salicide technology. Note that if W/L_G is larger than 20, then the required sheet resistance will be lower accordingly. (see Figs. 10.13 and 10.15).

10.2.10 Self-Aligned Silicide (Salicide) and Raised Source–Drain

To satisfy the low-resistance requirements for the gate and source–drain regions, self-aligned silicide (or salicide) that is simultaneously formed silicide on the gate and on the source–drain regions has been widely used. Among various silicide materials studied, titanium (Ti) and cobalt (Co) silicides are most commonly used.

Salicides

TiSi₂ Salicide TiSi₂ salicide has been the most prevalent used salicide for logic applications from about $0.8 \mu\text{m}$ down to about $0.25 \mu\text{m}$ CMOS technology. In this process, a deposited Ti film is thermally reacted in a N₂ ambient to form TiSi₂ on the exposed Si areas. A higher resistivity C49 phase ($60\text{--}90 \mu\Omega \cdot \text{cm}$) forms first. The unreacted TiN or Ti remaining on the oxide or nitride (e.g., field oxide or spacer material) are then stripped, and followed by a higher-temperature anneal to transform the TiSi₂ from the C49 phase to a lower resistance ($13\text{--}16 \mu\Omega \cdot \text{cm}$) C54 phase.

Compared to other materials (e.g., CoSi₂), TiSi₂ has several advantages that result in its prevalent usage: (1) relatively low bulk resistivity of $13\text{--}16 \mu\Omega \cdot \text{cm}$ for the C54 phase TiSi₂ compared to $15\text{--}18 \mu\Omega \cdot \text{cm}$ for CoSi₂ or NiSi; (2) not very sensitive to thin native oxide on the Si surface, which greatly widens the process margin; (3) Si is the diffusing species during the TiSi₂ silicide formation steps; thus, silicide spiking is less likely to happen and the diode is more immune to reverse-bias leakage problem; (4) less Si consumption per unit volume of silicide formed (Si : silicide ≈ 0.9 for TiSi₂, 0.8 for NiSi, and 1.0 for CoSi₂); and (5) reasonably good thermal stability ($\approx 850\text{--}900^\circ\text{C}$ for rapid thermal annealing).

The major drawback of TiSi₂ salicide is the *narrow-line effect*. The sheet resistance of TiSi₂ silicided poly increases with decreasing poly width. This is because the C49-to-C54 phase transformation is more difficult for narrower poly lines. In Figure 10.16, the conventional TiSi₂ case is an example of this narrow-line effect. The reason that the narrow lines have higher sheet resistance is because the C49 to C54 phase transformation originates from the “triple points” of the C49 grain boundaries.⁴¹ Because narrower lines have less such nucleation sites, it is more difficult to transform from the C49 to the lower-resistance C54 phase. Other factors such as the film thickness, and dopants in the underlying Si have also been observed to affect the C49 to C54 phase transition.

To overcome this narrow-line effect for TiSi_2 , numerous approaches have been proposed: (1) using TiN capping,⁴² (2) using preamorphization implant (PAI),⁴³⁻⁴⁶ (3) using a Mo implant before the TiSi_2 react,^{47,48} and (4) using combination of Mo and PAI implants.⁴⁹

The TiN capping process is relatively simple: adding only TiN deposition and TiN strip steps to a regular TiSi_2 sequence.⁴² The TiN capping can be applied at the first TiSi_2 reaction step, or at the second TiSi_2 anneal step, or both. The TiN capping is believed to enhance the thermal stability and increase the resistance to agglomeration for TiSi_2 . The effectiveness of this TiN capping process has been demonstrated on 0.35- μm technology.⁴²

Using preamorphization implant (PAI) before Ti deposition to improve the narrow-line effect has been widely studied.⁴³⁻⁴⁶ As mentioned above, the larger the grain size of C49 phase TiSi_2 , the more pronounced the narrow line effect due to lack of nucleation sites. On the other hand, the grain size of the C49 TiSi_2 is known to correlate with that of the underlying polysilicon. Therefore, by using a PAI implant, one can reduce the C49 grain size and thus improve the narrow line effect. Figure 10.16 shows an example of the gate sheet resistance of a PAI processed TiSi_2 , compared to a conventional non-PAI processed case.

However, there are several drawbacks of using PAI to improve the narrow-line effect of TiSi_2 salicide. Since the PAI is done right before the Ti deposition, the source-drain regions are also amorphized at the same time. This can lead to the following problems: (1) increase of R_{SD} (particularly for p-MOS) and thus degradation of I_{drive} ,⁴⁵ (2) enhanced V_T rolloff due to PAI induced boron transient enhanced diffusion,⁴⁶ (3) degradation of n^+ to n-well isolation also due to TED,⁵⁰ and (4) enhanced p-MOS GIDL leakage.⁴⁶ Because of these issues, it is necessary to carefully optimize the PAI condition.⁴⁵

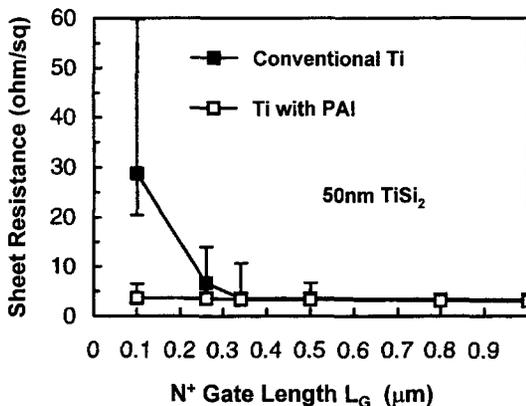


Figure 10.16 Sheet resistance versus n^+ poly gate length for 50-nm TiSi_2 . For the conventional TiSi_2 case, the sheet resistance increases sharply for gate length $\leq 0.25 \mu\text{m}$ as a result of the narrow-line effect, while for the case with PAI (preamorphization implant), the sheet resistance is nearly constant down to gate length of $\approx 0.1 \mu\text{m}$. (After Kittl et al., Ref. 44.)

Molybdenum (Mo) doping in the poly or crystalline Si has also proved effective in alleviating the narrow-line effect of TiSi_2 .^{47,48} In contrast to a regular TiSi_2 process with two RTP steps, Ti can react with Mo-doped poly and directly form the low-resistance C54 phase TiSi_2 with a one-step, low-temperature (650°C) RTP.⁴⁸ Mo is believed to serve as a catalyst that makes the transformation to C54-phase TiSi_2 much easier at low temperature. Figure 10.17 compares the narrow-line effect of several TiSi_2 processes (viz., As or Ge PAI, Mo-doping, and conventional) and a typical CoSi_2 process. The PAI cases have good narrow-line effect down to $L_G = 0.12\ \mu\text{m}$, but have the problems mentioned before. The Mo cases have reasonably good narrow-line effect and without many drawbacks, while the conven-

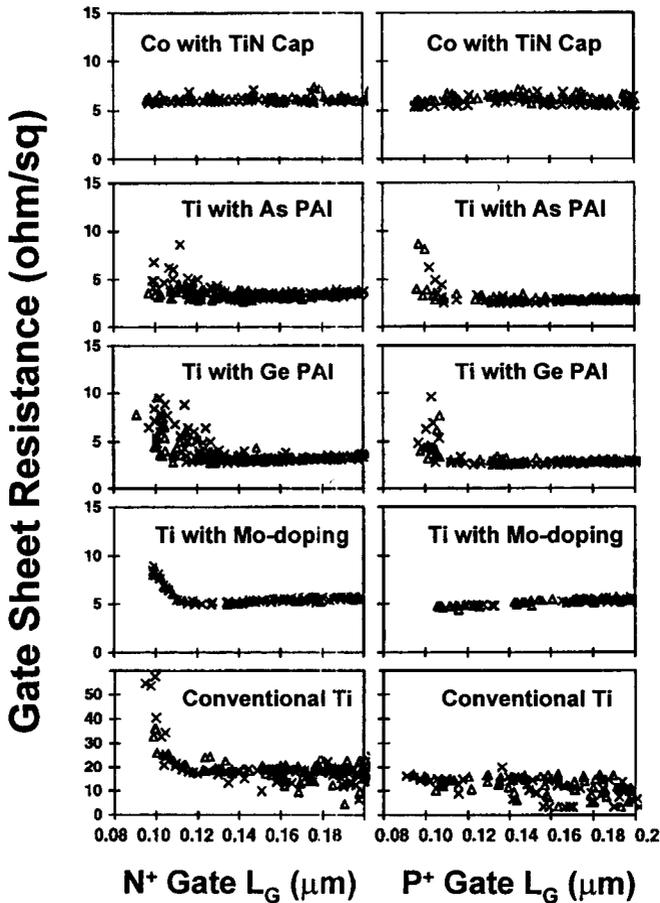


Figure 10.17 An extensive comparison of the linewidth-dependent poly sheet resistance for CoSi_2 (top two figures showing no linewidth dependence), and several variations of TiSi_2 : PAI can extend down to about $0.14\ \mu\text{m}$ (but with side effects; see text), Mo doping can be comparable to PAI without PAI-induced problems, and the conventional case has fairly high resistance for $L_G \leq 0.2\ \mu\text{m}$. (After Kittl et al., Ref. 48.)

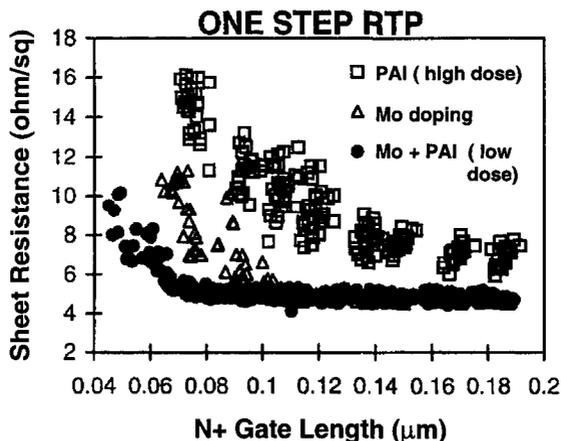


Figure 10.18 A further comparison of TiSi_2 narrow-line sheet resistance versus gate length using one-step RTP only. The Mo doping + low-dose PAI has the best resistance characteristics down to about $0.07 \mu\text{m}$. (After Kittl et al., Ref. 49.)

tional TiSi_2 has high resistance for L_G below $0.2 \mu\text{m}$. The CoSi_2 cases essentially do not have narrow-line effect, but have a sporadic diode leakage problem, which is discussed below.

Moreover, a TiSi_2 with a combination of Mo implant and PAI (Mo + PAI) has excellent narrow-line results with a one-step low-temperature RTP.⁴⁹ Figure 10.18 shows the gate sheet resistance versus L_G for the TiSi_2 with Mo + PAI, compared to a high-dose PAI and a Mo-only TiSi_2 salicide. Because only a relatively low-dose PAI is necessary, many of the drawbacks associated with the PAI process mentioned above are minimized, except that the Mo + PAI case has higher diode leakage compared to the Mo-only TiSi_2 or CoSi_2 cases.⁴⁹

CoSi_2 Salicide CoSi_2 is becoming increasingly attractive due to its linewidth-independent sheet resistance down to sub- $0.1\text{-}\mu\text{m}$ lines (see Fig. 10.17).⁵¹

The most serious problem for CoSi_2 salicide is the sporadically high diode leakage due to the nonuniform CoSi_2/Si interface or “CoSi spiking.”⁵² Several approaches were made to solve the problem, including high-temperature ($800\text{--}850^\circ\text{C}$) RTP for the second annealing.⁵² Figure 10.19 shows the effect of high-temperature RTP on CoSi_2 salicided diode leakage, which can be as low as that of a TiSi_2 salicided diode. The reason is that the CoSi_2/Si interface is smoother after the high-temperature annealing.^{52,53} In low-temperature annealing the interface is rough because Co diffusion in Co_2Si and CoSi_2 is relatively more favorable along the grain boundaries at low temperature.⁵³ Another approach consisted in high-temperature (450°C) Co deposition with in situ annealing (for about 5 min) at the same temperature.^{53,54}

Raised Source/Drain and Other Advanced Structures As discussed above, the choice of salicide material is heavily dependent upon the diode-leakage

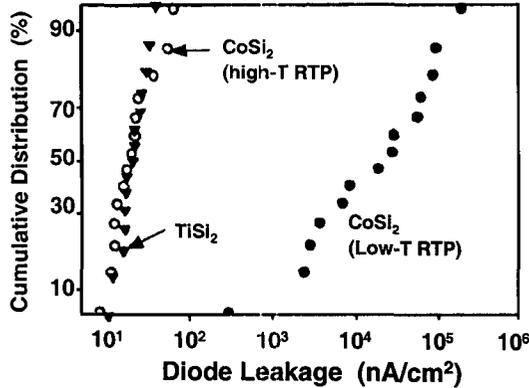


Figure 10.19 Distributions of diode leakages for high- and low-temperature annealed CoSi_2 and conventional TiSi_2 ($n^+x_j \approx 0.15 \mu\text{m}$). The low- T RTP enhances the Co diffusion along the grain boundaries (Co “spiking”) and has high diode leakage. The CoSi_2 with high- T RTP has about the same diode leakage as the conventional TiSi_2 . (After Hong et al., Ref. 54.)

characteristics, which is a function of salicide/Si interface roughness, silicide thickness, and the underlying junction depth (referred to as “deep x_j ”). The deep x_j needs to be scaled as technology progresses simply because the spacer thickness also needs to be scaled and it is necessary that the lateral diffusion of the deep x_j not to cover the S/D extensions. On the other hand, the silicide thickness pretty much is fixed to maintain the same sheet resistance (e.g., $\leq 7 \Omega/\text{square}$) for both gate and S/D regions. This means it is almost inevitable to run into the diode leakage problem in the future if we continue to scale the current device structure (i.e., salicide right after the S/D annealing). Figure 10.20 illustrates this problem for a conventional device structure.

Therefore, alternative device structures need to be studied before the issue becomes very serious. In the following sections, we discuss a few promising alternative device structures (raised source–drain, selective metal, and selective salicide) to achieve low-resistances gate and S/D regions without diode-leakage concerns.

The raised source/drain (R/SD) structure is made by using selective epitaxial-silicon growth (SEG) to “raise” up the source/drain regions.^{55–58} The poly gate may or may not have the selective silicon deposited, depending on whether the gate is exposed or covered during the selective silicon deposition. Figure 10.21a shows a schematic diagram of the raised Source/Drain transistor with gate exposed during the SEG step. The reason that diode leakage can be improved is simply because the SEG can separate the salicide/Si interface and the underlying junction depth, without increasing the effective junction depth (meaning the x_j measured from the channel). Figure 10.21b shows the diode leakage characteristics (for both diode bottom wall and edge leakage) for a n^+ diode (As-only junction). The conventional CoSi_2 (10 nm Co, low-temperature annealed, without R/SD) case has high diode

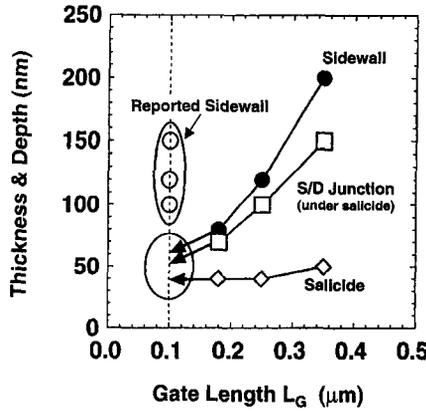


Figure 10.20 A scaling trend of sidewall thickness, S/D junction depth (under the silicide), and silicide thickness versus gate length. The silicide thickness needs to be roughly constant to maintain the same sheet resistance, while the sidewall thickness and thus S/D x_j need to scale to increase packing density. This trend indicates a serious diode leakage problem for the conventional salicide structure at gate length around 0.1 μm . (After Wakabayashi et al., Ref. 57.)

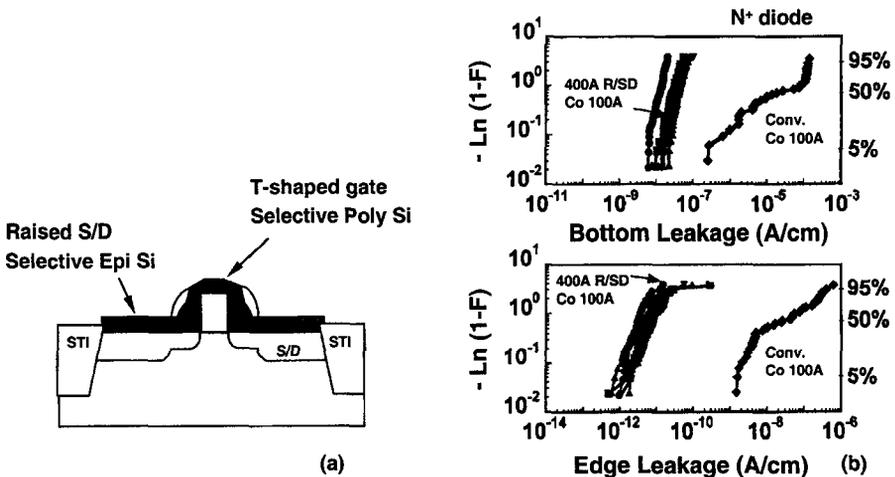


Figure 10.21 (a) A schematic diagram of the raised S/D device structure to alleviate the potential diode leakage problem shown in Figure 10.20a. In this example, the selective Si is deposited on the S/D regions as well as the gate. (b) Comparisons of diode leakage distributions for bottom wall and edge component for the raised source–drain and conventional device structures. For the conventional CoSi_2 (10-nm Co without raised S/D and with low-temperature annealing), the diode leakage is high. While the diode leakage of same CoSi_2 process is greatly improved with 40-nm raised S/D. (After Chao et al., Ref. 58.)

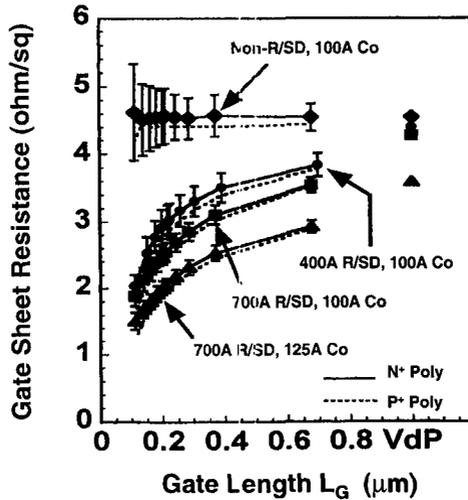


Figure 10.22 Gate resistance (CoSi_2 salicided) versus gate length for the raised S/D and conventional devices. The conventional CoSi_2 has a rough constant sheet resistance down to about $0.08\ \mu\text{m}$, while the raised S/D CoSi_2 cases have a “reverse linewidth” (i.e., lower resistance for narrower lines) effect due to the T-shaped gate (shown in Fig. 10.20*b*). (After Chao et al., Ref. 58.)

leakage as expected. With 40-nm R/SD (the circle symbol) and the same 10 nm Co process, the diode leakages are dramatically improved.

Another advantage of using the raised source/drain (with SEG on poly) is that the overgrowth of SEG on poly, forming a T-shaped gate (see Fig. 10.21*a*), can effectively increase the width and thus reduce the sheet resistance of the salicided poly (or a reverse narrow-line effect). This is very desirable from the gate resistance standpoint, as shown in Figure 10.22. The reverse narrow-line effect is due to the fact that the same amount of SEG overgrowth on the gate has a larger percentage increase in effective gate width on narrower gates than on wider ones.

Selective metal deposition is another promising way of alleviating the diode issue of the current salicide structure. In essence, with proper processing, the selective metal (e.g., tungsten) is only “selectively” deposited on the exposed Si regions (i.e., gate and S/D regions), but not on the field oxide regions.^{59,60} Because the metal layer is deposited on Si or a thin silicide layer, there is minimal Si consumption. Therefore, in principle the selective metal deposition can achieve both low resistance and low diode leakage simultaneously even when the junction depth is reduced.

Selective CVD TiSi_2 deposition potentially is another good alternative for replacing the current salicide process.⁶¹ In principle, the process can selectively deposit C54-phase TiSi_2 onto the exposed Si areas only. The Si consumption can be minimal ($\leq 10\ \text{nm}$) or none, thus the CVD TiSi_2 process is very attractive in terms of avoiding any potential diode leakage problem. Because the lower-resistance C54-phase TiSi_2 was directly deposited, rather than transformed from the higher-

resistance C49 phase as in conventional TiSi₂ salicidation, nearly constant 3 Ω/square of gate resistance down to a 0.18-μm poly line was demonstrated.⁶¹ If manufacturing equipment and process are developed for this technique, this CVD TiSi₂ process will be of importance in the future.

10.2.11 Poly Depletion

Poly depletion is also a major device design issue, particularly for sub-0.25-μm dual-gate (n⁺ gate for n-MOS and p⁺ gate for p-MOS) CMOS, which is prevalently used in the high-performance CMOS technologies because both the n-MOS and p-MOS can be maintained in the surface channel operation mode which is good for short-channel effect and thus device scaling. However, the dual-gate CMOS also requires some tradeoffs, most notably boron penetration (for p-MOS) and poly depletion.^{62,63} Boron penetration refers to the boron in the p⁺-doped poly diffusing through the gate oxide, which is a poor diffusion barrier for boron, and affects the V_{TP} , transistor performance, as well as reliability. The schemes for alleviating the boron penetration are discussed in more detail in the following section. The poly depletion problem for the dual-gate CMOS is due to the fact that when gate bias is applied to turn the transistor on, the electrical field in the gate oxide will not only invert the substrate but also penetrate into the gate electrode and deplete the gate poly. Therefore, poly depletion is dependent on the dopant concentration in the poly (and close to the poly/oxide interface) and the gate oxide thickness. Poly depletion is obviously an issue for high-performance CMOS because it reduces the drive current of the transistor.⁶⁴ Although for a high fan-out circuit the performance impact may not be very large because both drive current and gate capacitance are reduced by the same factor, lower drive current is certainly a critical issue for driving large interconnect capacitance (see Eq. 10.6) because the signal delay time in the interconnect network has become the performance bottle neck for high-performance VLSI circuits.⁶³

For a dual-gate CMOS, the relationship between poly depletion and poly concentration can be expressed as:

$$t_{\text{ox}}^2(\text{inv}) = t_{\text{ox}}^2(\text{acc}) + \frac{2\epsilon_{\text{ox}}^2 V_G}{q\epsilon_{\text{Si}}N_{\text{poly}}} \quad (10.25)$$

where $t_{\text{ox}}(\text{inv})$ is the oxide thickness measured in the inversion region ($= \epsilon_{\text{ox}}/C'_{\text{inv}}$, where C'_{inv} is the gate capacitance per unit area in the inversion region), $t_{\text{ox}}(\text{acc})$ is equal to $\epsilon_{\text{ox}}/C'_{\text{acc}}$, ϵ_{ox} and ϵ_{Si} are the dielectric constant of SiO₂ and Si, respectively; V_G is the gate bias, and N_{poly} is the poly concentration. Figure 10.23a shows the dependence of poly depletion, defined as the ratio of inversion capacitance and accumulation capacitance $C_{\text{inv}}/C_{\text{acc}}$, as a function of gate oxide thickness and poly concentration.² Figure 10.23b shows some typical n-MOS C - V curves with a reasonable n⁺ poly concentration of $7 \times 10^{19} \text{ cm}^{-3}$ and t_{ox} of 2 or 3 nm. For the $t_{\text{ox}} = 2 \text{ nm}$ case, the C_{inv} can be less than 70% of of the C_{acc} at 1 V gate bias, which defeats the purpose of using the thin oxide.

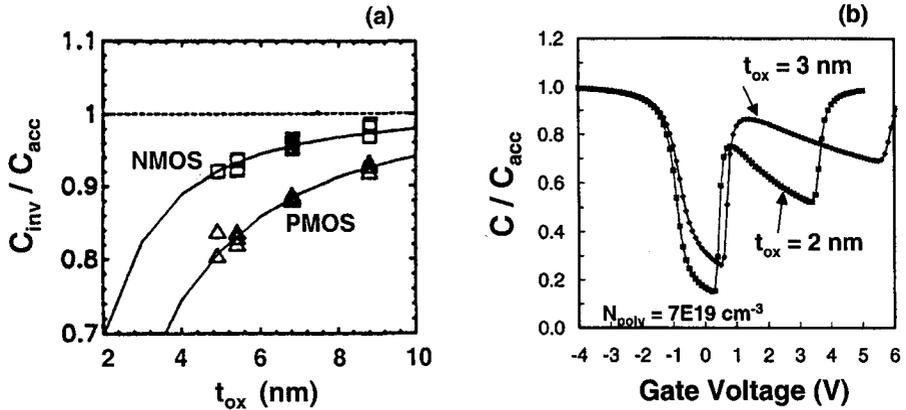


Figure 10.23 (a) C_{inv}/C_{acc} (ratio of inversion to accumulation gate capacitance) as a function of t_{ox} , showing that the poly depletion problem becomes progressively more serious with decreasing t_{ox} . The solid curves are calculated results using $N_{poly} = 8.3 \times 10^{19}$ and $2.9 \times 10^{19} \text{ 1/cm}^3$ for n^+ and p^+ poly, respectively. (After Rodder et al., Ref. 2.) (b) n-MOS CV curves for very thin ($2-3$ nm) t_{ox} (with fixed $N_{poly} = 7 \times 10^{19} \text{ cm}^{-3}$), showing that the excessively large poly depletion is a serious issue at t_{ox} around 2 nm. (After Chapman, Ref. 63.)

There are two approaches to alleviate the poly depletion problem: (1) polysilicon or poly-SiGe engineering and (2) using ϵ metal gate.

For conventional polysilicon dual-gate CMOS, increasing the implant dose is the simplest approach to relieve poly depletion. However, because of the poly grain boundaries and the different dopant segregation at the poly/SiO₂ interface, increasing the implant dose does not necessarily increase the dopant concentration proportionally. Also, unless the gate dielectrics has a good resistance to boron penetration so that one can dope the p^+ gate very heavily, simply increasing the boron dose usually cannot solve the problem of poly depletion.¹⁶

In addition to polysilicon, polycrystalline SiGe has also been used as the gate material because Si_{1-x}Ge_x has higher boron activation at a given temperature.^{65,66} With a fixed boron implant of $4 \times 10^{15} \text{ cm}^{-3}/20 \text{ keV}$ and a furnace anneal of 900°C for 40 min, Si_{0.6}Ge_{0.4} has two times (80 vs. 40%) the boron activation of poly Si,⁶⁵ which translates into higher C_{inc}/C_{acc} for the SiGe. Noted that the work function of p^+ SiGe decreases with increasing Ge content, which must be considered in the transistor design to set the V_{TP} correctly.^{55,66}

It is becoming increasingly difficult to satisfy the constraints of poly depletion and boron penetration. While short-term solutions are being generated and implemented, a relatively longer term fix is to use metal gate to replace the current poly gate. The metal gate was used previously in $\geq 3 \mu\text{m}$ p-MOS or n-MOS technology, and it becomes attractive again⁶⁷⁻⁷¹ because it has the following advantages: linewidth-independent sheet resistance,⁷⁰ compatible with low-temperature high-K dielectrics (Ta₂O₅, BST, etc.),⁶⁸ and it is free from gate depletion⁷⁰

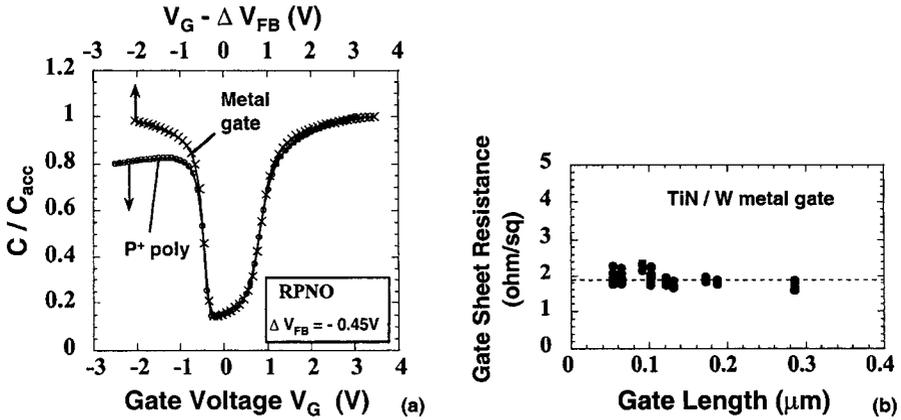


Figure 10.24 (a) A comparison of p-MOS CV curves using W/TiN metal gate versus conventional p⁺ poly gate ($t_{ox} = 3.3$ nm). The C_{inv}/C_{acc} of the W/TiN metal gate is close to 1 (at $V_G = -2$ V). Note that the Fermi level of TiN is close to the Si midgap, and a workfunction shift of 0.45 V has been applied to the metal gate CV. (After Hu et al., Ref. 70.) (b) The gate sheet resistance of W/TiN metal gate as a function of L_G showing the roughly constant 2 Ω /square of sheet resistance down to 0.05 μm . (After Hu et al., Ref. 70.)

and boron penetration problems. However, before a metal gate can be utilized, several challenges need to be overcome such as maintaining good thin gate oxide integrity and interface quality,⁶⁷ and designing transistor with the workfunction of the metal used.

The poly depletion issue associated with the current n⁺ and p⁺ gate can be solved by using a W/TiN gate stack on 3.3 nm SiO₂. Figure 10.24a compares the high-frequency CV of a p⁺ poly-gate p-MOS with the CV of a W/TiN metal-gate p-MOS ($t_{ox} = 3.3$ nm).⁷⁰ The poly depletion, C_{inv}/C_{acc} , is about 80% for the p⁺ poly, while it is close to 100% for the W/TiN gate, indicating no gate depletion. The sheet resistance of the W/TiN metal gate is independent of linewidth; roughly 2 Ω /square for the case shown in Figure 10.24b.

In addition to the metal-gate studies utilizing a conventional process sequence, a “replacement metal gate” process was also proposed and implemented recently.⁷¹ In essence, in the replacement gate process, the S/D formation and anneal are done before the gate dielectric formation and electrode deposition. This structure is considered essential for incorporating an unconventional gate dielectric (e.g., high- k materials) and electrode metal (e.g., Al), which cannot withstand the S/D annealing temperature, which is usually $\geq 850^\circ C$.

10.2.12 Gate-to-Drain Overlap Capacitance C_{GD}

Gate-to-drain overlap capacitance (C_{GD}) is an important device parameter for both analog and digital applications. We discuss analog applications later in the sections on high-frequency operation. In these applications the C_{GD} is a Miller capacitor,

meaning that the effective capacitance at the output node is equal to the product of C_{GD} and the gain of the transistor. For digital applications, the effect of C_{GD} is not as large as the Miller effect, but still larger than the normal gate capacitance. In essence, the gate capacitances (C_{gate} , which represents C_{gate}^n or C_{gate}^p) in Eq. 10.4 is expressed as

$$C_{\text{gate}} \approx C'_{\text{inv}} \cdot \left(L_{\text{eff}} + \frac{3}{2} \cdot \text{TLD} \right) \cdot W \quad (10.26)$$

where C'_{inv} is the inversion capacitance per unit area, L_{eff} is the effective channel length, W is the transistor width, and TLD is the total source and drain lateral diffusion for the S/D diffusion region underneath the gate, and can be expressed as

$$\text{TLD} = L_G - L_{\text{eff}} \quad (10.27)$$

$$C_{GD} \approx C'_{\text{inv}} \cdot \frac{\text{TLD}}{2} \cdot W \quad (10.28)$$

As mentioned before, for a fixed L_G , there is not much freedom to adjust L_{eff} because of the constraints of drive current, short-channel effects, and hot-carrier reliability.

10.2.13 Gate Dielectrics for Suppressing Boron Penetration

As was briefly mentioned in the section on poly depletion (Section 10.2.11), boron penetration is a major issue for dual-gate CMOS,⁶² because SiO_2 is a poor diffusion barrier for boron. Boron penetration will create many undesirable effects, such as shift V_{TP} , degrade p-MOS subthreshold swing, decrease hole mobility, increase oxide early failures and positive charge trapping,^{72,73} and reduce p-MOS drive current.^{18,73} Moreover, as CMOS technology scales, the boron penetration problem will become worse. This is because in order to maintain a certain acceptable level of p^+ poly depletion, higher boron concentration at the poly/gate oxide interface is needed as t_{ox} continues to scale;² but a thinner t_{ox} has lower resistance to boron penetration.⁷⁴ How to improve the gate oxide so that it can block boron diffusion and enhance p-MOS performance has been a major challenge for gate oxide engineers for many years.

Nitrogen-doped SiO_2 (or nitrided oxide) has been studied for a long time as a replacement for pure SiO_2 ,^{76–81} partially because it is known to be a better boron diffusion barrier than pure SiO_2 . In addition, the nitrided oxide also has characteristics that are beneficial to a high-performance CMOS logic technology: (1) higher resistance to channel-hot-carrier-induced degradation, (2) higher resistance to Fowler–Nordheim (FN) tunneling-induced interface trap generation, (3) lower FN stress-induced leakage current, (4) better charge trapping characteristics, and (5) better radiation hardness. The only common drawbacks of using nitrided oxide for logic CMOS are that hole mobility and low-field electron mobility are

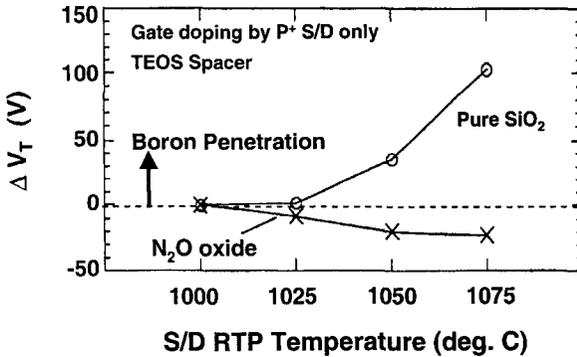


Figure 10.25 The resistance to boron penetration at various S/D RTA temperatures for a N_2O nitrided and a conventional thermal oxide ($t_{ox} \approx 3.3$ nm, oxide spacer was used). The N_2O oxide has high resistance to boron penetration to 1100°C , while the pure oxide has significant boron penetration at 1025 – 1050°C . (After Hu et al., Ref. 64.)

usually lower for nitrided oxides depending on the amount of nitridation, and that the saturation velocity is lower for nitrided oxides.⁷⁵

In the following sections, several approaches, mostly using nitrided oxides or nitride, to increase the resistance to boron penetration are briefly reviewed.

Nitridation of SiO_2 using N_2O (nitrous oxide) or NO (nitric oxide) are by far the most widely studied gate oxide process.^{76–78} Many sub- $0.25\text{-}\mu\text{m}$ dual-gate CMOS technologies reportedly chose N_2O or NO grown or annealed oxides as the gate dielectrics.^{14,15} Figure 10.25 compares the boron penetration of a 3.3-nm nitrided oxide grown in an N_2O ambient and followed by a furnace reoxidation and a 3.8-nm pure oxide p-MOS devices. This figure shows that the 3.3-nm N_2O nitrided oxide can block boron penetration within a reasonable range of RTP temperature (ranging from 1000 to 1100°C). Note that the spacer material (oxide or nitride) has an impact on boron penetration,⁶⁴ because hydrogen enhances boron penetration⁷⁴ and the nitride spacer deposition is done in a hydrogen-rich ambient.

Generally the impact of nitridation on carrier mobility, particularly hole mobility, is a tradeoff of the N_2O or NO nitrided oxide. Figure 10.26 compares typical hole mobility vs. effective electric field plots of an N_2O nitrided oxide, a pure oxide, and an oxide with nitrogen implant before oxide growth.⁷⁹ Compared to pure oxide, the N_2O nitrided oxide case has lower hole mobility for effective field ≥ 0.6 MV/cm.

The scheme of implanting nitrogen into the Si substrate before gate oxidation has been proposed as an alternative way of fabricating nitrided oxide.⁷⁹ Excellent results including high resistance to boron penetration were reported. For a nitrogen dose of $2 \times 10^{14} \text{ cm}^{-2}$, a 2.5-nm nitrided oxide with thermal anneal of 800°C furnace anneal for 40 min and RTA at 1050°C for 10 s can block boron penetration from the p^+ -poly.⁷⁹ However, care is required to use this technique. A high-dose ($\geq 1 \times 10^{14} \text{ cm}^{-2}$) nitrogen implant can significantly degrade the oxide charge-to-breakdown (Q_{BD}) distribution and reduce hole mobility.⁶⁴

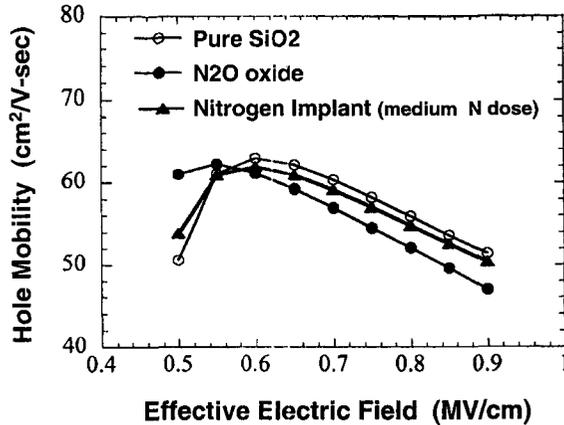


Figure 10.26 A comparison of effective hole mobility for the N_2O oxide shown in Figure 10.25, a thermal oxide grown on nitrogen-implanted (N I/I) substrate, and a conventional thermal oxide. Compared to the conventional oxide, the N I/I oxide has slightly lower hole mobility and the N_2O oxide has the lowest hole mobility. (After Hu et al., Ref. 64.)

The remote-plasma nitrided oxide (RPNO) technique uses a remote high-density N_2 plasma to nitridize the top SiO_2 surface (≈ 1 nm) of a thin thermal oxide at low temperature.⁸⁰ Because the nitridation is confined to the top surface of the gate oxide, it is ideal for blocking boron penetration while maintaining the SiO_2/Si interface properties including carrier surface mobility.⁷⁷ Excellent resistance to boron penetration was reported by using this RPNO with equivalent SiO_2 thickness ranging from 4 to 3 nm.^{19,80,81} It was also shown that both electron and hole mobility are hardly changed, even with a high concentration of [N] ($>15\%$ atomic percent), compared to pure SiO_2 .⁸¹

The jet-vapor-deposited (JVD) nitride^{82,83} is a “paradigm shift” from the pure or nitrided SiO_2 that the IC industry has been using since the late 1970s. The chemical composition of the JVD nitride (Si_3N_4) is the same as a commonly used CVD nitride, but the difference is that the deposition is done in a jet vapor environment at very low ($\leq 200^\circ C$) or even room temperature.⁸² The interface trap density D_{it} and bulk trapping properties of the JVD nitride are surprisingly good—much better than the regular CVD nitride and comparable to those of a thermal SiO_2 . It was shown that CMOS devices using JVD nitride as gate dielectrics have long-channel transconductances comparable to those of pure oxide cases,⁸³ which demonstrates the feasibility of using the JVD nitride in an ULSI technology.

The JVD nitride has an additional advantage—it can reduce the large direct-tunneling leakage current of a pure SiO_2 for an equivalent oxide thickness (≤ 3 nm).^{82,83} This is because the dielectric constant of Si_3N_4 (≈ 7.0) is about twice that of the SiO_2 (≈ 3.9). Therefore, for a given equivalent oxide thickness (≤ 3 nm), a JVD nitride film can be about twice as thick as a SiO_2 film, which is known to be very effective in reducing the direct-tunneling leakage.⁸⁴

10.3 LOW VOLTAGE/LOW POWER CONSIDERATIONS FOR DIGITAL APPLICATIONS

As the functionality and speed of state-of-the-art integrated circuits continue to increase at a staggering pace, the power consumption of the chips is also becoming a real concern for both portable or desktop electronics systems.^{85,86} For portable electronics, one of the major concerns is battery lifetime. Usually when the speed performance is above a certain acceptable level, it may be less important to continue to pursue faster performance than it is to lower the power consumption to prolong the battery. On the other hand, for the desktop systems, the demand for higher speed is much more than for the portable systems and there is no concern about battery. However, there is still an upper limit to the total power consumption because of the heat-conduction limitation for heat sinks. If the heat generated by the chip is higher than the heat conducted away by the heat sink, then the chip temperature will rise, which, in turn, will degrade the circuit performance.

The total power consumption of an integrated circuit can generally be broken down into three components: active switching power (P_a), leakage power (P_{leak}), and short-circuit power (P_{sc}).

$$\text{Total power consumption} = n \cdot [a \cdot (C \cdot V_{DD}^2 \cdot f + P_{sc}) + I_{OFF} \cdot V_{DD}] \quad (10.29)$$

where n is the number of transistors on a chip, a ($0 \leq a \leq 1$) is the activity factor, C is the total loading capacitance, f is the operating or switching frequency, I_{OFF} is the OFF-state leakage current, and P_{sc} , the short-circuit power, is the leakage power associated with the transient current when both n-MOS and p-MOS devices are conducting during switching.⁸⁷ For each individual transistor, the active power P_a is usually dominated by the term $(a \cdot C \cdot V_{DD}^2 \cdot f)$ and the leakage power P_{leak} , by the term $(I_{OFF} \cdot V_{DD})$.

Figure 10.27a shows the SIA Roadmap of power consumption/technology node. For the handheld battery-operated applications, the chip power consumption is

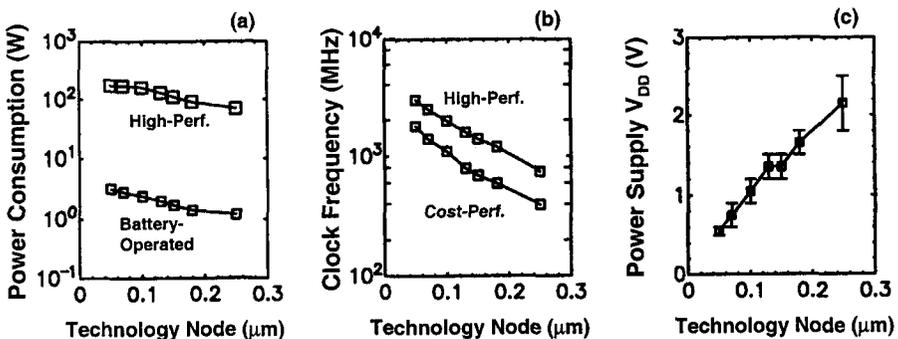


Figure 10.27 SIA roadmap on chip (a) power consumption, (b) clock frequency, and (c) power supply voltage V_{DD} . As technology scales, the power consumption (or dissipation) increases slightly, while clock frequency and number of transistors (not shown) increase rapidly, which demands a rapid reduction of V_{DD} . (From 1997 SIA Roadmap.)

limited to 1–3 W, while for the high-performance desktop applications the chip power consumption is limited to 70–130 W. Figure 10.27*b* shows the SIA *Roadmap* of the chip frequency to be used for each generation of logic technology. As can be seen, the clock frequency is forecast to increase at a staggering pace, especially for the high-performance applications. Therefore, if we assume that the P_{sc} and P_{leak} terms remain roughly the same, while the number of transistors and the clock frequency will increase drastically (see Fig. 10.27*b*), then lowering the power supply voltage V_{DD} is necessary in order to meet the power consumption goal. Figure 10.27*c* shows the V_{DD} /technology node, where one can see that V_{DD} decreases monotonically as technology scales. In fact, by simply scaling the V_{DD} according to the roadmap, a given circuit most likely cannot meet both the clock frequency and power consumption roadmaps. This means that the system architecture and circuit techniques also need to be improved so that a circuit can run faster and consume less power for a given device technology. In this chapter, we address the device approaches for achieving low-voltage / low-power applications only.

10.3.1 Device Approaches for Low-Voltage/Low Power Applications

The key challenge of low-power technologies is to meet both the speed performance (FOM) and power consumption targets at reduced V_{DD} . For conventional CMOS, lowering V_{DD} to reduce power consumption is a direct tradeoff with speed because if V_T is fixed and V_{DD} is reduced, the I_{drive} will be lower and performance will be reduced. On the other hand, if V_T is lowered to increase I_{drive} , then the I_{OFF} will increase drastically. For a V_T reduction of every 80–90 mV, the I_{OFF} will increase by an order of magnitude, and so will the standby power. Therefore, all approaches must address the I_{drive}/I_{OFF} tradeoff.

Besides this key challenge of I_{drive} and I_{OFF} , all the other device design considerations for high-performance CMOS (e.g., t_{ox} , short-channel effect, x_j , R_{SD} , gate sheet resistance, salicide, poly depletion, gate dielectrics) are all applicable to the low-voltage/low-power CMOS transistors design. It is also important to have as many common processing steps as possible between the two kinds of CMOS technology, so that they can easily be manufactured in the same plant.

Many of the device approaches are related to SOI, which is detailed in Chapter 5. To avoid duplication, the SOI devices are mentioned only briefly.

Dual- V_T CMOS

The most straight-forward and widely studied way of engineering the I_{drive} and I_{OFF} at low V_{DD} is simply to use dual- V_T values; one with nominal V_T and the other with lower than nominal V_T (referred to as *low- V_T*), on the same chip.^{88–90} The low- V_T transistors are used in the critical paths to improve the speed performance, while the nominal- V_T transistors are used in the rest of the circuit (SRAM array, etc.) to save standby power. In other words, the low- V_T transistors are used in the relatively few critical paths, where higher I_{drive} is important to improve the speed of the chip, and where the higher I_{OFF} generated by these low- V_T transistors will increase the overall

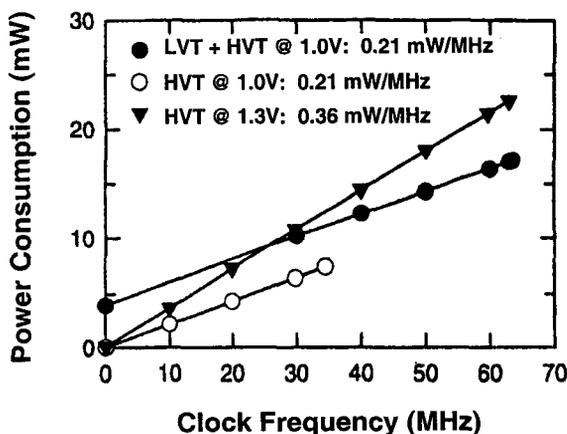


Figure 10.28 Power consumption versus clock frequency for a 1-V DSP chip using a dual- V_T (a low V_T and a standard high V_T) 0.35- μm CMOS technology. For the dual- V_T case, the DSP chip can be operated at 67 MHz at 1 V, while the high- V_T chip can run only to 35 MHz. On the other hand, if a higher power supply voltage (1.3 V) is applied to the high- V_T chip to operate at the same 67 MHz, then it will consume more power than the dual- V_T chip. (After Lee et al., Ref. 91.)

standby current only marginally. Additionally circuit techniques such as shutting off the power supply of a portion of the circuit during standby (or sleep) mode can be applied to further reduce the standby power.

Functional low-power circuits using the dual- V_T CMOS technology have been demonstrated. For example, a 1-V DSP has been implemented⁹¹ using a 0.35- μm dual- V_T technology with $L_G = 0.25 \mu\text{m}$ with reasonably good speed performance (63 MHz) at $V_{DD} = 1 \text{ V}$. Figure 10.28 compares the power consumption versus clock frequency of the 1-V DSP using either the dual- V_T (denoted as LVT + HVT) or a single- V_T (HVT) CMOS technology.⁹¹ At a fixed V_{DD} of 1 V, the dual- V_T DSP has almost 2 times the speed (63 MHz versus 35 MHz) of the HVT-only chip. On the other hand, if the HVT-only chip is allowed to have a higher V_{DD} value (e.g., 1.3 versus 1.0 V), then at 63 MHz the dual- V_T DSP chip has 32% lower power consumption than the HVT-only chip. The higher leakage power (at clock = 0 MHz) for the dual- V_T DSP chip is due to the higher I_{OFF} of low- V_T transistors, and can be reduced by circuit technique to shut off the leaky paths during standby mode.⁹¹

Low- V_T CMOS with an Adjustable Substrate Bias

This approach is really more of a circuit-design approach than a device approach. Single low- V_T transistors are used for all transistors on a chip. During the active operation mode, the low- V_T transistors can provide large I_{drive} for fast operation. However, during the standby mode, reverse-substrate and well biases are generated by on-chip generators to raise the V_T of the transistors and therefore reduce the I_{OFF} and, thus, standby power consumption.⁹² Since it is known that the L_G dependence of the body effect is different for the pocket-implanted or the conventional (non-

pocket-implanted) devices, the pocket-implanted devices (with almost constant body effect versus L_G) are better for this application to reduce the I_{OFF} with a given substrate bias. Because the dynamic substrate bias generation depends on the system operation mode, the associated circuit complexity is naturally the most challenging design task for this approach.

Dynamic Threshold Voltage CMOS (DTMOS)

A dynamic threshold voltage MOS (DTMOS) device, achieved by tying the gate to the substrate (see Fig. 10.29a), has been proposed for ultra-low-power applications.^{93–95} Let's use an n-MOS transistor to understand the dynamic V_T operation. During the ON state, the positive V_G (≤ 0.7 V) is also applied to the substrate to forward-bias the substrate slightly. Because of the body effect, the V_T at the ON state is lower than the nominal V_T for a grounded substrate and, thus, results in higher I_{drive} . During the OFF state, the V_T increases to its nominal value and, thus, low I_{OFF} can be obtained. Typical I - V curves of the DTMOS transistor are shown in Figure 10.29b, demonstrating the transistor's capability of achieving higher I_{drive} at a given I_{OFF} compared to a standard MOS transistor.

The V_{DD} of the DTMOS must be limited to 0.6–0.7 V, so that power is not wasted by a heavily forward-biased source-to-substrate diode. Because the body capacitance is added to the gate capacitance, SOI is naturally better for this dynamic V_T implementation because of its much lower body capacitance. Nonetheless, this dynamic V_T transistor can also be implemented in bulk CMOS using a triple-well construct with trench isolation.

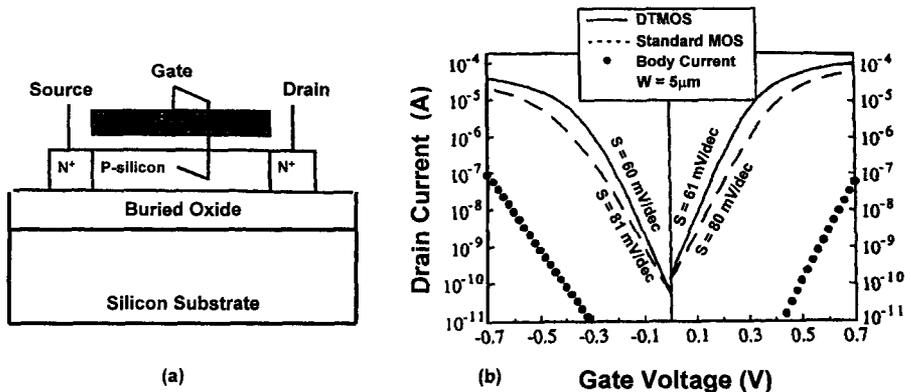


Figure 10.29 (a) The schematic diagram of a dynamic threshold MOS (DTMOS) made by connecting the gate to the body of the transistor for ultra-low-power applications. (After Assaderaghi et al., Ref. 93.) (b) IV characteristics of the DTMOS transistor. Because the body is connected to the gate, the subthreshold swing can achieve the ideal 60-mV/decade value, which effectively improved the I_{drive} at a given I_{off} compared to the conventional MOS device. The V_{DD} must be limited to about 0.6 V, to prevent excessive forward-bias diode leakage. (After Assaderaghi, Ref. 93.)

Fully or Partially Depleted SOI CMOS

Partially depleted (PD) or fully depleted (FD) SOI CMOS have been regarded as good candidates for low-power applications⁹⁵ because of the smaller subthreshold swing (for FD SOI) and lower junction capacitance (C_j), compared to a bulk CMOS. The smaller subthreshold swing can help achieve a higher I_{drive} at a given V_{DD} , or achieve a reasonable I_{drive} at reduced V_{DD} , compared to an equivalent bulk CMOS with the same I_{OFF} . The lower C_j helps to reduce the total loading capacitance (see Eq. 10.4). Both sharper subthreshold swing and lower C_j are beneficial for low-power applications. For more detailed discussion on SOI, please refer to Chapter 5.

SOI on Active Substrate (SOIAS)

SOI on active substrate (SOIAS) is a relatively new device structure for ultra-low-power applications.⁹⁶ It utilizes an SOI device with an additional electrode under the back-gate oxide (see Fig. 10.30a). In essence, during active operation a positive and negative back-gate bias can be applied to n-MOS and p-MOS, respectively, to lower the front-gate V_T and, thus, increase the I_{drive} . During standby operation the back-gate voltage can be set to zero and the front-gate V_T will rise to its nominal value, and have lower I_{OFF} . Figure 10.30b shows that with 3 V of back-gate bias (across a 100-nm back-gate oxide), the I_{OFF} can be modulated by four orders of magnitude and the I_{drive} can be modulated by a factor of 1.8 at $V_{DD} = 1$ V.

10.4 CUTOFF AND MAXIMUM OSCILLATION FREQUENCIES

The number of wireless subscribers worldwide is expected to reach 400 million by the year 2000. The global wireless market reaches more than 160 million, growing at a rate of 41% over the period from June 1996 to June 1997.⁹⁷ With this pace in the

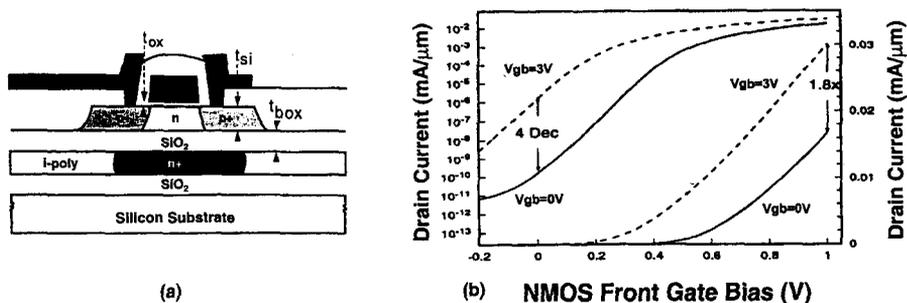


Figure 10.30 (a) The schematic diagram of a SOI with active substrate (SOIAS). The buried electrode underneath the SOI buried oxide can be biased to shift the V_T of the MOSFET on the surface. (After Yang et al., Ref. 96.) (b) IV characteristics of a n-MOS SOIAS device with and without a 3-V buried electrode bias. The 3-V back-gate bias can modulate I_{off} by four orders of magnitude and I_{drive} by a factor of 1.8. Similar results for p-MOS were also demonstrated. (After Yang et al., Ref. 96.)

mobile and satellite communications, the need for RF power transistors has grown rapidly. At the time of writing AlGaAs/GaAs HBTs (heterojunction bipolar transistors), GaAs MESFETs (metal semiconductor field-effect transistors), and the AlGaAs/InGaAs/GaAs PHEMT (pseudomorphic high-electron-mobility transistor) constitute the bulk of the market. Although GaAs parts are traditionally viewed as a high-cost/low-yield option, they have continued to maintain a strong presence in wireless applications because of the superior electrical properties inherent in the GaAs semiconductor and in the GaAs-based heterojunctions.⁹⁸ Nonetheless, there has been an industry trend to utilize silicon MOSFET for RF applications, which has the obvious advantage of integrating the baseband components as well as other digital signal processing capabilities on a single chip. The RF properties of the MOS transistors are the main focus of the remainder of this chapter. However, as alluded to, several other technologies such as HBT, MESFET, and PHEMT are being used and probably will be used in the near future. The equations discussed in the following, those based on the MOSFETs, can be applied (with some modification) to other field-effect transistors such as MESFETs and HEMTs. The related equations for bipolar transistors, either homojunction or heterojunction, have been detailed elsewhere.⁹⁹

Many high-frequency properties of a transistor can be understood with the analysis of its small-signal parameters, such as the y (admittance) and s (scattering) parameters. From a physics standpoint, it is conceptually straightforward to derive the MOSFET's y parameters by solving the charge continuity equation in the channel.¹⁰⁰ However, a direct measurement of the y parameters is difficult, if not impossible, especially at high frequencies. In practice, the s parameters based on the transmission and the reflected powers are measured. These parameters are then converted to y parameters when a comparison between the measurement results and the theoretical formulation is desired.

A solution to the continuity differential equation can be obtained with a parabolic-cylindrical function representation,¹⁰¹ with a Bessel series expansion,^{102,103} by a power-series self-consistency technique,¹⁰⁴ or by an iterative substitution technique.^{105,106} All of these methods are capable of producing the exact solution. However, they are all so algebraically intensive that none of the cited references sought for the exact solution. The most accurate solution among these references retained at most the second-order frequency terms. This solution, summarized elsewhere,¹⁰⁷ represents a major improvement over a first-order derivation that neglects the transit-time effects. Because the solution keeping only the first-order terms is referred to as the quasistatic (QS) solution,¹⁰⁸ we shall refer the solution with second-order frequency terms as the *non-quasi-static* (NQS) solution. It is generally accepted that QS solution is a good approximation to the NQS solution when the operating radian frequency (ω) is much lower than the intrinsic frequency of the transistor (ω_0 , to be discussed later). Based on either the QS or the NQS solution, the quasistatic y parameters or the NQS y parameters are obtained, respectively. We shall work mostly with the QS y parameters because they are simpler in form yet capture many important device properties. The occasional need to invoke the more complicated NQS forms will be mentioned when necessary.

A formal definition of the y parameter is

$$y_{xy} = \left. \frac{\partial i_x}{\partial v_y} \right|_{\text{node voltages other than } v_y \text{ are set to zero}} \quad (10.30)$$

where x and y refer to any of the four transistor terminals: gate (g), drain (d), source (s), and bulk (b). For example, y_{dg} denotes the ratio of the small-signal current at the drain node with respect to a small-signal voltage at the gate, while the drain, source and the bulk small-signal voltages are maintained at zero. There are a total of 16 y parameters, given by

$$\begin{aligned} y_{gg} &= j\omega C_{gg} & y_{dg} &= g_m - j\omega C_{dg} \\ y_{gd} &= -j\omega C_{gd} & y_{dd} &= g_d + j\omega C_{dd} \\ y_{gs} &= -j\omega C_{gs} & y_{ds} &= -g_d - g_m - g_{mb} - j\omega C_{ds} \\ y_{gb} &= -j\omega C_{gb} & y_{db} &= g_{mb} - j\omega C_{db} \\ y_{sg} &= -g_m - j\omega C_{sg} & y_{bg} &= -j\omega C_{bg} \\ y_{sd} &= -g_d - j\omega C_{sd} & y_{bd} &= -j\omega C_{bd} \\ y_{ss} &= g_d + g_m + g_{mb} + j\omega C_{ss} & y_{bs} &= -j\omega C_{bs} \\ y_{sb} &= -g_{mb} - j\omega C_{sb} & y_{bb} &= j\omega C_{bb} \end{aligned} \quad (10.31)$$

In these expressions, g_m is the mutual transconductance, g_d is the drain conductance, and g_{mb} is the bulk mutual transconductance. C_{xy} denotes the capacitance between any two nodes of the transistor. Their definitions are

$$g_m = \frac{\partial I_D}{\partial V_{GS}}; \quad g_d = \frac{\partial I_D}{\partial V_{DS}}; \quad g_{mb} = \frac{\partial I_D}{\partial V_{BS}}; \quad C_{xy} = \delta_{xy} \frac{\partial Q_x}{\partial V_y} \quad (10.32)$$

where Q_x is the charge at terminal x . δ_{xy} is a function whose value is equal to 1 if $x = y$ and -1 if otherwise. Not all of the 16 y parameters are required to characterize the transistor; only 9 of the 16 parameters are linearly independent. This is because the sums of the y parameters in any row and in any column of the y -parameter matrix need to be zero (e.g., $y_{gg} + y_{gd} + y_{gs} + y_{gb} = 0$).

In the common-source configuration, all of the terminal voltages are taken with respect to the source terminal. The gate and drain currents are expressed as

$$i_g = y_{gg}v_{gs} + y_{gd}v_{ds} + y_{gb}v_{bs} \quad (10.33)$$

$$i_d = y_{dg}v_{gs} + y_{dd}v_{ds} + y_{db}v_{bs} \quad (10.34)$$

A MOSFET designed for RF applications generally has the bulk internally shorted (short-circuited) to the source. Because $v_{bs} = 0$, the two-port y parameters are just a subset of the full y parameters:

$$[y] = \begin{bmatrix} y_{gg} & y_{gd} \\ y_{dg} & y_{dd} \end{bmatrix} = \begin{bmatrix} j\omega C_{gg} & -j\omega C_{gd} \\ g_m - j\omega C_{dg} & g_d + j\omega C_{dd} \end{bmatrix} \quad (10.35)$$

We write out the theoretical expressions of the transcapacitances appearing in the matrix. These expressions apply to transistors in strong inversion (either in the linear or saturation region), but are invalid at the subthreshold or accumulation regions:

$$C_{gg} = C_{ox} \left[\frac{2}{3} \frac{\alpha^2 + 4\alpha + 1}{(1 + \alpha)^2} + \frac{\delta}{3(1 + \delta)} \frac{(1 - \alpha)^2}{(1 + \alpha)^2} \right] \quad (10.36)$$

$$C_{gd} = C_{ox} \left[\frac{2}{3} \frac{\alpha^2 + 2\alpha}{(1 + \alpha)^2} \right] \quad (10.37)$$

$$C_{dg} = C_{ox} \left[\frac{1}{15} \frac{6\alpha^3 + 22\alpha^2 + 28\alpha + 4}{(1 + \alpha)^3} \right] \quad (10.38)$$

$$C_{dd} = C_{ox} \left[\frac{1 + \delta}{15} \frac{6\alpha^3 + 18\alpha^2 + 16\alpha}{(1 + \alpha)^3} \right] \quad (10.39)$$

where C_{ox} is the oxide capacitance. The parameter α , which has values between 0 and 1, is used to characterize the degree to which the transistor is in the linear region. It is defined as

$$\alpha = \begin{cases} 1 - \frac{V_{DS}}{V_{Dsat}} = 1 - \frac{(1 + \delta)V_{DS}}{V_{GS} - V_T} & \text{if } V_{DS} < V_{Dsat} \\ 0 & \text{if } V_{DS} \geq V_{Dsat} \end{cases} \quad (10.40)$$

where V_T is the threshold voltage. The parameter δ is the linearized slope of the bulk charge with respect to the bulk-to-source bias.¹⁰⁷ For convenience, δ is chosen to be a constant, having typical values ranging from 0.2 to 0.3. V_{Dsat} , the saturation voltage, is equal to $(V_{GS} - V_T)/(1 + \delta)$. [Previously, in Eq. 10.9, when velocity saturation was considered for the short-channel transistors with negligible bulk effects ($\delta = 0$), V_{Dsat} was given in a different form.] When the drain-to-source bias (V_{DS}) exceeds V_{Dsat} , the device operates in saturation; otherwise, the device is in the linear operation region.

The y parameters of Eq. 10.35 are those of the intrinsic transistor. An actual device consists of series parasitic resistances at the terminals, as shown in Figure 10.31. To establish the RF properties of the overall transistor, we incorporate the effects of these parasitics by first converting the intrinsic y parameters to z parameters. The conversion is performed with some well-established formula.¹⁰⁹ Next, the series resistances are appended to the $[z]$ matrix. The final $[z']$ matrix for the overall transistor is given as

$$[z'] = [z] + \begin{bmatrix} R_G + R_S & R_S \\ R_S & R_D + R_S \end{bmatrix} = \begin{bmatrix} R_G + R_S + \frac{y_{22}}{\Delta y} & R_S - \frac{y_{12}}{\Delta y} \\ R_S - \frac{y_{21}}{\Delta y} & R_D + R_S + \frac{y_{11}}{\Delta y} \end{bmatrix} \quad (10.41)$$

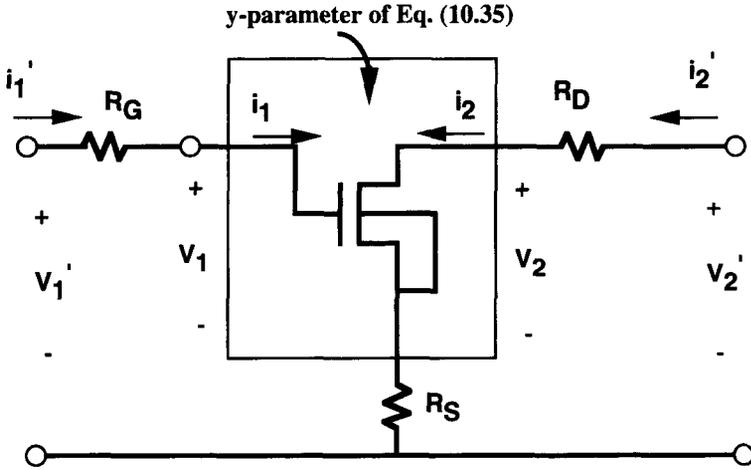


Figure 10.31 A schematic representation of a practical transistor consisting of the intrinsic portion and the terminal resistances.

where Δy is

$$\Delta y = y_{11}y_{22} - y_{12}y_{21} = j\omega(C_{gg}g_d + C_{gd}g_m) + \omega^2(C_{gd}C_{dg} - C_{gg}C_{dd}) \quad (10.42)$$

We are ready to derive two of the most important figures of merit in characterizing the RF characteristics of the transistor. One is the cutoff frequency (f_T), defined as the frequency at which the magnitude of the forward-current gain (h_{21}) is equal to one. Another is the maximum oscillation frequency (f_{max}), the frequency at which the unilateral power gain (U) is equal to 1. Because f_T characterizes the switching speed and f_{max} measures the amount of power gain, f_T is more critical to digital circuit design while f_{max} is the more important figure of merit for RF applications.

According to a conversion table,¹⁰⁹ h_{21} is equal to $-z_{21}/z_{22}$. Or, working with the $[z']$ matrix given in Eq. 10.41, h'_{21} of the primed two-port of Figure 10.31 is $-z'_{21}/z'_{22}$. After a lengthy derivation and neglecting the second- and higher-order terms of ω , we find the forward current gain of the overall transistor to be

$$h'_{21} = -\frac{g_m - j\omega(R_S C_{gg} g_d + R_S C_{gd} g_m + g_m)}{j\omega[C_{gg} + (R_S + R_D)C_{gd} g_d + (R_S + R_D)C_{gd} g_m]} \quad (10.43)$$

We omit the $j\omega$ term in the numerator of Eq. 10.43 to acquire a simple expression for the cutoff frequency. Substituting $\omega = 2\pi f_T$ into the equation and setting the

condition $|h'_{21}| = 1$, we find f_T as

$$\frac{1}{2\pi f_T} = \frac{C_{gg}}{g_m} + \frac{C_{gg}}{g_m} (R_S + R_D)g_d + (R_S + R_D)C_{gd} \quad (10.44)$$

In the ideal case wherein the drain and source resistances are zero, Eq. 10.44 leads to the well-known (but approximate) expression

$$\omega_T = 2\pi f_T = \frac{g_m}{C_{gg}} \quad (10.45)$$

Because g_m is proportional to the inverse of gate length while C_{gg} is proportional to the gate length, f_T increases in each generation of CMOS devices. This statement is not necessarily applicable to f_{\max} , whose value is more sensitive to the parasitics as well as to processing details.

A unilateral gain expression in terms of the device z parameters has been derived.¹¹⁰ Expressing the variables in terms of the more fundamental y parameters with a matrix conversion procedure,⁹⁹ we obtain the unilateral power gain as

$$U = \frac{\left| \frac{y_{21} - y_{12}}{\Delta y} \right|^2}{4 \left[\operatorname{Re} \left(R_S + R_G + \frac{y_{22}}{\Delta y} \right) \operatorname{Re} \left(R_S + R_D + \frac{y_{11}}{\Delta y} \right) - \operatorname{Re} \left(R_S - \frac{y_{12}}{\Delta y} \right) \operatorname{Re} \left(R_S - \frac{y_{21}}{\Delta y} \right) \right]} \quad (10.46)$$

After some algebraic manipulations, the unilateral gain can be expressed as

$$\begin{aligned} \frac{1}{U} = & \frac{4\omega^2}{g_m^2} \{ R_G [(R_S + R_D)(C_{gg}g_d + C_{gd}g_m)^2 + C_{gg}(C_{gg}g_d + C_{gd}g_m)] \\ & + R_D [R_S(C_{gg}g_d + C_{gd}g_m)^2 + C_{gd}(C_{dg}g_d + C_{dd}g_m)] \\ & + R_S [g_m(C_{gg} - C_{gd})(C_{gd} - C_{dd}) + g_d(C_{gg} - C_{gd})(C_{gg} - C_{dg})] \} \end{aligned} \quad (10.47)$$

Generally, the drain and source resistances are minimized by doping the drain and source heavily. However, because of the resistance associated with the polysilicon gate material (R_{sh}), there is always a finite gate resistance, especially when the gate length is below $0.25 \mu\text{m}$. R_{sh} has a typical value of $3 \Omega/\text{square}$ in a $0.35\text{-}\mu\text{m}$ technology, but can increase to $10 \Omega/\text{square}$ as the gate length shrinks below $0.18 \mu\text{m}$. A thorough derivation of R_G based on the solution of a partial differential equation has been carried out.¹¹¹ The analysis identifies a $\frac{1}{3}$ factor that accounts for the distributed nature of the gate current flow:

$$R_G = \frac{1}{3} R_{sh} \frac{W}{L_{\text{gate}}} \cdot \frac{1}{N} \quad (10.48)$$

where W and L_{gate} are the gate width and length in a gate finger, respectively, and N is the total number of the fingers in the transistor. The factor $\frac{1}{3}$ can also be derived from a two-dimensional resistive network analog; however, such a derivation would not reveal the time evolution of the gate voltage as both a function of time and position in the finger.

We neglect the latter two terms in Eq. 10.47 to establish a simple equation for the maximum oscillation frequency. Replacing U by 1 and ω by $2\pi \cdot f_{max}$, we find

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_G C_{gd} \left(1 + \frac{2\pi f_T}{C_{gd}} \Psi\right)}} \quad (10.49)$$

where Ψ is

$$\Psi = (R_D + R_S) \frac{C_{gg}^2 g_d^2}{g_m^2} + (R_D + R_S) \frac{C_{gg} C_{gd} g_d}{g_m} + \frac{C_{gg}^2 g_d}{g_m^2} \quad (10.50)$$

Equations 10.49 and 10.50 indicate that, if the parasitics resistances were zero, f_{max} would attain an infinite value. This is in contrast to f_T whose value remains finite even in an ideal transistor without parasitics.

This analysis is sufficient for the f_T and f_{max} calculations when the transistor operates in strong inversion. It is desirable to establish a physical, scalable model capable of simulating both dc I - V characteristics and the s -parameter characteristics, and applicable to all bias ranges. A de facto industry standard SPICE model (BSIM3v3) is a potential candidate,¹¹² having demonstrated accuracy and scalability in the devices' dc characteristics. When it is applied to simulate the s parameters at RF, we find BSIM3 to be somewhat problematic and two modifications are necessary to obtain a good fit. The first modification involves the addition of a bulk resistive network, achievable with a simple circuit extension to the existing BSIM3 model. The second improvement accounts for the finite channel resistance, whose origin lies in a NQS analysis. Since BSIM3 (as well as the many other models) employs a quasistatic assumption, this improvement, while important, involves some changes in the model source code.

We first concentrate on the fitting of s_{22} , from which we demonstrate the need of a bulk resistive network. s_{22} characterizes the output impedance when the input port (gate) is terminated with $R_0 = 50 \Omega$. Figure 10.32 illustrates measured s_{22} when the transistor is in OFF state. The data lie on a semicircle characterized by $r = 0.6$ (r is shown in the figure), indicating that the real part of the output impedance (R_{out}) is $0.6 \cdot R_0 = 30 \Omega$ and is independent of frequency. Figure 10.33a presents an intuitive but wrong approach to determine the value of R_{out} . Because the transistor was OFF, the impedance path to ground seen at the output port consisted of a serial connection of the overlap gate-drain capacitance and the terminating resistance of 50Ω . R_{out} , in this example, would be identical to 50Ω . However, the circuit of Figure 10.33a neglects the presence of a larger parasitic capacitance, the bulk-drain junction capacitance C_{jdb} . Because the default BSIM3 configuration tied the bulk to the

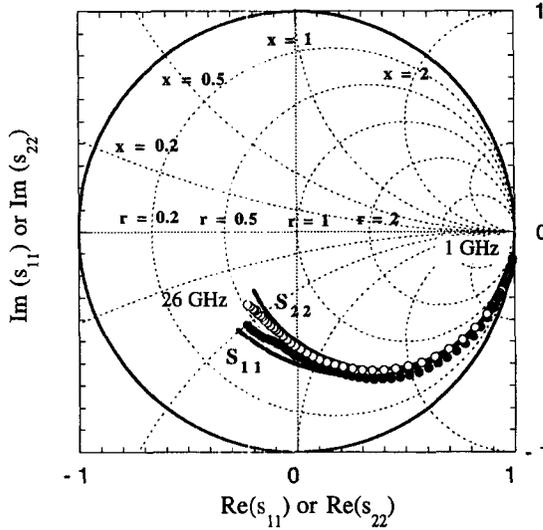


Figure 10.32 Measured s_{11} and s_{22} of a RF MOS transistor when the transistor is in OFF state ($L_{\text{eff}} = 0.29 \mu\text{m}$; $W_G = 256 \mu\text{m}$; $V_{GS} = V_{DS} = 0$). Solid dots are measured s_{11} ; open circles are measured s_{22} ; solid lines are simulation results.

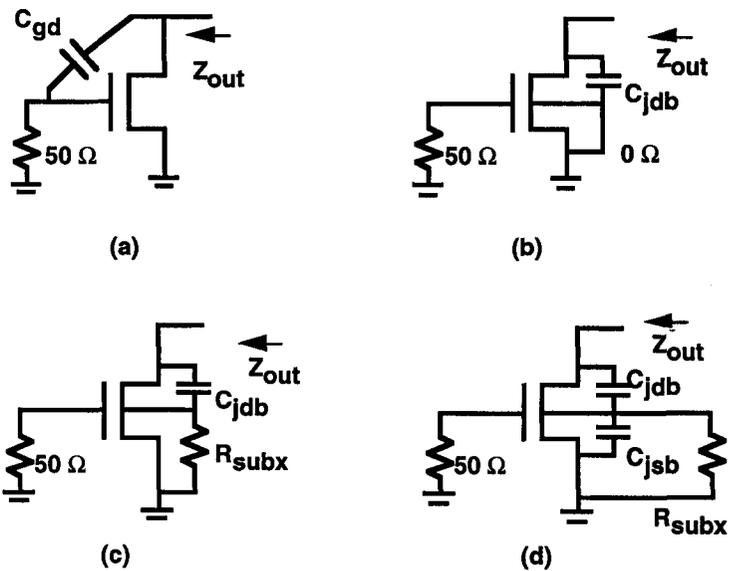


Figure 10.33 Schematic representations of the transistor for the discussion of s_{22} : (a) only the overlap gate-drain capacitance is considered; (b) the drain–bulk junction capacitance is considered; (c) a substrate resistance is included; (d) the full representation in which both the drain–bulk and source–bulk junction capacitances are tied to the substrate resistance. (After Liu et al., Ref. 113, © IEEE, reprinted with permission.)

source, the path connecting the output to the ground would give rise to a R_{out} of $0\ \Omega$ as shown in Figure 10.33*b*.

An approach to amend this problem could be to connect a resistor to the bulk node. As shown in Figure 10.33*c*, the output impedance would then consist of C_{jdb} in series with R_{subx} . R_{out} would therefore be equal to R_{subx} .

Unfortunately, this reasoning is still fallacious. Figure 10.33*d* gives a more accurate representation of the transistor. Besides C_{jdb} , the source–bulk junction capacitance (C_{jsb}) also exists at the bulk node. The output impedance consists of C_{jdb} in series with a parallel combination of R_{subx} and C_{jsb} . At high frequencies, C_{jsb} bypasses the signal that would otherwise flow through R_{subx} , reducing R_{out} from its low-frequency value of R_{subx} toward 0 at high frequencies. The measurement, in contrast, reveals that R_{out} remains relatively constant with frequency. Merely adding a substrate resistance to the bulk node of the transistor clearly does not lead to a good fit of s_{22} across all frequencies.

The thought process reflected above reveals the importance of inserting a resistance between C_{jdb} and C_{jsb} . Without such a resistance, C_{jsb} always shunts the signal flow from the output to ground at high frequencies. In the BSIM3 model (as well as many other models), C_{jdb} and C_{jsb} are hard-wired to the bulk node, not allowing for an addition of resistance between the two ends of the capacitors. We get around this problem by declaring the transistor's junction areas and junction peripheries to be zero and thereby eliminate C_{jdb} and C_{jsb} in the transistor. In return, semiconductor capacitances tantamount to C_{jdb} and C_{jsb} are declared externally. This way, resistances ($R_{sub2} + R_{sub3}$) can be inserted between the junction capacitances, as shown in Figure 10.34*a*. The basis of this model is made clear by the physical representation of the transistor shown in Figure 10.34*b*. In accordance with the symmetry criteria, $R_{sub2} = R_{sub3}$ and $R_{sub1} = R_{sub4}$.

The substrate resistance network is required for an accurate fit in s_{22} for all biases. Additionally, the resistive network is important for the fitting in s_{11} when the transistor is not yet turned ON. Analogous to s_{22} , s_{11} characterizes the input impedance when the output port (drain) is terminated with $R_0 = 50\ \Omega$. When the channel inversion has not taken place, the lack of shielding by the channel charges leads to a relatively high value for the gate-to-bulk capacitance (C_{gb}). Figure 10.35 is basically identical to Figure 10.34*a*, except that we replace the MOS transistor by an equivalent circuit. The substrate resistive network has not been altered. As shown, C_{gb} provides a path by which the input signal traverses through R_{sub2} . The signal eventually couples through C_{jdb} and exits to the output port where it encounters the termination resistance of $50\ \Omega$. For this discussion, we first neglect the entire resistive network and simplify the equivalent circuit of Figure 10.35 to that of Figure 10.36. In this scenario, C_{gb} , which connects to the grounding bulk terminal, is parallel to the gate-to-source capacitance (C_{gs}). The input impedance consists of two main paths in parallel. One is that formed with the sum of C_{gs} and C_{gb} in series with R_S . The remaining path consists of C_{gd} in series with the terminating $50\ \Omega$ (g_m , g_d of the transistor can be omitted since the transistor is in OFF state.) In an ideal transistor where the gate–drain capacitance (C_{gd}) approaches 0 (no parasitic overlap capacitance), the input resistance reduces to $R_G + R_S$, which was established to be

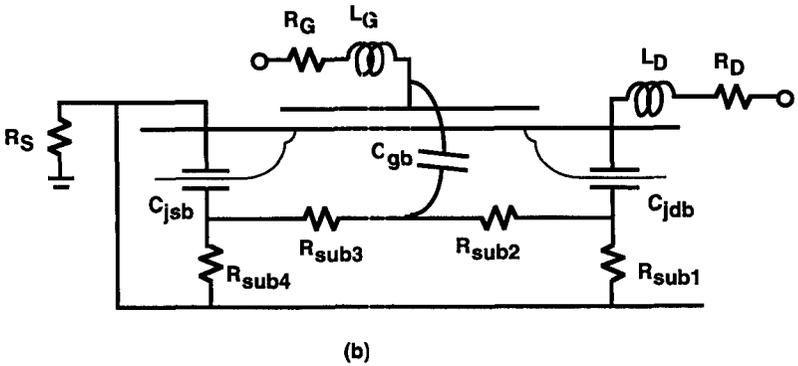
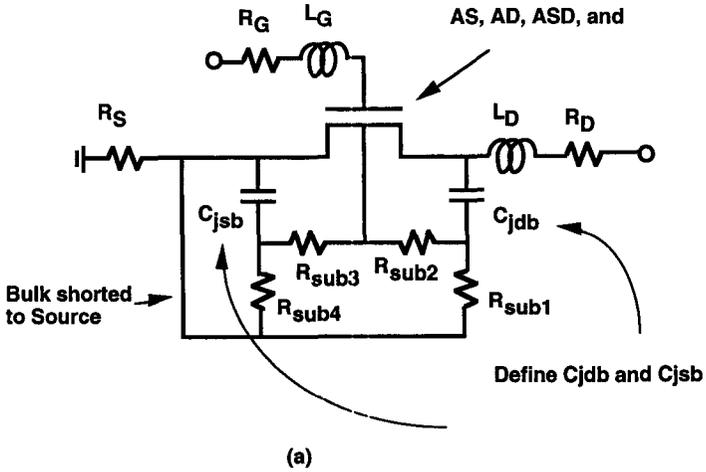


Figure 10.34 (a) An equivalent-circuit representation of a MOS transistor. The drain–bulk and the drain–source junction capacitances are detached from the bulk node to facilitate the insertion of resistances; (b) the physical representation of the MOS transistor, showing the origin of the various components shown in (a). (After Liu et al., Ref. 113, © IEEE, reprinted with permission.)

7.2 Ω. In contrast, the measured R_{out} according to Figure 10.32 is $0.6 \cdot R_0 = 30 \Omega$. This large discrepancy indicates that the path formed with the overlap gate–drain capacitance should be considered. It is a straightforward algebraic exercise to determine the real part of the input impedance when both the $C_{gb} + C_{gs}$ and the C_{gd} paths are considered:

$$\begin{aligned}
 \text{Re}(Z_{in})|_{\text{at low freq}} &= R_G + \frac{C_{gs}^2 R_S + C_{gd}^2 R_0}{(C_{gs} + C_{gd})^2} \\
 &= R_G + \frac{R_S + R_0}{4} \quad \text{when } C_{gs} = C_{gd} \quad (10.51)
 \end{aligned}$$

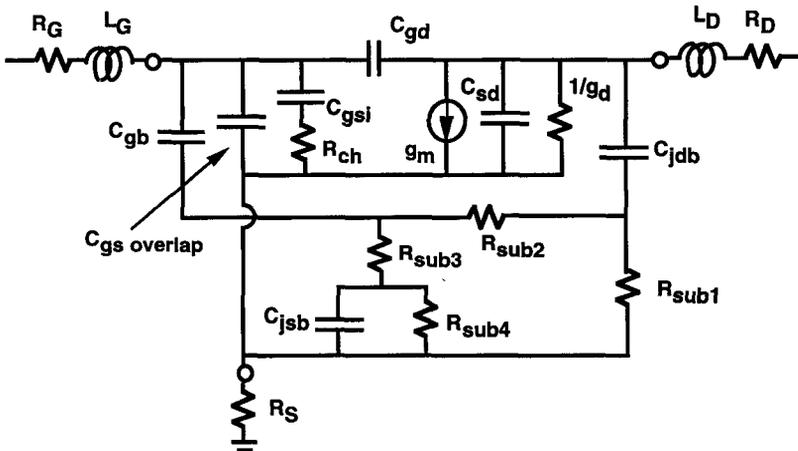


Figure 10.35 A more elaborate representation of the MOS transistor shown in Figure 10.34a. The transistor itself is replaced by an equivalent circuit. The terminal resistances and inductances, whose values are extracted from a shorted test structure, are added for completeness ($R_G = 6 \Omega$, $R_S = 1.2 \Omega$, $R_D = 2.5 \Omega$, $L_G = 0.08$ nH, $L_D = 0.1$ nH). While the serial impedances are mostly associated with the feedthroughs of the RF probing structure, R_G contains an 1- Ω component that is due to the poly-gate sheet resistance. (After Liu et al., Ref. 113, © IEEE, reprinted with permission.)

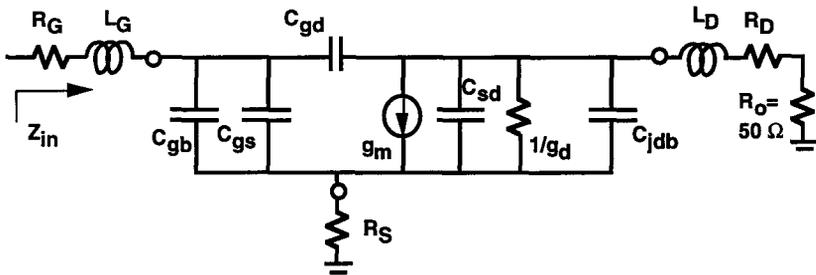


Figure 10.36 A simplified circuit representation of Figure 10.35 when the substrate resistive network is omitted. This representation, although intuitive, does not yield the correct s_{22} . (After Liu et al., Ref. 113, © IEEE, reprinted with permission.)

The low-frequency R_{out} is $R_G + (R_0 + R_S)/4 = 19 \Omega$. Although it is certainly higher than the 7.2Ω obtained when the C_{gd} path is disconnected, the $19\text{-}\Omega$ value still differs from the measured $30\text{-}\Omega$ value. To account for this difference, we now reconsider the resistive network developed previously.

Figure 10.37 is a partial extraction of the complete equivalent circuit of Figure 10.35. It includes the path between the coupling capacitances of C_{gb} at the input and C_{jdb} at the output. This path runs through a substrate resistance R_{sub2} and terminates

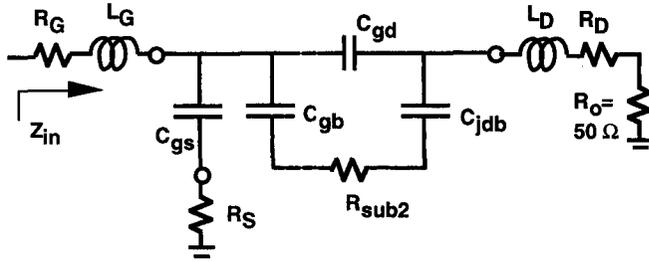


Figure 10.37 A subcircuit of the equivalent circuit of Figure 10.35. The essence of the substrate conduction path is captured with the inclusion of the C_{gb} - R_{sub2} - C_{jdb} path. (After Liu et al., Ref. 113, © IEEE, reprinted with permission.)

at $R_o = 50 \Omega$. Unlike the representation of Figure 10.36, it is difficult to write out a mathematical expression for the input resistance here, mainly due to the presence of a loop at the output node. Nonetheless, the input resistance is easily evaluated from a SPICE simulation and is found to have a high value of 30Ω . The actual input resistance of the entire circuit of Figure 10.35 is smaller because of the shunting paths to ground provided by R_{sub1} and $R_{sub4} || C_{jsb}$, resulting in the close match with measurements (Figure 10.32).

At higher biases near or below V_T , the substrate resistance network, though still important to the accurate modeling of s_{22} , is less critical to s_{11} since C_{gb} decreases with V_{GS} . However, there is a widening error in fitting s_{11} . This is because the BSIM3 model, as well as many other models, is based on quasistatic analysis. It does not account for the channel resistance, whose position in an equivalent-circuit model relative to other elements was shown in Figure 10.35. It is in series with the intrinsic gate-source capacitance (C_{gsi}), which in turn, is in parallel with the overlap capacitance. According to a NQS analysis, the y parameter associated with the intrinsic gate-source terminal can be written as

$$y_{gsi} = -j\omega C_{gsi} \left(\frac{1 + j\omega\tau_2}{1 + j\omega\tau_1} \right) \tag{10.52}$$

If the frequency terms in the parentheses of Eq. 10.52 are omitted, then y_{gsi} simplifies to $-j\omega C_{gsi}$, which is exactly the QS expression used in BSIM3 and shown in Eq. 10.31.

In Eq. 10.52, the parameters τ_1 and τ_2 are given by¹⁰⁷

$$\tau_1 = \frac{4}{15} \frac{1}{\omega_0} \frac{\alpha^2 + 3\alpha + 1}{(1 + \alpha)^3} \tag{10.53}$$

$$\tau_2 = \frac{1}{15} \frac{1}{\omega_0} \frac{5\alpha^2 + 8\alpha + 2}{(1 + \alpha)^2(1 + 2\alpha)} \tag{10.54}$$

where the intrinsic frequency of the transistor is given by

$$\omega_0 = \frac{\mu_n(V_{GS} - V_T)}{(1 + \delta)L_{\text{eff}}^2} \quad (10.55)$$

where μ_n is the electron mobility in the channel and L_{eff} is the effect channel length, which differs from the mask dimension L_{gate} . We have mentioned that ω_0 is used as a reference frequency. When the operating frequency is lower than ω_0 , $\tau_1 \approx \tau_2 \approx 0$. Therefore, the quasistatic y parameters of Eq. 10.31 are considered accurate. This statement, while generally true (especially for digital circuits), is not entirely correct when the resistance associated with the channel is important. More specifically, y_{gs} of Eq. 10.31 is entirely imaginary, suggesting that the input impedance does not contain a real part. If instead the NQS expression of Eq. 10.52 is used, a finite resistance is found to exist between the gate and the source terminal, even as frequency tends to zero:

$$R_{\text{ch}} = \lim_{\omega \rightarrow 0} \text{Re} \left(-\frac{1}{y_{gsi}} \right) = \lim_{\omega \rightarrow 0} \frac{-\omega(\tau_2 - \tau_1)}{\omega C_{gs}(1 + \omega^2 \tau_2^2)} = \frac{(\tau_1 - \tau_2)}{C_{gs}}. \quad (10.56)$$

Similar NQS expressions (e.g., Eq. 10.52) have been derived for other y parameters. If the motivation for adopting the NQS equations is to account for the channel resistance, it is not necessary to convert all of the BSIM3 y parameters to their respective NQS counterparts. Of the 16 y parameters, only $y_{gg}, y_{gd}, y_{gs}, y_{dg}, y_{dd}, y_{sg}$, and y_{ss} need to be modified without violating the requirement that the sums of the y parameters in a row and in a column be zero.

We have never stated it explicitly, but the NQS and QS expressions (Eqs. 10.31 and 10.52) are derived for long-channel devices. The continuity equation accounting for the velocity saturation in short-channel devices is more complicated. A practical compromise to obtain the short-channel parameters is to modify the solution based on the long-channel devices. In this regard, we introduce a τ -factor given by (see Problem 10.7):

$$\tau_1 = \frac{C_{dg} - C_{gd}}{g_m} \cdot \tau\text{-factor} \quad (10.57)$$

When the τ -factor is 0, τ_1 goes to zero and the whole expression reduces to the quasi-static expressions. When the τ -factor is 1, the channel resistance is properly accounted for in long channel devices. For short-channel devices, the τ -factor exceeds 1. A τ -factor of 3 is found to yield a good match between the measured and simulated s -parameters at various bias conditions.¹¹³

10.5 LARGE-SIGNAL POWER AND EFFICIENCY

MOSFETs are used for large-signal operations in which the device characteristics cannot be well linearized. Small-signal parameters such as the y and s parameters are

not meaningful under such operations. Unlike the small-signal power amplifier design, the design of large-signal power amplifiers is complicated because the transistor cannot be represented by a linear two-port network. An accurate design requires the use of computer-aided design (CAD) tools that incorporate large-signal models to represent the transistor. However, even without such tools, considerable understanding of the large-signal operation of the amplifier remain possible.

A schematic bias configuration for a large-signal amplifier is shown in Figure 10.38. (In an actual circuit, there will be a matching circuit between the transistor and a 50-Ω load. The load R_L in Figure 10.38a or $R_L || L || C$ shown in Figure 10.38b is assumed to be the impedance seen at the output of the transistor, at all frequencies including the harmonics.) The input of the power amplifier consists of a sinusoidal voltage $v_{gs} \cdot \cos \omega t$ and a dc bias voltage V_{GS} . The sum of these two components is denoted as v_{GS} . We assume that the drain voltage is biased at a high value such that the transistor stays in the saturation whenever it is turned on. We will further assume that the transfer characteristics do not depend on frequency. In essence, the transfer characteristics are those obtained from a dc measurement, disregarding the presence of the capacitances and time delays intrinsic to an actual transient. Therefore, the drain current in response to a gate-source voltage v_{GS} obeys the following:

$$i_D(t) = \begin{cases} \frac{W}{L_{\text{eff}}} \mu_n C'_{\text{ox}} \frac{(v_{GS}(t) - V_T)^2}{2(1 + \delta)} & \text{when } v_{GS} \geq V_T \\ 0 & \text{when } v_{GS} < V_T \end{cases} \quad (10.58)$$

$$v_{GS}(t) = V_{GS} + v_{gs} \cos(\omega t). \quad (10.59)$$

where C'_{ox} is the oxide capacitance per unit-area.

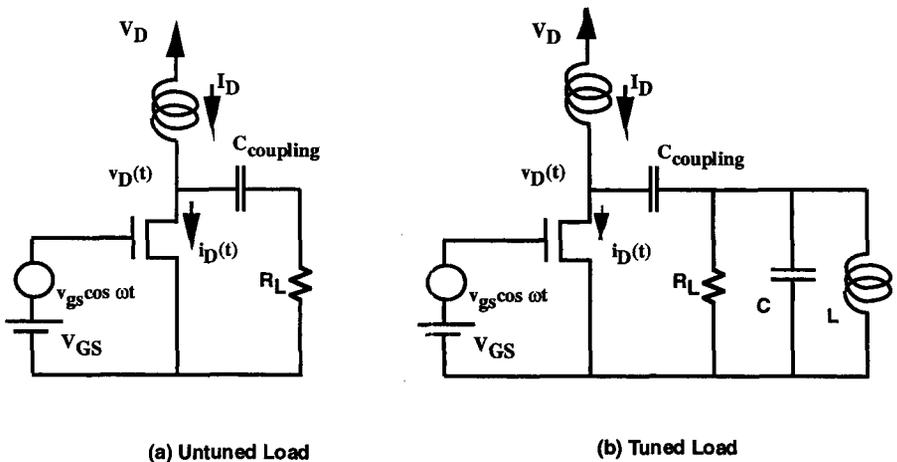


Figure 10.38 Schematic bias configurations for a large-signal amplifier; (a) untuned amplifier in which the load is purely resistive; (b) tuned amplifier with the load designed to allow the passage of only the fundamental current component.

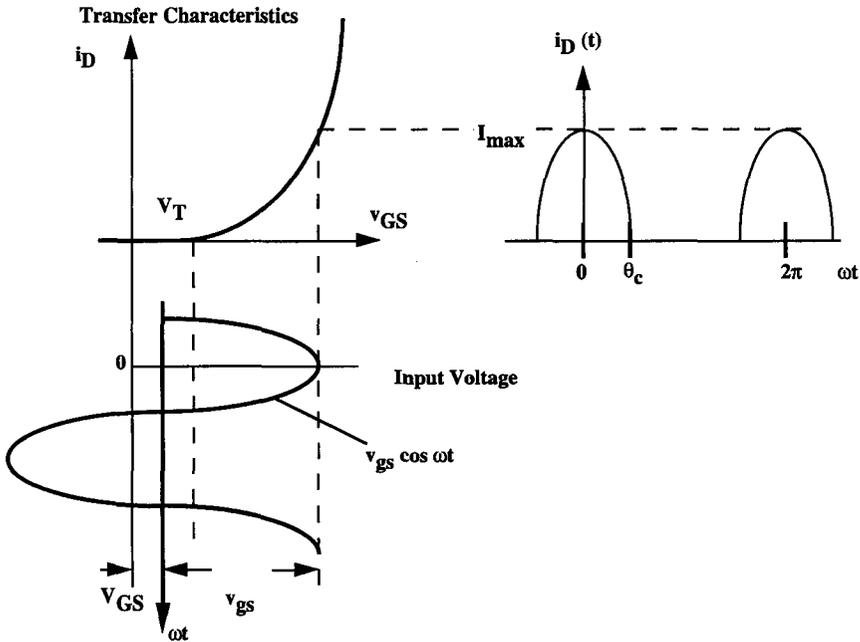


Figure 10.39 Transfer characteristics of a large-signal amplifier whose $I-V$ characteristics are described by Eq. 10.61. The figure shows a class AB operation in which the conduction angle is between 0 and 180° .

Depending on the relative magnitudes of v_{gs} and V_{GS} , the output current can exceed zero all the time or during only a portion of the period. Figure 10.39 shows the relationship between the input voltage and the output current waveforms, a relationship that is governed by Eq. 10.58. The figure is drawn with $V_{GS} > 0$, although the equations would work at $V_{GS} < 0$. As the sinusoidal voltage increases above and decreases below zero, the magnitude of total input voltage oscillates with time. As long as the total input voltage exceeds V_T , some current flows through the drain terminal. The drain current decreases to zero when $V_{GS} + v_{gs} \cdot \cos \omega t = V_T$. Therefore, as shown in the figure, the output current waveform resembles pulse trains. The conduction angle (θ_c), whose definition is shown in Figure 10.39, is a measure the amount of time that the transistor is ON. It is obtained by equating $i_D = 0$ when ωt is set to θ_c :

$$\theta_c = \cos^{-1} \left(\frac{V_T - V_{GS}}{v_{gs}} \right) \quad (10.60)$$

If $V_{GS} - V_T > v_{gs}$, the output current is greater than zero at all times, instead of being trains of current pulses. This situation resembles a small-signal operation in which the ac signal perturbs the current only slightly about the quiescent value. The transistor operation is categorized to different classes according to the

conduction angle. It is a class A operation if $\theta_c = 180^\circ$, class B if $\theta_c = 90^\circ$. When $90^\circ < \theta_c < 180^\circ$, it is a class AB operation. When $\theta_c < 90^\circ$, it is a class C operation. The bias condition depicted in Figure 10.39 results in a class AB operation.

Suppose that during a cycle the maximum drain current is I_{\max} , then I_{\max} takes place when $\omega t = 0$. Using I_{\max} and θ_c as the parameters, we express the drain current as

$$i_D(t) = \begin{cases} I_{\max} \frac{(\cos \omega t - \cos \theta_c)^2}{(1 - \cos \theta_c)^2} & \text{when } v_{GS} \geq V_T \\ 0 & \text{when } v_{GS} < V_T \end{cases} \quad (10.61)$$

To facilitate the following analysis, we write $i_D(t)$ in terms of its Fourier series, expressing it as a sum of a dc current and the harmonic components, $I_d(n)$:

$$i_D(t) = I_{dc} + \sum_{n=1}^{\infty} I_d(n) = I_{dc} + I_{\max} \sum_{n=1}^{\infty} a_n \cos(n\omega t) \quad (10.62)$$

The dc component I_{dc} as well as the Fourier coefficients a_1, a_2 , and a_n values (for $n \geq 3$) are given by

$$I_{dc} = \frac{I_{\max}}{2\pi(1 - \cos \theta_c)^2} [\theta_c + 2\theta_c \cos^2 \theta_c - 3\cos \theta_c \sin \theta_c]; \quad (10.63)$$

$$a_1 = \frac{2[\cos^2 \theta_c \sin \theta_c + 2 \sin \theta_c - 3\theta_c \cos \theta_c]}{3\pi(1 - \cos \theta_c)^2}; \quad (10.64)$$

$$a_2 = \frac{2 \cos^3 \theta_c \sin \theta_c - 5 \cos \theta_c \sin \theta_c + 3\theta_c}{6\pi(1 - \cos \theta_c)^2}; \quad (10.65)$$

$a_n =$

$$\frac{4[2\sin n \theta_c \cos^2 \theta_c + \sin n \theta_c + n^2 \sin n \theta_c \cos^2 \theta_c - n^2 \sin n \theta_c - 3n \cos n \theta_c \cos \theta_c \sin \theta_c]}{\pi(1 - \cos \theta_c)^2(n-2)n(n+2)(n-1)(n+1)} \quad (10.66)$$

We examine the output voltage at the drain terminal. In the absence of any sinusoidal excitation at the input, the dc current I_{dc} flows through the ideal inductor without a potential drop. Thus, $v_D(t)$ is the dc bias voltage V_D . When $v_{gs} \cdot \cos \omega t$ is superimposed on top of V_{GS} , the harmonic currents $I_d(n)$ also flow through the transistor. Because the inductance between the power supply and the drain provides a large impedance at high frequencies, these harmonic components necessarily come from the load through the coupling capacitance. The value of the coupling capacitance between the load and the power supply has to be large so that I_{dc} is blocked from the load while allowing the harmonics to pass to the transistor without a potential drop.

We shall state without proof that, to achieve the largest possible output power, the load impedance should be purely real, denoted by R_L as shown in Figure 10.38a. When an imaginary component exists in the load, the loadline of the amplifier (the trace of the operating i_D and V_D) can traverse through high-voltage and high-current regions simultaneously. This kind of operation is prone to device breakdown. The detailed analysis comparing amplifiers with purely real or partially imaginary loads can be found elsewhere.⁹⁹

The instantaneous drain voltage $v_D(t)$ is a sum of the dc bias voltage and the ac voltage drop brought about by the harmonic currents. Since the load is purely real, we obtain

$$v_D(t) = V_D - I_{\max} \sum_{n=1}^{\infty} a_n \cos(n\omega t) R_L \quad (10.67)$$

With both the instantaneous current and voltage waveforms known, the average power dissipated by the device (P_{diss}) is evaluated. Using $i_D(t)$ of Eq. 10.62 and $v_D(t)$ of Eq. 10.67, and noting that the products harmonics of different frequencies are averaged to zero in a given period (T), we find P_{diss} to be

$$P_{\text{diss}} \equiv \frac{1}{T} \int_0^T i_D(t) \cdot v_D(t) dt = I_{\text{dc}} V_D - \sum_{n=1}^{\infty} \frac{I_{\max}^2 a_n^2}{2} R_L. \quad (10.68)$$

If P_{diss} is positive, the transistor dissipates power. If instead P_{diss} is negative, the transistor delivers power to the rest of the circuit. The latter condition is precisely what the amplifier is designed for. More specifically, we desire the output power at the fundamental frequency (ω , which is the frequency of the input sinusoidal). We define P_{out} as the output power at the fundamental frequency delivered to the load. It is given by

$$P_{\text{out}} = \frac{I_{\max}^2 a_1^2}{2} R_L = \frac{2I_{\max}^2 [\cos^2 \theta_c \sin \theta_c + 2 \sin \theta_c - 3\theta_c \cos \theta_c]^2}{9\pi^2 (1 - \cos \theta_c)^4} R_L \quad (10.69)$$

For a given θ_c , it is clear that both I_{\max} and R_L should be maximized to maximize P_{out} given by Eq. 10.69. However, neither I_{\max} nor R_L can be increased indefinitely without adverse effects. We consider first the limit placed on the value of R_L , a limit that relates to the knee voltage of the MOSFET. As R_L increases, Eq. 10.67 shows that the minimum operating voltage decreases while the maximum operating voltage increases. The increase in the overall range of voltage swing is the fundamental reason why P_{out} generally increases with R_L . However, when R_L increases to an extreme, the minimum operating voltage becomes negative. Realistically, when $v_D \leq V_{\text{knee}}$ (typically ~ 0.5 V), the transistor enters the linear region wherein i_D decreases from its maximum value. The output power expression given by Eq. 10.69 would then cease to be accurate and P_{out} is expected to decrease. In essence, V_{knee} is

the minimum required operating voltage at the collector when $i_D(t) = I_{\max}$. From Eq. 10.62, $i_D(t) = I_{\max}$ when $\omega t = 0$. Substituting this condition into Eq. 10.67, we find V_{knee} to be

$$V_{\text{knee}} = v_D(t)|_{\omega t=0} = V_D - I_{\max} R_L \sum_{n=1}^{\infty} a_n \quad (10.70)$$

The optimal load resistance to maximize P_{out} is obtained by rearranging Eq. 10.70:

$$R_{L\text{opt}} = \frac{(V_D - V_{\text{knee}})}{I_{\max}} \left(\sum_{n=1}^{\infty} a_n \right)^{-1} \quad (10.71)$$

How about the limit placed on the drain current? Equation (10.58) shows that the drain current continues to increase as the gate-to-source bias increases. Realistically, the transistor can be operated safely only below a certain maximum current I_{\max} . We assume that, if the operating drain current exceeds I_{\max} , then g_m suddenly plummets and P_{out} decreases.

The optimal output power that is obtained with the optimal load resistance is given by Eq. 10.67. Using the result of Eq. 10.71, we have

$$P_{\text{out-opt}} = I_{\max} (V_D - V_{\text{knee}}) \frac{a_1^2}{2} \left(\sum_{n=1}^{\infty} a_n \right)^{-1} \quad (10.72)$$

We denote P_{out} as the general RF output power. $P_{\text{out-opt}}$, on the other hand, is the output power obtained when $R_L = R_{L\text{opt}}$ and the current level is I_{\max} .

The drain efficiency (ε_{ff}) is defined as the average RF power of the fundamental frequency measured at the output over the dc power dissipation of the entire circuit (P_{dc}):

$$\varepsilon_{\text{ff}} = \frac{P_{\text{out}}}{P_{\text{dc}}} = \frac{P_{\text{out}}}{I_{\text{dc}} \cdot V_D} \quad (10.73)$$

The optimal drain efficiency ($\varepsilon_{\text{ff-opt}}$) is the efficiency when P_{out} in this expression is replaced by $P_{\text{out-opt}}$. Using Eq. 10.72, we obtain

$$\varepsilon_{\text{ff|opt}} = \frac{I_{\max} (V_D - V_{\text{knee}})}{V_D} \frac{a_1^2}{2I_{\text{dc}}} \left(\sum_{n=1}^{\infty} a_n \right)^{-1} \quad (10.74)$$

Again, we denote ε_{ff} as the general drain efficiency, while $\varepsilon_{\text{ff-opt}}$ means the output power obtained when $R_L = R_{L\text{opt}}$.

According to Eq. 10.74, V_{knee} should be minimized to increase the drain efficiency. A large V_{knee} reduces the available operating region of the transistor,

limiting the magnitude of the output voltage oscillation. Consequently, the optimum load resistance is decreased from that achievable when V_{knee} was zero. As $R_{L\text{opt}}$ decreases in value, $P_{\text{out-opt}}$ and $\varepsilon_{\text{ff-opt}}$ are also reduced. The effects of V_{knee} are especially detrimental when the transistor's dc bias voltage is small. Minimizing V_{knee} is therefore an important part of the device design.

Previously we determined $R_{L\text{opt}}$ from the condition that the minimum value of $v_D(t)$ is V_{knee} , as shown in Eq. 10.70. Besides its minimum value, we are interested in knowing the maximum value of $v_D(t)$ during one RF cycle. It is clear from the loadline examples that $v_D(t)$ reaches its maximum value when $\omega t = \pi$. When ωt is equal to π , Eq. 10.67 shows that the maximum value of $v_D(t)$, v_{DM} , is equal to

$$v_{DM} = v_D(t)|_{\omega t = \pi} = V_D - R_L \sum_{n=1}^{\infty} a_n (-1)^n \quad (10.75)$$

We need to make sure that v_{DM} is smaller than the drain-source breakdown voltage (BV_{DS}) so that the normal transistor operation is maintained during the entire RF cycle. If the calculated v_{DM} exceeds BV_{DS} , V_D must be reduced so that the transistor does not break down during an RF cycle. Let us consider again the optimal condition under which $R_L = R_{L\text{opt}}$. The maximum collector voltage in the optimal condition, $v_{DM\text{-opt}}$ is obtained as:

$$v_{DM\text{-opt}} = v_D(\omega t = \pi)|_{\text{opt}} = V_D - R_{L\text{opt}} \sum_{n=1}^{\infty} a_n (-1)^n = V_D + (V_D - V_{\text{knee}}) \mathbf{f}_{\text{VDM}} \quad (10.76)$$

where the factor \mathbf{f}_{VDM} is given by,

$$\mathbf{f}_{\text{VDM}} = - \left(\sum_{n=1}^{\infty} a_n \right)^{-1} \sum_{n=1}^{\infty} a_n (-1)^n \quad (10.77)$$

We calculate $P_{\text{out-opt}}$, $\varepsilon_{\text{ff-opt}}$, $R_{L\text{opt}}$, \mathbf{f}_{VDM} as a function of the conduction angle from Eqs. 10.72, 10.74, 10.71, and 10.77, respectively. Figure 10.40 shows $P_{\text{out-opt}}$ normalized by $I_{\text{max}} \cdot (V_D - V_{\text{knee}})$. Figure 10.41 shows $\varepsilon_{\text{ff-opt}}$ normalized by $(V_D - V_{\text{knee}})/V_D$. Figure 10.42 shows $R_{L\text{opt}}$ normalized by $(V_D - V_{\text{knee}})/I_{\text{max}}$. Figure 10.43 shows \mathbf{f}_{VDM} . As is clear from Figures 10.41 and 10.42, a desire to simultaneously maximize P_{out} and ε_{ff} naturally leads to biasing the transistors in the class AB operation. If a high efficiency is desired, class C operation can be considered, at the cost of reduced P_{out} .

As an example, we design a MOS amplifier with a V_{knee} of 0.5 V for the class A operation. Its maximum operable current is $I_{\text{max}} = 50$ mA. According to Eqs. 10.63–10.66, when $\theta_c = \pi$, $I_{\text{dc}} = \frac{3}{8} \times I_{\text{max}}$, $a_1 = \frac{1}{2}$, $a_2 = \frac{1}{8}$, and a_n 's are zero for $n \geq 3$. $R_{L\text{opt}}$ from Eq. 10.71 is therefore $1.6 \times (V_D - V_{\text{knee}})/I_{\text{max}}$. If we bias the circuit at $V_D = 5$ V, and $I_{\text{max}} = 50$ mA, then $R_{L\text{opt}} = 144 \Omega$. This is much smaller

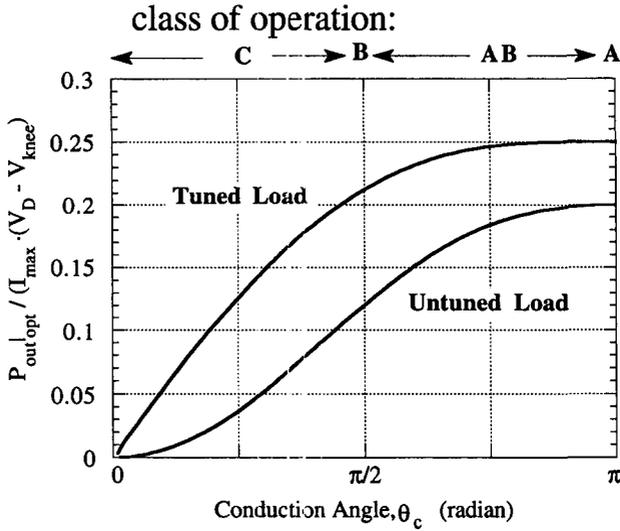


Figure 10.40 Large-signal output power as a function of the conduction angle when the optimal load is used. The MOS is assumed to be a linear g_m device.

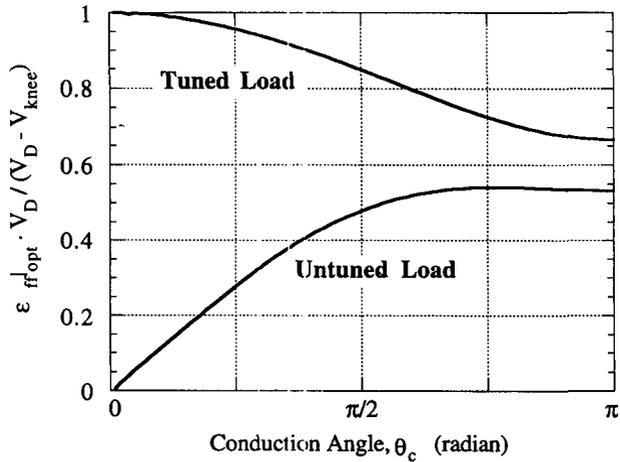


Figure 10.41 Drain efficiency as a function of the conduction angle when the optimal load is used. The MOS is assumed to be a linear g_m device.

than the optimum load resistance used in small-signal amplifiers. Therefore, if a circuit's load resistance is designed for small-signal operation, it will deliver a power that is less than the maximum possible value under a large-signal operation.

Note that when $\omega t = 0$, $i_D(t)|_{opt}$ reaches its maximum value of I_{max} , and at $\omega t = \pi$, $v_D(t)|_{opt}$ reaches its maximum value. At the time when $v_D(t)$ attains its dc value of $V_D = 5$ V, $i_D(t)|_{opt}$ attains its dc value of $0.375 \times I_{max} = 18.75$ mA. $P_{out-opt}$

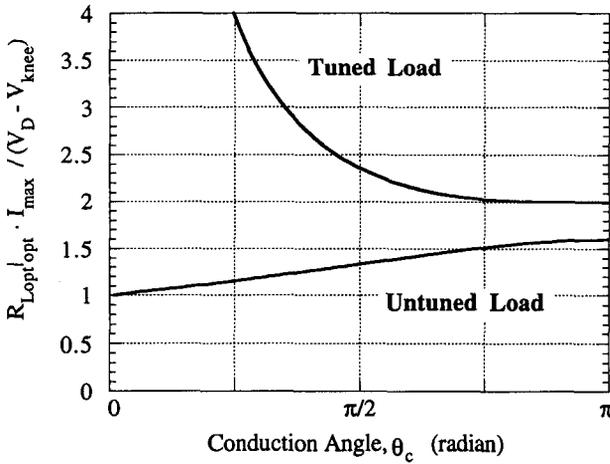


Figure 10.42 Optimal load resistance as a function of the conduction angle. The MOS is assumed to be a linear g_m device.

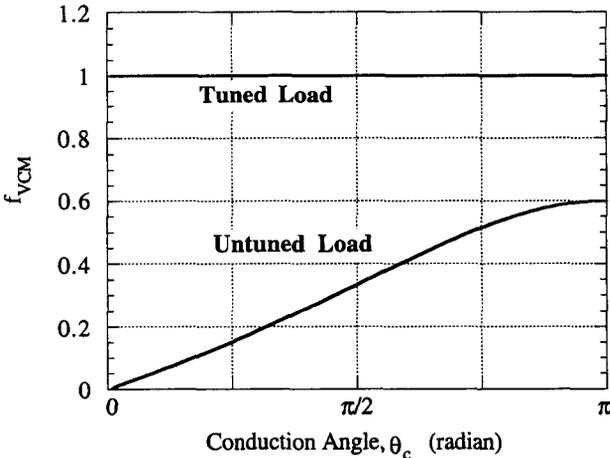


Figure 10.43 f_{VDM} as a function of the conduction angle. The MOS is assumed to be a linear g_m device.

calculated from Eq. 10.69 is $I_{\max}^2/8 \times R_{Lopt} = 0.045 \text{ W}$. $\epsilon_{\text{ff-opt}} = P_{\text{out-opt}} / (I_{\text{dc}} \cdot V_D) = 48\%$.

Suppose now that the same MOS amplifier is operated in class B instead of class A. According to Figure 10.42, at $\theta_c = \pi/2$, R_{Lopt} is $1.33 \times (V_D - V_{\text{knee}}) / I_{\max} = 119.7 \Omega$. $P_{\text{out-opt}}$ from Figure 10.40 is $0.12 \times (V_D - V_{\text{knee}}) \times I_{\max} = 0.027 \text{ W}$. From Figure 10.41, $\epsilon_{\text{ff-opt}} = 48\% \times (V_D - V_{\text{knee}}) / V_D = 43.2\%$.

These calculations demonstrate that both P_{out} and ϵ_{ff} are higher when the device operates in class A rather than class B. This conclusion, however, is restricted to

devices with I - V characteristics governed by Eq. 10.58. Because the drain current varies parabolically with V_{GS} , the transconductance according to Eq. 10.32 increases linearly with V_{GS} . These transistors are classified as the linear g_m transistors. In sub-0.25- μm transistors where velocity saturation is more likely to occur, the drain current often varies linearly with V_{GS} and g_m is a constant. For these constant g_m transistors, P_{out} and ε_{ff} are comparable in both class A and class B operations. The derivation of $P_{\text{out-opt}}$, $\varepsilon_{\text{ff-opt}}$, $R_{L\text{opt}}$, f_{VDM} in the constant g_m transistors resemble that for a linear bipolar transistor whose the output current is linearly proportional to the input voltage.⁹⁹ For a constant g_m MOS operating with $I_{\text{max}} = 50 \text{ mA}$, $V_D = 5 \text{ V}$, and $V_{\text{knee}} = 0.5 \text{ V}$, the power-added efficiencies are 45 and 52% for class A and class B operations, respectively. The large-signal output powers are 0.056 and 0.041 W, respectively.

The conclusion that P_{out} and ε_{ff} are higher in class A for linear g_m MOSFETs is also subject to error when the output load is tuned. Tuning is a circuit technique in which the harmonic contents of the time-varying drain current are selected. Figure 10.38*b* illustrates the circuit arrangement of a tuned amplifier. The load consists of a parallel RLC network (instead of just R). We consider the ideal case in which the LC resonates and becomes open at the fundamental frequency, but acts as a short at other harmonic frequencies. Therefore, while the fundamental component of $i_D(t)$ still flows through the resistive load R_L , the other components of $i_D(t)$ were directly supplied from the ground through the LC circuit, bypassing the resistance. Because all the nonfundamental components of currents are diverted to the ground, only the fundamental component affects $v_D(t)$. The amplifier behaves as though all the harmonic Fourier components were zero. Essentially, all the previous large-signal equations remain valid for the tuned circuit, except that a_2 , a_3 , and so forth appearing in these equations are replaced by zero.

As an example of calculation of the tuned transistor, we again consider the linear g_m MOS transistor operated in class A with $I_{\text{max}} = 50 \text{ mA}$, $V_D = 5 \text{ V}$ and $V_{\text{knee}} = 0.5 \text{ V}$. For this transistor at $\theta_c = \pi$, $I_{\text{dc}} = \frac{3}{8} \times I_{\text{max}} = 18.75 \text{ mA}$ and $a_1 = \frac{1}{2}$. Although some $a_n \neq 0$ for $n \geq 2$, the tuned circuit behaves as if these a_n terms were zero. $R_{L\text{opt}}$ from Eq. 10.71 is therefore $2 \times (V_D - V_{\text{knee}})/I_{\text{max}} = 180 \Omega$. $P_{\text{out-opt}}$ calculated from Eq. 10.69 is $I_{\text{max}}^2/8 \cdot R_{L\text{opt}} = 0.05625 \text{ W}$. $\varepsilon_{\text{ff-opt}} = P_{\text{out-opt}}/(I_{\text{dc}} \cdot V_D) = 60\%$.

If the linear g_m transistor is biased in class B, we find from Eqs. 10.63–10.66 that $I_{\text{dc}} = I_{\text{max}}/4 = 12.5 \text{ mA}$ and $a_1 = 0.42441$. Although the harmonic Fourier components are nonzero, only the fundamental component is passed through the load resistor. The amplifier behaves as though these Fourier components were zero. Setting a_n terms to zero for $n \geq 2$, we find $R_{L\text{opt}}$ from Eq. 10.71 to be $(V_D - V_{\text{knee}})/I_{\text{max}}/0.42441 = 212 \Omega$. $P_{\text{out-opt}}$ calculated from Eq. 10.69 is $0.09 \times I_{\text{max}}^2 \cdot R_{L\text{opt}} = 0.0477 \text{ W}$. Finally, $\varepsilon_{\text{ff-opt}} = P_{\text{out-opt}}/(I_{\text{dc}} \cdot V_D) = 76.3\%$.

Just as with the untuned loads, the performance of the constant g_m MOSFET with tuned loads has been analyzed.⁹⁹ With $I_{\text{max}} = 50 \text{ mA}$, $V_D = 5 \text{ V}$ and $V_{\text{knee}} = 0.5 \text{ V}$, the power-added efficiencies are 45 and 70.6% for class A and class B operations, respectively. The large-signal output powers are both 0.056 W in either class of operation. These results indicate that for constant g_m MOSFET, class B operation

offers superior performance if not competitive, especially when the output load is tuned to permit the current flow of the fundamental component.

The large-signal analyses leave out many details such as the junction capacitances. In general, a sound circuit design requires commercial simulation tools to determine the large-signal performance of the transistor. Nonetheless, this analysis provides insight to the large-signal transistor operation.

10.6 NOISE FIGURE

An amplifier is designed to deliver large-signal output power. Sometimes, the design is subject to the constraint that the electrical noise of the amplifier be below a certain level. We consider the noise performance of a MOSFET whose gate is connected to a power source v_p with a source resistance R_p and whose output is connected to a load resistance R_L . The noise sources and the small-signal model for the transistor are shown in Figure 10.44a. R_p is the resistance associated with the power source, to be distinguished from the source terminal resistance R_S . R_D is the drain resistance. R_G is the poly gate resistance given by Eq. 10.48, not including the channel resistance. We represent C_{gd} between the input and the output nodes by an effective Miller capacitance of $C_{gd} \times (1 + g_m R_L)$ at the input. We shall from here onward refer to the sum of this Miller capacitance and C_{gs} as C_π , as indicated in the figure. As far as the effective resistance seen at the source node is concerned, we replace R_S in Figure 10.44a with an effective source resistance as shown in Figure 10.44b to separate the input from the output. The effective source resistance is $(1 + g_m/j\omega C_\pi)R_S$, amplified by a factor to account for the fact that both the current from the input and from the $g_m \cdot v_\pi$ output current source flow through R_S . If we consider only a small-signal bandwidth (Δf) at a particular frequency f , then the mean-square, noise-signal sources appearing in the figure are given as

$$\overline{v_p^2} = 4kTR_p \Delta f \quad (10.78)$$

$$\overline{v_g^2} = 4kTR_G \Delta f \quad (10.79)$$

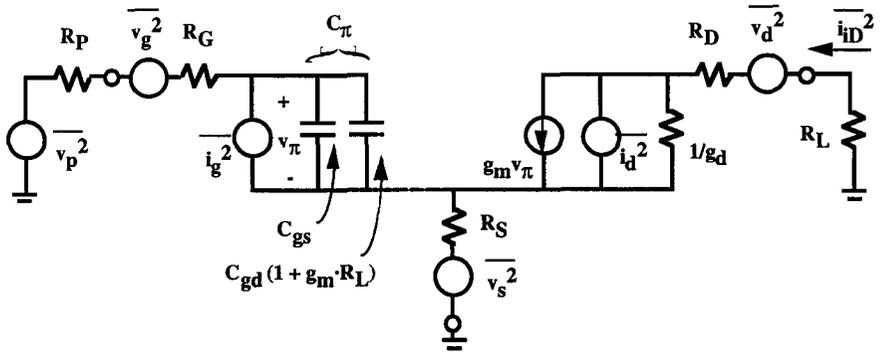
$$\overline{v_d^2} = 4kTR_D \Delta f \quad (10.80)$$

$$\overline{v_s^2} = 4kTR_S \Delta f \quad (10.81)$$

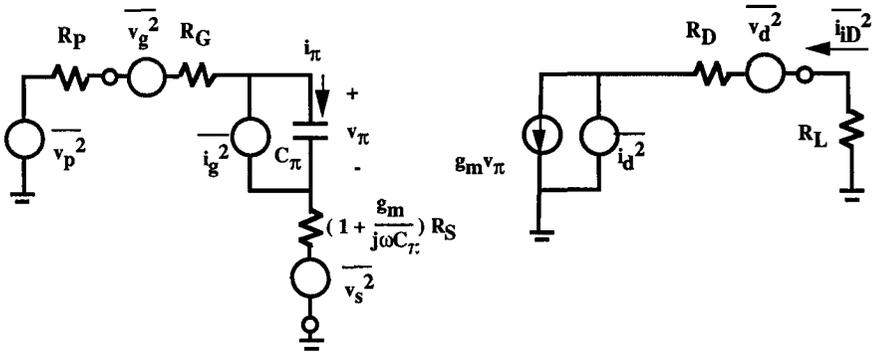
$$\overline{i_d^2} = \frac{K_1}{f^\gamma} \Delta f + \frac{4kT}{R_{td}} \Delta f \quad (10.82)$$

$$\overline{i_g^2} = 2qI_G \Delta f + \frac{4kT}{R_{tg}} \left(\frac{\omega}{\omega_T} \right)^2 \Delta f \quad (10.83)$$

Only the first noise source expressed in Eqs. 10.78–10.83 is extrinsic to the device whereas the rest are intrinsic. γ appearing in Eq. 10.82 is the power exponent of the frequency f , having a value near unity. R_{td} , the thermal-noise channel resistance associated with the drain current, and R_{tg} , the thermal noise channel resistance



(a)



(b)

Figure 10.44 Noise sources and small-signal model for a MOSFET. The noise power of $\overline{v_p^2}$ is associated with the input power resistance R_p . (a) The capacitance C_{gd} between the input and the output nodes is represented by an effective Miller capacitance of $C_{gd} \times (1 + g_m R_L)$ at the input. The sum of this Miller capacitance and C_{gs} is denoted as C_π . (b) An effective source resistance is used to separate the input from the output. The effective source resistance is amplified by a factor to account for the fact that both the current from the input and from the $g_m \cdot v_\pi$ output current source flow through R_S .

associated with the induced gate current, are given by:^{114,115}

$$R_{id} = \left[\frac{2}{3} g_{d0} \frac{1 + \alpha + \alpha^2}{1 + \alpha} \right]^{-1} \tag{10.84}$$

$$R_{ig} = \left[\frac{8C_{ox}^2}{135C_{gg}^2} \frac{g_m^2}{g_{d0}} \times \frac{2\alpha^4 + 10\alpha^3 + 21\alpha^2 + 10\alpha + 2}{(1 + \alpha)^5} \right]^{-1} \tag{10.85}$$

where g_{d0} denotes the drain conductance defined in Eq. 10.32 as $V_{DS} \rightarrow 0$. g_{d0} is equal to $g_m \times (1 + \delta)$ when the transistor is in saturation. These equations apply to long-channel devices. For short-channel devices, some expressions of R_{td} accounting for the electron velocity saturation were reported.^{116,117} Another assumption embodied in Eqs. 10.84 and 10.85 is the neglect of the hot-carrier effects at high electric field.¹¹⁸ The induced charge fluctuations produced by the diffusion noise in the velocity saturation region of the channel has been analyzed elsewhere.¹¹⁹

The terms $\overline{v_p^2}$, $\overline{v_g^2}$, $\overline{v_d^2}$, and $\overline{v_s^2}$ are the thermal noise voltages associated with the power source resistance, and the terminal resistances in the gate, drain and source, respectively. $\overline{i_d^2}$ is the noise associated with the drain current, consisting of two components. The first is the so-called *1/f noise* whose precise origin has not been unequivocally identified,¹²⁰ and the second is the *thermal noise* associated with the conducting channel. The channel resistance is a device resistance whose value is modified by the transistor control mechanism, in contrast to the terminal resistances, which have fixed values independent of the voltages across the two terminals of the resistors. Therefore, the formula for the second component of $\overline{i_d^2}$, modeled with an effective thermal resistance R_{td} , differs from the channel resistance established from the non-quasi-static analysis (see Problem 10.8). $\overline{i_g^2}$, the noise associated with the gate current, also consists of two independent components. The first one is due to the *shot noise* associated with the small but finite gate leakage current. The shot noise is governed by well-established formulas.¹²¹ The second component appearing in $\overline{i_g^2}$ is related to the aforementioned thermal fluctuations in the channel. A fluctuation in the channel charge induces an equal and opposite fluctuation in the charge on the gate electrode, causing a fluctuation current in the gate. The effective noise resistance as result of this capacitive coupling is modeled with R_{tg} .¹²² Because the noises associated with R_{td} and R_{tg} both originate from the fluctuation in the channel, the gate and the drain noise current generators are partially correlated. That is, $\overline{i_g^* i_d} \neq 0$. For simplicity, we shall neglect such a correlation.

We determine the output noise current, assuming that all of six noise sources are uncorrelated because they originate from independent physical mechanisms. We neglect the effects of the output resistance ($1/g_d$) throughout the rest of the section. This does not lead to significant error because the output resistance usually has a large value. As a first step in applying the superposition principle, we short out the noise voltage source in the load as well as all the current sources, and consider the response due to the noise voltage associated with R_p . From Figure 10.44b, the root-mean-square voltage at the input node is

$$\overline{(v_\pi^2)}^{1/2} = \frac{1}{j\omega C_\pi} \frac{1}{R_G + R_P + \frac{1}{j\omega C_\pi} + \left(1 + \frac{g_m}{j\omega C_\pi}\right) R_S} \overline{(v_p^2)}^{1/2} \quad (10.86)$$

The mean-square output drain current due to $\overline{v_p^2}$ is therefore,

$$\overline{i_{iD1}^2} = g_m^2 \frac{\overline{v_p^2}}{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P + R_S)^2} \quad (10.87)$$

Next, we find the mean-square output current due to the gate resistance noise voltage source. This time we short out the source and the load-voltage noise sources and open up the current sources as before. Because $\overline{v_p^2}$ and $\overline{v_g^2}$ are in the same branch, the result is nearly identical to Eq. 10.87:

$$\overline{i_{iD2}^2} = g_m^2 \frac{\overline{v_g^2}}{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P + R_S)^2} \quad (10.88)$$

The contribution of output noise current due to the source noise voltage is obtained in a similar fashion:

$$\overline{i_{iD3}^2} = g_m^2 \frac{\overline{v_s^2}}{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P + R_S)^2} \quad (10.89)$$

To find the output noise current due to the gate current source, we short out all voltage sources and open up the drain current noise source. Now, the branch $R_G + R_P$ is in parallel to the path of C_π and R_S . The current flowing into the branch formed by C_π and $(1 + g_m/j\omega C_\pi)R_S$ is

$$(\overline{i_\pi^2})^{1/2} = (\overline{i_g^2})^{1/2} \frac{R_G + R_P}{R_G + R_P + \frac{1}{j\omega C_\pi} + \left(1 + \frac{g_m}{j\omega C_\pi}\right)R_S} \quad (10.90)$$

Therefore, the output noise current due to such input voltage fluctuation is

$$\overline{i_{iD4}^2} = g_m^2 \frac{(R_G + R_P)^2 \overline{i_g^2}}{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P + R_S)^2} \quad (10.91)$$

The mean square output current due to the drain current noise source is obtained by noting that $\overline{v_\pi^2}$, obtained when the voltage noise sources as well as the gate current noise source are removed, is zero. Therefore, the mean-square output voltage is

$$\overline{i_{iD5}^2} = \overline{i_d^2} \quad (10.92)$$

In the calculation of the mean-square output current due to the drain resistance noise source, we shut off all the other power sources. Since both $g_m \cdot v_\pi$ and $\overline{i_{id}^2} = 0$, there is no current flow through R_D . Hence

$$\overline{i_{iD6}^2} = 0 \quad (10.93)$$

The actual mean-square output current is the sum of the mean-square currents calculated when only one of the noise sources is present. It is therefore

$$\overline{i_{iD}^2} = \sum_{n=1}^6 \overline{i_{iDn}^2} = g_m^2 \frac{\overline{v_p^2} + \overline{v_g^2} + \overline{v_s^2} + (R_G + R_P)^2 \overline{i_g^2}}{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P + R_S)^2} + \overline{i_d^2} \quad (10.94)$$

The spectral density of the output collector current is obtained by substituting Eqs. 10.78 to 10.83 into Eq. 10.94:

$$S_{iD}(f) = \frac{\overline{i_{iD}^2}}{\Delta f} = g_m^2 \frac{4kT(R_P + R_G + R_S) + \left(2qI_G + \frac{4kT}{R_{tg}} \left(\frac{\omega}{\omega_T}\right)^2\right) \times (R_G + R_P)^2}{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P)^2} + \frac{4kT}{R_{td}} + \frac{K_1}{f^\gamma} \quad (10.95)$$

Note that the symbol $S_{iD}(f)$ denotes the spectral density of the transistor drain current. The symbol $S_{id}(f)$ would refer to the spectral density of the shot noise current at the drain. That is, $S_{id}(f)$, according to Eq. 10.82, is equal to $4kT/R_{td} + K_1/f^\gamma$.

It is often convenient to refer the output current noise spectral density to the input. We are interested in finding the equivalent input voltage noise spectral density such that it produces the same output current noise spectral density as shown in Eq. 10.95. Essentially, we combine the effects of the four independent noise sources into one equivalent noise voltage source at the input. The circuit schematic is shown in Figure 10.45. The equivalent input noise voltage satisfies the following relationship:

$$\frac{g_m^2}{|(1 + g_m R_S) + j\omega C_\pi (R_G + R_P + R_S)|^2} \overline{v_{vG}^2} = \overline{i_{iD}^2} \quad (10.96)$$

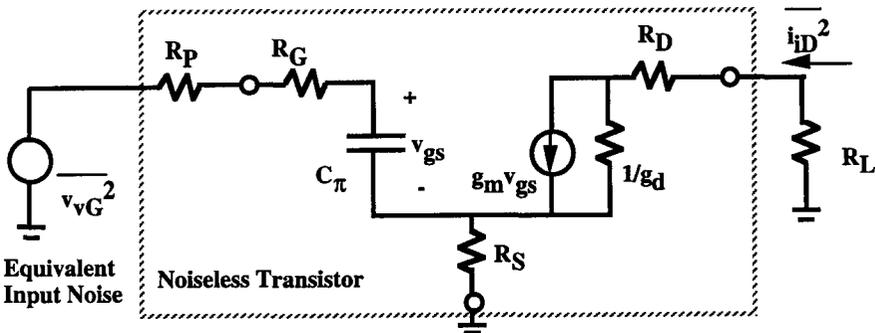


Figure 10.45 A circuit diagram showing the position of the equivalent input-voltage-noise source, in relation to the rest of the circuit.

The equivalent input voltage noise spectral density is obtained by replacing the $\overline{i_{iD}^2}$ of Eq. 10.94 into Eq. 10.96:

$$S_{v_G}(f) = \frac{\overline{v_{vG}^2}}{\Delta f} = 4kT(R_P + R_G + R_S) + \left[2qI_G + \frac{4kT}{R_{ig}} \left(\frac{\omega}{\omega_T} \right)^2 \right] \times (R_G + R_P)^2 + \frac{(1 + g_m R_S)^2 + \omega^2 C_\pi^2 (R_G + R_P + R_S)^2}{g_m^2} \left(\frac{4kT}{R_{id}} + \frac{K_1}{f^\gamma} \right) \quad (10.97)$$

To simplify this expression, we approximate g_m/C_π as ω_T in accordance with Eq. 10.45. We are off by some factor since C_π is not exactly equal to C_{gg} . Therefore

$$S_{v_G}(f) = 4kT \left(R_P + R_G + R_S + \frac{(1 + \xi_m R_S)^2}{g_m^2 R_{id}} \right) + 2qI_G (R_G + R_P)^2 + \frac{(1 + g_m R_S)^2 K_1}{g_m^2 f^\gamma} + \left[(R_G + R_P)^2 \frac{4kT}{R_{ig}} + (R_G + R_P + R_S)^2 \left(\frac{4kT}{R_{id}} + \frac{K_1}{f^\gamma} \right) \right] \left(\frac{\omega}{\omega_T} \right)^2 \quad (10.98)$$

We have separated the frequency dependent term.

An alternative input-noise representation is shown in Figure 10.46. This representation uses two input noise sources: $\overline{E_n^2}$ and $\overline{I_n^2}$, which are inserted between the input voltage source and the gate terminal. The noise voltage associated with the power source resistance is separated out so that $\overline{E_n^2}$ and $\overline{I_n^2}$ represent the noise sources due purely to the intrinsic transistor, unlike the representation of Figure 10.46. By comparing the circuit configurations of Figure 10.47a,b, we find that

$$(\overline{v_{vG}^2})^{1/2} = (\overline{v_p^2})^{1/2} + (\overline{E_n^2})^{1/2} + (\overline{I_n^2})^{1/2} R_P \quad (10.99)$$

The correlation between $\overline{E_n^2}$ and $\overline{I_n^2}$ is assumed to be negligible; hence

$$\overline{v_{vG}^2} = \overline{v_p^2} + \overline{E_n^2} + \overline{I_n^2} R_P^2 \quad (10.100)$$

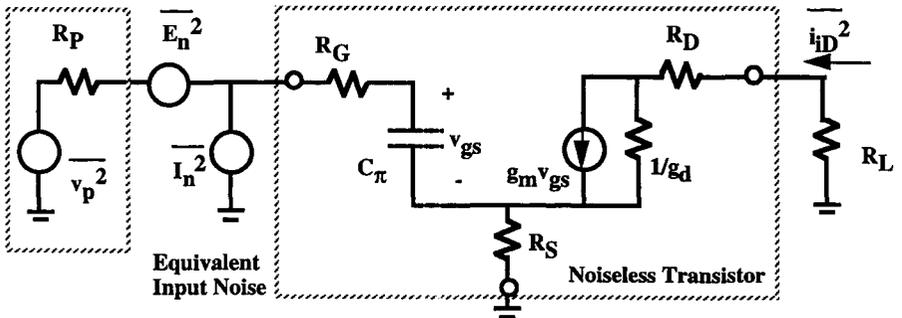


Figure 10.46 An alternative input noise representation using two input-noise sources: $\overline{E_n^2}$ and $\overline{I_n^2}$ inserted between the input voltage source and the gate terminal.

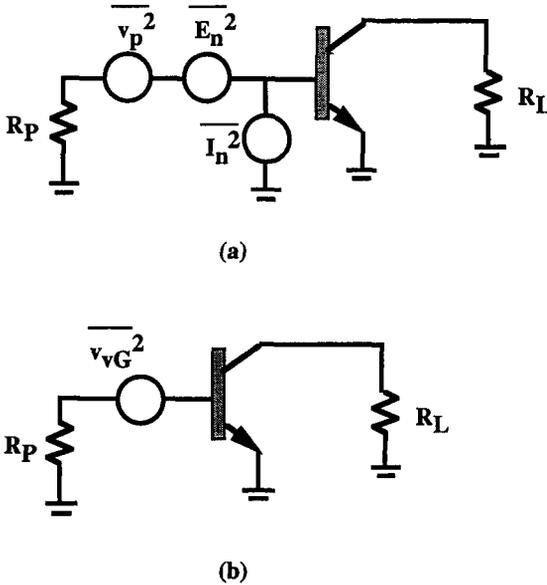


Figure 10.47 (a) Circuit diagram showing the relative position of the noise power source $\overline{v_p^2}$ in relation to the equivalent input noises of $\overline{E_n^2}$ and $\overline{I_n^2}$; (b) circuit diagram showing the replacement of $\overline{v_p^2}$, $\overline{E_n^2}$ and $\overline{I_n^2}$ by $\overline{v_{vG}^2}$.

Representing the noise circuit by $\overline{E_n^2}$ and $\overline{I_n^2}$ has the advantage that they are easily measurable. For example, if R_p is purposely made to be zero, both $\overline{v_p^2}$ and $\overline{I_n^2} \cdot R_p^2$ become zero. The input equivalent-noise spectral density then becomes $\overline{E_n^2}$. In an actual measurement, however, the measurable quantity is the output noise spectral density associated with the load resistance at the output. We therefore need to divide the output noise voltage spectral density by the voltage gain to obtain the input equivalent-noise spectral density, which in turn, is $\overline{E_n^2}$. How about the equivalent input noise current source $\overline{I_n^2}$? We note that $\overline{E_n^2}$ is independent of the source resistance, $\overline{v_p^2}$ is proportional to R_p , and $\overline{I_n^2} \cdot R_p^2$ is proportional to R_p^2 . To determine $\overline{I_n^2}$, we choose a source resistance of an arbitrarily large value. We measure the output noise and divide the result by the voltage gain to obtain $\overline{v_{vG}^2}$ in such a situation. Finally, $\overline{I_n^2}$ is set to be equal to $\overline{v_{vG}^2}$ divided by R_p^2 .

According to the preceding discussion, $\overline{E_n^2}$ is obtained from $S_{vG}(f)$ in the absence of the power-source noise. Taking the limit of $R_p \rightarrow 0$, Eq. 10.98 yields

$$\begin{aligned} \frac{\overline{E_n^2}}{\Delta f} &= 4kT \left(R_G + R_S + \frac{(1 + g_m R_S)^2}{g_m^2 R_{id}} \right) + 2qI_G R_G^2 + \frac{(1 + g_m R_S)^2 K_1}{g_m^2 f^\gamma} \\ &+ \left[R_G^2 \frac{4kT}{R_{id}} + (R_G + R_S)^2 \left(\frac{4kT}{R_{ig}} + \frac{K_1}{f^\gamma} \right) \right] \left(\frac{\omega}{\omega_T} \right)^2 \end{aligned} \tag{10.101}$$

The equivalent input current spectral density is obtained by dividing $S_{vG}(f)$ by R_p^2 and then taking the limit as R_p approaches infinity. Therefore

$$\frac{\overline{I_n^2}}{\Delta f} = 2qI_G + \left(\frac{4kT}{R_{id}} + \frac{4kT}{R_{ig}} + \frac{K_1}{f^\gamma} \right) \left(\frac{\omega}{\omega_T} \right)^2 \tag{10.102}$$

We caution that, when we place Eqs. 10.101 and 10.102 back into Eq. 10.100, the equality is valid only for the two extreme values of $R_p(0 \text{ and } \infty)$. Between these extremes, Eq. 10.100 is actually not correct. A part of the reason is the assumption that the correlation coefficient between $\overline{E_n^2}$ and $\overline{I_n^2}$ is zero at the very beginning of the derivation. Nonetheless, Eq. 10.100 is often used and assumed to be valid for all ranges of power source resistances.

In a typical MOSFET with a reasonable power source resistance R_p , the output noise current is mostly due to $\overline{E_n^2}$ rather than $\overline{I_n^2}$. This is the reason why sometimes only $\overline{E_n^2}/\Delta f$ is reported in the study of FET noise properties. Without $\overline{I_n^2}/\Delta f$, the noise properties of the transistor cannot be analyzed accurately. However, it is still a good approximation to neglect $\overline{I_n^2}/\Delta f$ completely and concentrate on the analysis of the equivalent input voltage spectral density. Figure 10.48 plots the variation of $\overline{E_n^2}/\Delta f$ with frequency, showing the key characteristics of the $1/f$ noise at low frequencies, the white noise that is determined by the thermal noises at medium frequencies, and the increasing noise at high frequencies due to the gain rolloff. In the medium frequencies where the noise is the minimum, the noise is dominated primarily by the thermal noise of the channel and the gate resistance.

Although the equivalent-noise sources are important to the study of the device, often the performance of the transistor is quantified in another figure of merit: the noise figure (NF). It is defined as the ratio (in decibels) of the input signal-to-noise ratio (SNR) to the output SNR. Mathematically, it is expressed

$$NF = 10 \log_{10} \frac{S_o/N_o}{S_i/N_i} \tag{10.103}$$

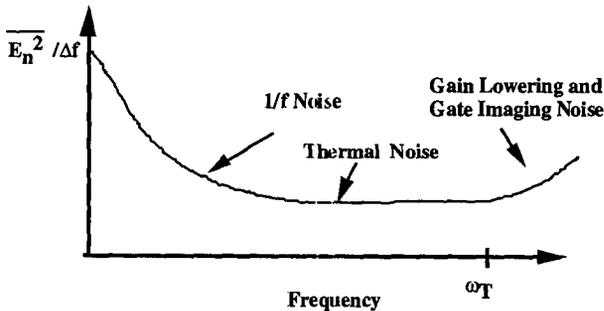


Figure 10.48 Schematic variation of $\overline{E_n^2}/\Delta f$ with frequency, showing the key characteristics of the $1/f$ noise at low frequencies, the thermal noise at medium frequencies, and the increasing noise at high frequencies due to the gain rolloff.

where N_o is the total output noise at the output. It can be taken to be $\overline{i_{iD}^2}/\Delta f$, including the contribution from the noise sources intrinsic to the device as well as the noise associated with the power source resistance. N_i refers to the input noise at the input, excluding the contribution from the device itself. In the context of our discussion, N_i is taken to be $\overline{v_p^2}/\Delta f$. With these choices of input and output noises, we choose the input signal to be the square of the voltage source magnitude ($S_i = \overline{v_p^2}$), and the output signal to be the square of the drain current magnitude ($S_o = \overline{i_D^2}$). Note, for example, $i_D^2 \neq \overline{i_D^2}$. The former refers to the square of the drain current and the latter denotes the noise in the drain current. By examining the circuit of Figure 10.44b, we find the output and input signals to obey the same relationship as Eq. 10.92, with $\overline{v_{vG}^2}$ replaced by S_i and $\overline{i_{iD}^2}$ replaced by S_o . Therefore, we can write

$$\begin{aligned} \text{NF} &= 10 \log_{10} \frac{|(1 + g_m R_S) + j\omega C_\pi (R_G + R_P + R_S)|^2 \overline{i_{iD}^2}/\Delta f}{g_m^2 \overline{v_p^2}/\Delta f} \\ &= 10 \log_{10} \frac{\overline{v_p^2} + \overline{E_n^2} + R_P^2 \overline{I_n^2}}{\overline{v_p^2}} \end{aligned} \quad (10.104)$$

Substituting Eq. 10.100 into Eq. 10.104, we obtain

$$\text{NF} = 10 \log_{10} \left[1 + \frac{1}{4kTR_P} \frac{\overline{E_n^2}}{\Delta f} + \frac{R_P}{4kT} \frac{\overline{I_n^2}}{\Delta f} \right] \quad (10.105)$$

When R_P is large, the third term in the brackets of Eq. 10.105 dominates and the noise figure is proportional to R_P . When R_P decreases to a small value, the second term becomes significant and the noise figure also increases. Between the extreme large and small values of R_P , there is an optimum value of R_P such that the noise figure is the minimum. This optimum resistance is unrelated to the optimum resistance for large-signal power transfer discussed previously.

We express the noise figure in terms of the fundamental device parameters. As mentioned previously, Eq. 10.100 is really just an approximation. To be more accurate, we start out with Eq. 10.98, not equating $\overline{v_{vG}^2}$ to the sum of $\overline{v_p^2}$, $\overline{E_n^2}$, and $\overline{I_n^2} \cdot R_P^2$. Because $\overline{v_p^2}$ is $4kTR_P$, we find the noise figure as

$$\begin{aligned} \text{NF} &= 10 \log_{10} \left[\frac{\overline{v_{vG}^2}}{\overline{v_p^2}} \right] = 10 \log_{10} \left[1 + \frac{R_G + R_S}{R_P} + \frac{(1 + g_m R_S)^2}{g_m^2 R_{iD} R_P} + \frac{2qI_G (R_G + R_P)^2}{4kTR_P} \right. \\ &\quad \left. + \frac{(1 + g_m R_S)^2 K_1}{4kTR_P g_m^2 f^\gamma} + \left(\frac{(R_G + R_P)^2}{R_P R_{iG}} + \frac{(R_G + R_P + R_S)^2}{4kTR_P} \left(\frac{4kT}{R_{iD}} + \frac{K_1}{f^\gamma} \right) \right) \left(\frac{\omega}{\omega_T} \right)^2 \right] \end{aligned} \quad (10.106)$$

The following is an example of applying Eq. 10.106 to find the noise figure. Consider a 8-finger CMOS, where each finger is $0.29 \times 32 \mu\text{m}$. The source and drain

resistances are negligible. The oxide thickness, which is on the order of 60 Å, is thick enough that the gate leakage current is deemed negligible. The gate sheet resistance is 3 Ω/square. The threshold voltage of this technology is around 0.4 V, and the substrate doping is such that the body-effect factor δ is 0.2. During the measurement, the source resistance of the voltage source is chosen to be $R_P = 50 \Omega$. At a bias of $V_{DS} = 2.5 \text{ V}$ and $V_{GS} = 1.4 \text{ V}$, the device attains a g_m of 80 mS and a cutoff frequency of 20 GHz. The transistor operates at 2 GHz, at which frequency the $1/f$ noise is generally considered to be negligible. Therefore, the terms involving K_1/f^γ are omitted in the calculation.

Because $V_{DS} \gg V_{GS} - V_T$, the transistor is biased in saturation. Therefore, $g_{d0} = g_m \times (1 + \delta) = 96 \text{ mS}$ and α obtained from Eq. 10.40 is zero. According to Eq. 10.85, R_{ig} under this bias condition is the inverse of $\frac{16}{135} \times g_m^2/g_{d0} \times C_{ox}^2/C_{gg}^2$. The ratio of C_{ox}/C_{gg} , obtained from Eq. 10.36, is $\frac{3}{2}$. Therefore, R_{ig} is 56.3 Ω. R_{id} , calculated from Eq. 10.85, is equal to $(2/3 \times g_{d0})^{-1}$, or 15.6 Ω. The gate resistance is calculated from Eq. 10.48 as $1/3 \times 3 \times 32/0.29/8 = 13.8 \Omega$. We summarize the parameters needed to calculate the noise figure here: $g_m = 0.08 \text{ 1}/\Omega$; $g_{d0} = 0.096 \text{ 1}/\Omega$; $R_G = 13.8 \Omega$; $R_S = 0$; $R_D = 0$; $R_{id} = 15.6 \Omega$; $R_{ig} = 56.3 \Omega$; $R_P = 50 \Omega$; $I_G = 0$; $K_1 = 0$; $f_T = 20 \text{ GHz}$; $f = 2 \text{ GHz}$; and $kT = 0.0258 \text{ V}$. We calculate the terms included inside the square brackets of Eq. 10.106:

$$\begin{aligned} \frac{R_G}{R_P} &= \frac{13.8}{50} = 0.276 \\ \frac{1}{g_m^2 R_{id} R_P} &= \frac{1}{0.08^2 \times 15.6 \times 50} = 0.2 \\ \frac{2qI_G(R_G + R_P)^2}{4kTR_P} &= 0 \\ \frac{K_1}{4kTR_P g_m^2 f^\gamma} &= 0 \\ \frac{(R_G + R_P)^2}{4kTR_P} \left(\frac{4kT}{R_{id}} + \frac{4kT}{R_{ig}} + \frac{K_1}{f^\gamma} \right) \left(\frac{\omega}{\omega_T} \right)^2 &= \frac{(R_G + R_P)^2 (R_{id} + R_{ig})}{R_P R_{id} R_{ig}} \left(\frac{\omega}{\omega_T} \right)^2 \\ &= \frac{(13.8 + 50)^2 \times (15.6 + 56.3)}{50 \times 15.6 \times 56.3} \left(\frac{2}{20} \right)^2 = 0.067 \end{aligned}$$

Adding 1 to the sum of these numbers, we find the noise figure is 1.87 dB. For this 8-finger device with 32 μm-wide fingers, the noise contribution from R_g is comparable to that from R_{id} . For a 32-finger device with the same total device width (only 8 μm per finger), the noise contribution from the gate resistance becomes negligible.

10.7 SUMMARY AND FUTURE TRENDS

In this chapter, the CMOS design considerations for digital high-performance, low-voltage low-power, and high-frequency operations are reviewed.

For digital CMOS high-speed operations, we first introduced a speed-performance figure of merit (FOM), which in essence is dependent on the ratios of n/p-MOS drive currents over total charge and the RC time constants of n/p-MOS gate electrodes. From the FOM formula and a simple velocity saturation model for the drive current, the device parameters affecting the FOM are identified: L_{gate} , V_{DD} , $I_{\text{OFF}}(\text{max})$, L_{gate} CD control, L_{eff} , V_T , t_{ox} , poly (or gate) depletion, short-channel effects (which is a function of channel dopant profile, t_{ox} , and x_j of the S/D extension), S/D series resistance R_{SD} , gate and S/D region sheet resistance (which need to be $< 7 - 8 \Omega/\text{square}$ for $W/L \leq 20$, and is heavily dependent on the salicide process), effective channel mobility, carrier saturation velocity, gate-to-drain overlap capacitance, junction capacitance, and transistor-width reduction. In addition, the following issues are also very important, normally, boron penetration and direct-tunneling leakage current through the thin gate dielectrics, S/D diode leakage (which is dependent on deep S/D x_j and silicide thickness, and can be improved by advanced structures such as raised S/D, selective salicide, or metal deposition), gate-induced drain leakage, random dopant fluctuation, and device reliability. Because of length limitations and to avoid duplication, most of the device design parameters and the issue of boron penetration are discussed and some current approaches to overcome the limitations are presented in his chapter, while the rest are discussed in other chapters of this book.

For low-voltage low-power CMOS applications, the major challenge is to achieve a certain acceptable speed performance at reduced power supply voltage V_{DD} . The tradeoff between performance and power consumption is certainly application/product-dependent. The considerations for speed performance are the same as those mentioned above, while the addition device design efforts had been concentrated on improving the $I_{\text{drive}}/I_{\text{OFF}}$ tradeoff to improve performance at a given standby power, or to maintain a given performance but with much reduced standby power, and reducing the parasitic capacitance to reduce active power at a given performance. Dual- V_T , low- V_T with adjustable substrate bias, dynamic V_T (DTMOS), fully or partially depleted SOI, SOI on active substrate (SOIAS), and low-temperature CMOS are examples of addressing either or both improving $I_{\text{drive}}/I_{\text{OFF}}$ tradeoff and reducing the parasitic junction capacitance.

The CMOS design considerations for RF applications are different from those for digital high-speed applications. Two figures of merit quantifying the transistor's high-frequency performance are discussed: the cutoff frequency and the maximum oscillation frequency. The small-signal y parameter analysis demonstrates the importance of the drain-to-bulk capacitance in bypassing the output signal into the substrate. A proper substrate resistance network and the non-quasi-static modeling of the channel resistance are crucial to the accurate modeling of the s parameters.

The large-signal operation of a CMOS power amplifier is described using a Fourier analysis. The analysis is based on a CMOS device exhibiting linear g_m transfer characteristics, although the general analytical principles can be applied to arbitrary transfer characteristics. Figures for the output power, the drain efficiency, and the optimum load resistance are presented for any given class of operation (or any conduction angle). The tuning of the harmonic signal at the output load

improves the amplifier performance, but often comes at the expense of increased complexity in circuit design. The power performance is related to the transistor parameters such as the maximum drain current, the knee voltage, and the breakdown voltage.

Finally, the CMOS noise properties are described. The critical noise sources include the channel thermal noise, induced gate noise, the $1/f$ noise, as well as noises associated with the parasitic resistances. The noise figure due to these noise sources is established from an equivalent-circuit analysis. The importance of the noise contribution from the gate resistance of wide transistors is pointed out through an example.

REFERENCES

1. M. Rodder, A. Chatterjee, D. Boning, and I.-C. Chen "Transistor Design with TCAD Tuning and Device Optimization for Process/Device Synthesis," *Proc. 1993 Int. Symp. VLSI Technology, Systems, and Applications*, June 1993, p. 29.
2. M. Rodder, S. Iyer, S. Aur, A. Chatterjee, J. McKee, R. Chapman, and I.-C. Chen, "Oxide Thickness Dependence of Inverter Delay and Device Reliability for 0.25 μm CMOS Technology," *Tech. Digest 1993 Int. Electron Devices Meeting (IEDM)*, Dec. 1993, p. 879.
3. A. Chatterjee, M. Rodder, M. Nandakumar, and I.-C. Chen, "An improved Figure-of-Merit Metrics for CMOS Transistor Performance and its Application to 0.25 μm CMOS Technologies" *Proc. Microelectronic Device and Multilevel Interconnection Technology, 1995 SPIE Symp. Microelectronic Manufacturing (SPIE Vol. 2636)*, Oct. 1995, p. 115.
4. M. Nandakumar, A. Chatterjee, M. Rodder, and I.-C. Chen, "A Device Design Study of 0.25 μm Gate Length CMOS for 1V Low Power Applications," *Tech. Digest 1995 IEEE Symp. Low Power Electronics*, Aug. 1995, p. 80.
5. M. Rodder, A. Amerasekera, S. Aur, and I.-C. Chen, "A Study of Design/Process Dependence of 0.25 μm Gate Length CMOS for Improved Performance and Reliability," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 71.
6. C. Sodini, P. K. Ko, and J. L. Moll, "The Effect of High Fields on MOS Device and Circuit Performance," *IEEE Trans. Electron Devices* **ED-31**(10), 1386 (1984).
7. M.-C. Jeng, J. Chung, J. E. Moon, G. May, P. K. Ko, and C. Hu, "Design Guidelines for Deep-Submicrometer MOSFETs," *Tech. Digest 1988 IEDM*, Dec. 1988, p. 386.
8. S. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.
9. R.-H. Yan, A. Ourmazd, and K. Lee "Scaling the Si MOSFET: from Bulk to SOI to Bulk," *IEEE Trans. Electron Devices* **ED-39**, 1704 (1992).
10. H. Hu, L. T. Su, I. Yang, D. A. Antoniadis, and H. I. Smith, "Channel and Source/Drain Engineering in High-Performance Sub-0.1 μm NMOSFETs Using X-ray Lithography," *Tech. Digest 1994 Symp. VLSI Technology*, June 1994, p. 17.
11. K. Noda, T. Uchida, T. Tatsumi, T. Aoyama, K. Nakajima, H. Miyamoto, T. Hashimoto, and I. Sasaki, "0.1 μm Delta-Doped MOSFET Using Post Low-energy Implanting Selective Epitaxy," *Tech. Digest 1994 Symp. VLSI Technology*, June 1994, p. 19.

12. C. F. Codella and S. Ogura, "Halo Doping Effects in Submicron DI-LDD Device Design," *Tech. Digest 1985 IEDM*, Dec. 1985, p. 230.
13. T. Hori, "A 0.1 μ m CMOS Technology with Tilt-Implanted Punchthrough Stopper (TIPS)," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 75.
14. B. Davari, "CMOS Technology Scaling, 0.1 μ m and Beyond," *Tech. Digest 1996 IEDM*, Dec. 1996, p. 555.
15. L. Su, S. Subbanna, E. Crabbe, P. Agnello, E. Nowak, R. Schulz, S. Rauch, H. Ng, T. Newman, A. Ray, M. Hargrove, A. Acovic, J. Snare, S. Crowder, B. Chen, J. Sun, and B. Davari, "A High-Performance 0.08 μ m CMOS," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 12.
16. M. Luo, P. Tsui, W.-M. Chen, P. Gilbert, B. Maiti, A. Sitaram, and S.-W. Sun, "A 0.25 μ m CMOS Technology with 45 Å NO-nitrided Oxide," *Tech. Digest of 1995 IEDM*, Dec. 1995, p. 691.
17. M. Rodder, S. Aur, and I.-C. Chen, "A Scaled 1.8 V, 0.18 μ m Gate Length CMOS Technology: Device Design and Reliability Considerations," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 415.
18. M. Rodder, Q. Z. Hong, M. Nandakumar, S. Aur, J. C. Hu, and I.-C. Chen, "A sub-0.18 μ m Gate Length CMOS Technology for High-performance (1.5V) and Low Power (1.0V)," *Tech. Digest of 1996 IEDM*, Dec. 1996, p. 563.
19. M. Rodder, M. Hanratty, D. Rogers, T. Laaksonen, S. Murtaza, J. C. Hu, C.-P. Chao, S. Hattangady, S. Aur, A. Amerasekera, and I.-C. Chen, "A 0.1 μ m Gate Length CMOS Technology with 30 Å Gate Dielectric for 1.0–1.5 V Applications," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 223.
20. H. Kawaguchi, H. Abiko, K. Inoue, Y. Saito, T. Yamamoto, Y. Hayashi, S. Masuoka, A. Ono, T. Tamura, K. Tokunaga, Y. Yamada, Y. Yoshida, and I. Sasaki, "A Robust 0.15 μ m CMOS Technology with CoSi₂ Salicide and Shallow Trench Isolation," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 125.
21. Z. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Trans. Electron Devices* **40**(1), 86 (1993).
22. C. H. Wann, K. Noda, T. Tanaka, M. Yoshida, and C. Hu, "A Comparative Study of Advanced MOSFET Concepts," *IEEE Trans. Electron Devices* **43**(10), 1742 (1996).
23. S. Venkatesan, J. W. Lutze, C. Lage, and W. J. Taylor, "Device Drive Current Degradation Observed with Retrograde Channel Profiles," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 1742.
24. T. Ohguro, S. Nakamura, M. Saito, M. Ono, H. Harakawa, E. Morifuji, T. Yoshitomi, T. Morimoto, H. S. Momose, Y. Katsumata, and H. Iwai, "Ultra-Shallow Junction and Salicide Techniques for Advanced CMOS Devices," *Electrochemical Society Proc.*, Vol. 97-3, 275 (1997).
25. S. Shishiguchi, A. Mineji, T. Hayashi, and S. Saito, "Boron Implanted Shallow Junction Formation by High-Temperature/ Short-Time/ High-Ramping-Rate (400°C/sec) RTA," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 89.
26. D. Scott, W. Hunter, and H. Shichijo, "A Transmission Line Model for Silicided Diffusions: Impact on the Performance of VLSI Circuits," *IEEE Trans. Electron Devices* **29**(4), 651 (1982).

27. C. S. Rafferty, H.-H. Vuong, S. A. Eshraghi, M. D. Giles, M. R. Pinto, and S. J. Hillenius, "Explanation of Reverse Short-Channel Effect by Defect Gradients," *Tech. Digest 1993 IEDM*, Dec. 1993, p. 311.
28. A. Boussetta, J. van den Berg, D. Arnour, and P. Zalm, "Si Ultrashallow p⁺n junctions using Low-Energy Boron Implantation," *Appl. Phys. Lett.* **58**(15), 1626 (1991).
29. G. Sai-Halasz and H. Harrison, "Device-Grade Ultra-Shallow Junctions Fabricated with Antimony," *IEEE Electron Device Lett.* **7**, 534 (1986).
30. K. Goto, J. Matsuo, T. Sugii, H. Minakata, I. Yamada, and T. Hisatsugu, "Novel Shallow Junction Technology using Decaborane (B₁₀H₁₄)," *Tech. Digest 1996 IEDM*, Dec. 1996, p. 435.
31. S. Subbanna, et al, "200 nm Process Integration for a 0.15 μm Channel-Length CMOS Technology Using Mixed X-Ray / Optical Lithography," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 695.
32. J. C. Hu, M. Rodder, and I.-C. Chen, "Shallow Source/Drain Extension Formation Using Antimony (Sb) and Indium (In) Pre-amorphization Schemes for 0.18-0.13 μm CMOS Technologies," *Proc. Microelectronic Device Technology, 1997 SPIE Symp. Microelectronic Manufacturing (SPIE Vol. 3212)*, Oct. 1997, p. 136.
33. B. Mizuno, M. Takase, I. Nakayama, and M. Ogura, "Plasma Doping of Boron for Fabricating the Surface Channel Sub-quarter micron PMOSFET," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 56.
34. Y. Nakahara, K. Takeuchi, T. Tasumi, Y. Ochiai, S. Manoko, S. Samukawa, and A. Furukawa, "Ultra-shallow In-situ Doped Raised Source/Drain Structure for Sub-tenth Micron CMOS," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 174
35. Y. Mitani, I. Mizushima, S. Kambayashi, H. Koyama, M. T. Takagi, and M. Kashiwagi, "Buried Source and Drain (BSD) Structure for Ultra-shallow Junction Using Selective Deposition of Highly Doped Amorphous Silicon," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 176.
36. T. Yoshitomi, M. Saito, T. Ohguo, M. Ono, H. Momose, and H. Iwai, "Silicided Silicon-Sidewall Source and Drain (S⁴D) Structure for High-Performance 75 nm Gate Length pMOSFET," *Tech. Digest 1995 Symp. VLSI Technology*, June, 1995, p. 11.
37. M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, and H. Iwai, "Sub-50 nm Gate Length n-MOSFET with 10 nm Phosphorus Source and Drain Junctions," *Tech. Digest 1993 IEDM*, Dec. 1993, p. 119.
38. M. Saito, T. Yoshitomi, M. Ono, Y. Akasaka, H. Nii, S. Matsuda, H. Momose, Y. Katsumata, Y. Ushiku, and H. Iwai, "An SPDD p-MOSFET Structure Suitable for 0.1 and sub-0.1 μm Channel Length and Its Electrical Characteristics," *Tech. Digest 1992 IEDM*, Dec. 1992, p. 897.
39. K. Kramer, S. Talwar, A. McCarthy, and K. Weiner, "Characterization of Reverse Leakage Components for Ultrashallow p⁺/n Diodes Fabricated using Gas Immersion Laser Doping," *IEEE Electron Device Lett.* **17**(10), 461 (1996).
40. M. Mehrotra, A. Chatterjee, and I.-C. Chen, "Sheet Resistance Requirements for the Source/Drain Regions of 0.11 μm Gate Length CMOS Technology," *Proc. Microelectronic Device Technology, 1997 SPIE Symp. Microelectronic Manufacturing, (SPIE Vol. 3212)*, Oct. 1997, p. 162.

41. J. Kittl, D. Prinslow, P. Apte, and M. Pas, "Kinetics and Nucleation Model of the C49 to C54 Phase Transformation in TiSi_2 Thin Films on Deep Submicron n-Type Polycrystalline Silicon Lines," *Appl. Phys. Lett.* **67**(16), 2308 (1995).
42. P. Apte, A. Paranjpe, and G. Pollack, "Use of a TiN cap to Attain Low Sheet Resistance for Scaled TiSi_2 on Sub-Half-Micron Polysilicon Lines," *IEEE Electron Device Lett.* **17**(11), 506 (1996).
43. I. Sakai, H. Abiko, H. Kawaguchi, T. Hirayama, L. Johansson, and K. Okabe, "A New Salicide Process (PASET) for Sub-half Micron CMOS," *Tech. Digest 1992 Symp. VLSI Technology*, June 1992, p. 66.
44. J. A. Kittl, Q.-Z. Hong, M. Rodder, D. Prinslow, and G. Misium, "A Ti Salicide Process for 0.10 μm Gate Length CMOS Technology," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 14.
45. J. A. Kittl, A. Chatterjee, I.-C. Chen, G. A. Dixit, P. P. Apte, D. A. Prinslow, and Q.-Z. Hong, "Study of Integration Issues of Ti Salicide Process with Pre-Amorphization for sub-0.18 μm Gate Length CMOS Technologies," *Tech. Digest 1997 International Symp. VLSI Technology, Systems, and Applications (VLSI-TSA)*, June 1997, p. 23.
46. J.-Y. Tsai and S. Yeh, "Device Degradation Associated with Pre-Amorphization Implant (PAI) of the Ti Salicide Process," *Proc. 1997 Int. Symp. VLSI Technology, Systems, and Applications*, June 1997, p. 28.
47. R. Mann, G. Miles, T. Knotts, D. Rakowski, L. Clevenger, J. Harper, F. D'Heurle, and C. Cabral, "Reduction of the C54- TiSi_2 Phase Transformation Temperature Using Refractory Metal Ion Implantation," *Appl. Phys. Lett.* **67**(25), 3729 (1995).
48. J. A. Kittl, Q.-Z. Hong, C.-P. Chao, I.-C. Chen, N. Yu, S. O'Brien, and M. Hanratty, "Salicides for 0.1 μm Gate Lengths: A Comparative Study of One-step RTP Ti with Mo Doping, Ti with Pre-amorphization, and Co Processes," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 103.
49. J. A. Kittl, Q.-Z. Hong, M. Rodder, and T. Breedijk, "Novel One-Step RTP Ti Salicide Process with Low Sheet Resistance 0.06 μm Gates and High Drive Current," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 111.
50. S. Murtaza, A. Chatterjee, P. Mei, A. Amerasekera, P. Nicollian, J. Kittl, T. Breedijk, M. Hanratty, S. Nag, I. Ali, D. Rogers, and I.-C. Chen, "A Shallow Trench Isolation Study for 0.18 μm CMOS Technology with Special Emphasis on the Effects of Well Design, Channel Stop Implants, Trench Depth, and Salicide Process" *Tech. Digest 1997 Int. Symp. VLSI Technology, Systems, and Applications (VLSI-TSA)*, June 1997, p. 133.
51. T. Yamazaki, K. Goto, T. Fukano, Y. Nara, T. Sugii, and T. Ito, "21 psec Switching 0.1 μm -CMOS at Room Temperature using High-performance Co Salicide Process," *Tech. Digest 1993 IEDM*, Dec. 1993, p. 906.
52. K. Goto, A. Fushida, J. Watanabe, T. Sukegawa, K. Kawamura, T. Yamazaki, and T. Sugii, "Leakage Mechanism and Optimized Conditions of Co Salicide Process for Deep-Submicron CMOS Devices," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 449.
53. Q.-Z. Hong, W. T. Shiau, H. Yang, J. A. Kittl, C. P. Chao, H. L. Tsai, S. Krishnan, I.-C. Chen, and R. H. Havemann, " CoSi_2 with Low Diode Leakage and Low Sheet Resistance at 0.065 μm Gate Length," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 107.
54. K. Inoue, K. Mikagi, H. Abiko, and T. Kikkawa, "A New Cobalt Salicide technology for 0.15 μm CMOS Using High-Temperature Sputtering and In-situ Vacuum Annealing," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 445.

55. T. Mogami, H. Wakabayashi, Y. Saito, T. Matsuki, T. Tatsumi, and T. Kunio, "A Novel Salicide Process (SEDAM) for Sub-quarter Micron CMOS Devices," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 687.
56. H. Abiko, A. Ono, R. Ueno, S. Masuoka, S. Shishiguchi, K. Nakajima, and I. Sasaki "A Channel Engineering Combined with Channel Epitaxy Optimization and TED Suppression for 0.15 μm n-n Gate CMOS Technology," *Tech. Digest 1995 Symp. VLSI Technology*, June 1995, p. 23.
57. H. Wakabayashi, T. Yamamoto, T. Tatsumi, K. Tokunaga, T. Tamura, T. Mogami, and T. Kunio, "A High-Performance 0.1 μm CMOS with Elevated Salicide using Novel Si-SEG Process," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 99.
58. C.-P. Chao, K. Violette, S. Unnikrishnan, M. Nandakumar, R. Wise, J. A. Kittl, Q.-Z. Hong, and I.-C. Chen, "Low Resistance Ti or Co Salicided Raised Source/Drain Transistors for Sub-0.13 μm CMOS Technologies," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 103.
59. M. Sekine, K. Inoue, H. Ito, I. Honma, H. Miyamoto, K. Yoshida, H. Watanabe, K. Mikagi, Y. Yamada, and T. Kikkawa, "Self-aligned Tungsten Strapped Source/Drain and Gate Technology Realizing the Lowest Sheet Resistance for Sub-quarter Micron CMOS," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 493.
60. D. Hisamoto, K. Umeda, Y. Nakamura, N. Kobayashi, S. Kimura, and R. Nagai, "High-Performance Sub-0.1 μm CMOS with Low-Resistance T-Shaped Gates Fabricated by Selective CVD-W," *Tech. Digest 1995 Symp. VLSI Technology*, June 1995, p. 115.
61. R. Achutharaman, P. Hey, and J. L. Regolini, "Selective CVD of Titanium Silicide for Raised Source/Drains," *Semiconduct. Inter.* (Oct. 1996).
62. C. Hu, "Gate Oxide Scaling Limits and Projections," *Tech. Digest 1996 IEDM*, Dec. 1996, p. 319.
63. R. A. Chapman, "Trends in CMOS Process Integration," *Electrochemical Society Proc.*, Vol. 97-3, 1997, p. 413.
64. J. C. Hu, J. Kuehne, T. Grider, M. Rodder, and I.-C. Chen, "A Comparative p MOS Study of 33 \AA Nitrided Oxides Prepared by either N₂O or Nitrogen Implant before Gate Oxidation for 0.18-0.13 μm CMOS Technologies," *Proc. 1997 Int. Symp. VLSI Technology, Systems, and Applications*, June 1997, p. 167.
65. T.-J. King, J. Pfister, J. Shott, J. McVittie, and K. Saraswat, "A Polycrystalline-Si_{1-x}Ge_x-Gate CMOS Technology," *Tech. Digest 1990 IEDM*, Dec. 1990, p. 253.
66. Y. Ponomarev, C. Salm, J. Schmitz, P. Woerlee, and D. Gravesteijn, "High-Performance Deep Submicron MOSTs with Polycrystalline-(Si,Ge)," *Proc. 1997 Int. Symp. VLSI Technology, Systems and Applications*, June 1997, p. 311.
67. D. Lee, K. Yeom, M. Cho, N. Kang, and T. Shim, "Gate Oxide Integrity (GOI) of MOS Transistors with W/TiN Stacked Gate," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 208.
68. Y. Momiyama, H. Minakata, and T. Sugii. "Ultra-Thin Ta₂O₅/SiO₂ Gate Insulator with TiN Gate Technology for 0.1 μm MOSFETs," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 135.
69. H. Yang, G. Brown, J. C. Hu, J.-P. Lu, A. Rotondaro, R. Kraft, I.-C. Chen, J. D. Luttmer, and R. A. Chapman, "A Comparison of TiN Processes for CVD W/TiN Gate Electrode on 3 nm Gate Oxide," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 459.

70. J. C. Hu, H. Yang, R. Kraft, A. L. P. Rotondara, S. Hattangady, W. W. Lee, R. A. Chapman, C.-P. Chao, A. Chatterjee, M. Hanratty, M. Rodder, and I.-C. Chen "Feasibility of Using W/TiN as Metal Gate for Conventional 0.13 μm CMOS Technology and Beyond," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 825.
71. A. Chatterjee, R. A. Chapman, G. Dixit, J. Kuehne, S. Hattangady, H. Yang, G. A. Brown, R. Aggarwal, U. Erdongan, Q. He, M. Hanratty, D. Rogers, S. Murtaza, S. Fang, R. Kraft, A. Rotondara, J. Hu, M. Terry, W. Lee, C. Fernando, A. Konecni, G. Wells, D. Frystak, C. Bowen, M. Rodder, and I.-C. Chen, "Sub-100 nm Gate Length Metal Gate NMOS Transistors Fabricated by a Replacement Gate Process," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 821.
72. K. Uwasawa, T. Mogami, T. Kunio, and M. Fukuma, "Scaling Limitations of Gate Oxide in p^+ Polysilicon Gate MOS Structure for Sub-Quarter Micron CMOS Devices," *Tech. Digest 1993 IEDM*, Dec. 1993, p. 895.
73. B. Y. Kim, I. M. Liu, B. W. Min, H. F. Luan, M. Gardner, J. Fulford, and D. L. Kwang, "Impact of Boron Penetration of Gate Oxide Reliability and Device Lifetime in p^+ -poly PMOSFETs," *Proc. 1997 Int. Symp. VLSI Technology, Systems, and Applications*, June 1997, p. 182.
74. R. Fair, "Unified Model of Boron Diffusion in Thin Gate Oxides: Effects of F, H_2 , N, Oxide thickness, and Injected Si Interstitials," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 85.
75. F. Assaderaghi, D. Sinitsky, H. Gaw, J. Bokor, P. Ko, and C. Hu, "Saturation Velocity and Velocity Overshoot of Inversion Layer Electrons and Holes," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 479.
76. B. Y. Kim, D. Wristers, and D. L. Kwong, "Materials and Processing Issues in the Development of $\text{N}_2\text{O}/\text{NO}$ -based Ultra Thin Oxynitride Gate Dielectrics for CMOS ULSI Applications," *Proc. 1996 Microelectronic Device and Multilevel Interconnect (SPIE Vol. 2875)*, Oct. 1996, p. 188.
77. E. Hasegawa, M. Kawata, K. Ando, M. Makabe, M. Kitakata, A. Ishitani, L. Manchanda, M. Green, K. Krisch, and L. Feldman, "The Impact of Nitrogen Profile Engineering on Ultra-Thin Nitrided Oxide Films for Dual-Gate CMOS ULSI," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 327.
78. Y. Okada, P. Tobin, K. Reid, R. Hegde, B. Maiti, and S. Ajuria, "Furnace Grown Gate Oxynitride Using Nitric Oxide (NO)," *IEEE Trans. Electron Devices* **41**(9), 1608 (1994).
79. C. T. Liu, Y. Ma, K. P. Cheung, C. Chang, L. Fritzing, J. Becerro, H. Lutman, H. Vaidya, J. Colonell, A. Kamgar, J. Minor, R. Murry, W. Lai, C. Pai, and S. Hillenius, "25 \AA Gate Oxide without Boron Penetration for 0.25 μm and 0.3 μm pMOSFETs," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 18.
80. S. Hattangady, R. Kraft, D. Grider, M. Douglas, G. Brown, P. Tiner, J. Kuehne, P. Nicollian, and M. Pas, "Ultra Nitrogen-Profile Engineered Gate Dielectric Films," *Tech. Digest IEDM 1996*, Dec. 1996, p. 495.
81. D. Grider, S. Hattangady, R. Kraft, P. Nicollian, J. Kuehne, G. Brown, S. Aur, R. H. Eklund, M. Pas, W. Hunter, and M. Douglas, "A 0.18 μm CMOS Process Using Nitrogen Profile-Engineered Gate Dielectrics," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 47.
82. X. W. Wang, Y. Shi, T. Ma, G. Cui, T. Tamagawa, J. Golz, B. Halpern, and J. Schmitt, "Extending Gate Dielectric Scaling Limit by Use of Nitride of Oxynitride," *Tech. Digest 1995 Symp. VLSI Technology*, June 1995, p. 109.

83. M. Khare, X. Guo, W. Wang, T.P. Ma, C. Cui, T. Tamagawa, B. Halpern, and J. Schmitt, "Ultra-Thin Silicon Nitride Gate Dielectric for Deep-Sub-Micron CMOS Devices," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 51.
84. C. Bowen, C. L. Fernando, G. Klimeck, A. Chatterjee, R. Lake, D. Blank, J. Davis, M. Kulkarni, S. Hattangady, and I.-C. Chen, "Physical Oxide Thickness Extraction and Verification using Quantum Mechanical Simulation," *Tech. Digest 1997 IEDM*, Dec. 1997, p. 869.
85. Z. Lemnios, "Manufacturing Technology Challenges for Low Power Electronics," *Tech. Digest 1995 Symp. VLSI Technology*, June 1995, p. 5.
86. R.-H. Yan, D. Monroe, J. Weis, A. Mujtaba, and E. Westerwick, "Reducing Operating Voltage from 3, 2, to 1 Volt and Below — Challenges and Guidelines for Possible Solutions," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 55.
87. A. Chatterjee, M. Nandakumar, and I.-C. Chen, "An Investigation of the Impact of Technology Scaling on Power Wasted as Short-Circuit Current in Low Voltage Static CMOS Circuits," *Tech. Digest 1996 Inter. Symp. Low Power Electronics and Design*, Aug. 1996, p. 145.
88. M. Nandakumar, A. Chatterjee, G. Stacey, and I.-C. Chen, "A 0.25 μm Gate Length CMOS Technology for 1V Low Power Applications — Device Design and Power/Performance Considerations," *Tech. Digest 1996 Symp. VLSI Technology*, June 1996, p. 68.
89. Z. Chen, C. Diaz, J. Plummer, M. Cao, and W. Greene, "0.18 μm Dual V_T MOSFET Process and Energy-Delay Measurement," *Tech. Digest 1996 IEDM*, Dec. 1996, p. 851.
90. S. Thompson, I. Young, J. Greason, and M. Bohr, "Dual Threshold Voltages and Substrate Bias: Keys to High-performance, Low Power, 0.1 μm Logic Designs," *Tech. Digest 1997 Symp. VLSI Technology*, June 1997, p. 69.
91. W. Lee, P. Landman, B. Barton, S. Abiko, H. Takahashi, H. Mizuno, S. Muramatsu, K. Tashiro, M. Fusumada, L. Pham, F. Bouteaud, E. Ego, G. Gallo, H. Tran, C. Lemonds, A. Shih, M. Nandakumar, R. Eklund, and I.-C. Chen, "A 1V Programmable DSP for Wireless Communications," *IEEE J. Solid-State Circ.* **32**(11), 1766 (1997).
92. Z. Chen, J. Burr, J. Shott, and J. Plummer, "Optimization of Quarter Micron MOSFETs for Low Voltage/Low Power Applications," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 63.
93. F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. K. Ko, and C. Hu, "A Dynamic Threshold Voltage MOSFET (DTMOS) for Ultra-low Voltage Operation," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 809.
94. T. Andoh, A. Furukawa, and T. Kurio, "Design Methodology for Low-Voltage MOSFETs," *Tech. Digest 1994 IEDM*, Dec. 1994, p. 79.
95. J.-P. Colinge, "Status and Trends in SOI CMOS Technology," *Proc. 1997 VLSI Technology, Systems, and Applications*, June 1997, p. 118.
96. I. Yang, C. Vieri, A. Chandrakasan, and D. A. Antoniadis, "Back Gated CMOS on SOIAS for Dynamic Threshold Voltage Control," *Tech. Digest 1995 IEDM*, Dec. 1995, p. 877.
97. I. Channing, ed., "Advanced Cordless Communications," *Financial Times*, Aug. 11, 1997.
98. R. Schneiderman, "GaAs Continues to Gain in Wireless Applications," *Wireless Syst. Design*, 14–16 (March 1997).

99. See for example, W. Liu, *Fundamentals of III-V Devices: HBTs, MESFETs, and HFETs/HEMTs*, Wiley, New York, 1999 and W. Liu, *Handbook of III-V Heterojunction Bipolar Transistors*, Wiley, New York, 1998.
100. W. Shockley, "A Unipolar Field-Effect Transistor," *Proc. IRE*, **40**, 1365–1376 (1952).
101. J. A. Geurst, "Calculation of High-Frequency Characteristics of Field-Effect Transistors," *Solid-State Electron.* **8**, 563–565 (1965). (This paper deals with JFET or MESFET, rather than MOSFET directly. The governing equations for the MESFET and MOSFET are slightly different, mainly because the channel electric fields have different expressions. The general derivation of this paper, however, is noteworthy.)
102. J. W. Hasleett, and F. N. Trofimenkoff, "Small-Signal, High-Frequency Equivalent Circuit for the Metal-Oxide Semiconductor Field-Effect Transistor," *IEE Proc.* **116**, 699–702 (1969). (This paper assumes the MOSFET to be a three-terminal device, although the solution technique is readily applicable to the four-terminal MOSFET. The general solution given in the paper, in its Bessel series form, is correct. However, the application of the solution to determine the small-signal parameters may have an algebraic error when one high-order frequency term is omitted in the series expansion.)
103. J. J. Paulos and D. A. Antoniadis, "Limitations of Quasi-Static Capacitance Models of the MOS Transistor," *IEEE Electron Device Lett.* **4**, 221–224, (1983).
104. A. Van der Ziel, "Small-Signal, High-Frequency Theory of Field-Effect Transistors," *IEEE Trans. Electron.* **11**, 128–135 (1964). [Like Ref. 101, this paper is concerned with JFET or MESFET, rather than MOSFET. The solution technique to solve a given differential equation, however, is powerful and applicable to other devices as well. This technique has been applied to solve the differential equation for the MODFET, whose governing equation is the same as for a three-terminal MOSFET (neglecting the bulk node). See P. Roblin, S. Kang, A. Ketterson, and H. Morkoc, "Analysis of MODFET Microwave Characteristics," *IEEE Trans. Electron Devices* **34**, 1919–1927 (1987). In this latter paper, k corresponds to $1 - \alpha$ of this chapter. The derivation can be shown to be identical to that Ref. 103 for a three-terminal MOSFET.]
105. J. A. Van Nielen, "A Simple and Accurate Approximation to the High-Frequency Characteristics of IGFETs," *Solid-State Electron.* **12** 826–829, (1969).
106. M. Bagheri and Y. Tsvividis, "A Small-Signal DC-to-High-Frequency Nonquasi-Static Model for the Four-Terminal MOSFET Valid in All Regions of Operation," *IEEE Trans. Electron Devices* **32**, 2383–2391 (1982).
107. Y. P. Tsvividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987.
108. D. E. Ward and R. W. Dutton, "A Charge-Oriented Model for MOS Transistor Capacitance," *IEEE J. Solid-State Circ.* **13**, 703–707 (1978).
109. See, for example, Fig. 8-26 of W. Liu, *Handbook of III-V Heterojunction Bipolar Transistors*, Wiley, New York, 1998. Alternatively, G. Vendelin, *Design of Amplifiers and Oscillators by the S-parameter Method*, Wiley, New York, 1982; p. 13; or G. Gonzalez, *Microwave Transistor Amplifiers*, Prentice-Hall, Englewood Cliffs, NJ, 1984, p. 25, or R. Carson, *High-Frequency Amplifiers*, 2nd ed., Wiley, New York, 1982, p. 200. (Note: There is a sign error in a s_{11} expression in the last reference.)
110. S. J. Mason, "Power Gain in Feedback Amplifier," *Trans. IRE*, **CT-1**, 20–25 (June 1954).

111. W. Liu and M. Chang, "Transistor Transient Studies Including Transcapacitive Current and Distributive Gate Resistance for Inverter Circuits," *IEEE Trans. Circ. Syst. (Part. I: Fundamental Theory Appl.)* **45**, 416–422 (1998).
112. BSIM3 Manual, Dept. Electrical Engineering and Computer Science, Univ. California, Berkeley, 1995.
113. W. Liu, R. Gharpurey, M. C. Chang, U. Erdogan, R. Aggarwal, and J. P. Mattia, "R.F. MOSFET Modeling Accounting for Distributed Substrate and Channel Resistances with Emphasis on the BSIM3v3 SPICE Model," *IEEE Int. Electron Device Meeting Digest*, 1997, pp. 309–312.
114. M. Shoji, "Analysis of High-Frequency Thermal Noise of Enhancement Mode MOS Field-Effect Transistors," *IEEE Trans. Electron Devices* **13**, 520–523 (1966). [The formula given in this chapter is modified from that of this reference because we include a factor $(\omega/\omega_T)^2$ in the equation. In this paper, $\xi_{00}^{1/2}$ can be taken to be 1 and $\xi_L^{1/2}$ can be taken to correspond to α of this chapter. I_s (not really defined in the paper) likely means the maximum current when transistor is in saturation, while I_0 is the typical drain current expression. Therefore, I_0/I_s can be taken to correspond to $(1-\alpha^2)$ of this chapter. The y parameters derived in this paper agree with those of Ref. 107 (Appendix M). The coefficient n_{gg2} appears different because this reference normalizes y_{gg} against g_{d0} while Ref. 107 normalizes y_{gg} against ω_0 . The derivation of R_{td} and R_{tg} is based on the NQS solution of the frequency response of the MOS transistor. A simpler derivation leading to the same result is found in the next reference.]
115. H. Johnson, "Noise in Field-Effect Transistors," in J. T. Wallmark and H. Johnson, eds., *Field-Effect Transistors: Physics, Technology and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1966 Chapter 6. [The parameter η used in the reference is precisely $(1-\alpha)$ of this chapter.]
116. B. Wang, J. Hellums, and C. Sodini, "MOSFET Thermal Noise Modeling for Analog Integrated Circuits," *IEEE J. Solid-State Circ.* **29**, 833–835 (1994). (It seems that this derivation implicitly assumes that the velocity saturation occurs right when the device pinches off at the drain. The fact that the velocity saturation occurs at some distance in front of drain does not seem to be considered.)
117. D. P. Triantis, A. N. Birbas, and D. Kondis, "Thermal Noise Modeling for Short-Channel MOSFET's," *IEEE Trans. Electron Devices* **43**, 1950–1955 (1996). (It seems that this paper does not consider the linear region, but is concerned exclusively with transistors operating in saturation.)
118. F. M. Klaassen, "On the Influence of Hot Carrier Effects on the Thermal Noise of Field-Effect Transistors," *IEEE Trans. Electron Devices* **17**, 858–861 (1970).
119. R. Pucel, H. Haus, and H. Stantz, "Signal and Noise Properties of Gallium Arsenide Microwave Field-Effect Transistors," in L. Marton, ed., *Advances in Electronics and Electron Physics*, Vol. 38, Academic Press, New York, 1975, pp. 195–265.
120. For a list of various theories on the origin of the $1/f$ noise in MOSFETs, see, for example, the following and its references cited there in: G. Reimbold, "Modified $1/f$ Trapping Noise Theory and Experiments in MOS Transistors Biased from Weak to Strong Inversion — Influence of Interface States," *IEEE Trans. Electron Devices* **31**, 1190–1198 (1984).
121. C. D. Motchenbacher and F. C. Fitcher, *Low-Noise Electronic Design*, Wiley-Interscience, New York, 1973.

122. A. Van der Ziel, "Gate Noise in Field Effect Transistors at Moderately High Frequencies," *IEEE Proc.* **51**, 461–467 (1963).
123. A. Van der Ziel, *Noise in Solid State Devices and Circuits*, Wiley, New York, 1986, p. 89.

PROBLEMS

10.1 Calculate the performance figure of merit (FOM) of a CMOS device.

Assume:

- (a) $L_{Gn} = L_{Gp} = 0.25 \mu\text{m}$.
- (b) $t_{\text{ox}}(\text{acc}) = 4 \text{ nm}$ (no poly depletion).
- (c) $V_{DD} = 2.5 \text{ V}$.
- (d) $I_{\text{drive}}^n = 630 \mu\text{A}/\mu\text{m}$, $I_{\text{drive}}^p = 250 \mu\text{A}/\mu\text{m}$.
- (e) $C_{\text{interconnect}} = 1 \text{ fF}$.
- (f) $C_j^{\text{STI sidewall}} = C_j^{\text{gate sidewall}} = 0.4 \text{ fF}/\mu\text{m}$ (for both n^+ and p^+ junctions).
- (g) $C_j^{\text{bottom wall}} = 1 \text{ fF}/\mu\text{m}^2$ (for both n^+ and p^+ junctions).
- (h) $W_p = 2 \cdot W_n$.
- (i) $W_n = L_G/1 \times 10 \mu\text{m}$.
- (j) Distance between gate edge and S/D edge = $2.5 \times L_G$.
- (k) Fanout (FO) = 3.
- (l) The effects of C_{GD} and gate sheet resistance are negligible. Use Eqs. 10.1–10.6 to calculate $C_{\text{gate}}(\text{FO} = 3) = ?$, $C_j = ?$, $C_{\text{total}} = ?$, $\text{FOM}_n = ?$, $\text{FOM}_p = ?$, and FOM (for CMOS) = ?

10.2 Assume that the best fit to the FOM data (shown in Fig. 10.2) can be expressed as $\log_{10}(\text{FOM}) = -\log_{10}(L_G) + 0.453$. Calculate the FOM values for $L_G = 0.25, 0.18, 0.13$, and $0.10 \mu\text{m}$.

10.3 Using the assumptions in questions Problems 10.1 and 10.2 whenever applicable, and assuming that $I_{\text{drive}}^n = 2.3 \times I_{\text{drive}}^p$, calculate the I_{drive}^n and I_{drive}^p for the following nodes:

- (a) $L_G = 0.25 \mu\text{m}$, $V_{DD} = 2.5 \text{ V}$, $t_{\text{ox}}(\text{acc}) = 5 \text{ nm}$.
- (b) $L_G = 0.18 \mu\text{m}$, $V_{DD} = 1.8 \text{ V}$, $t_{\text{ox}}(\text{acc}) = 4 \text{ nm}$.
- (c) $L_G = 0.13 \mu\text{m}$, $V_{DD} = 1.5 \text{ V}$, $t_{\text{ox}}(\text{acc}) = 3.3 \text{ nm}$.
- (d) $L_G = 0.10 \mu\text{m}$, $V_{DD} = 1.2 \text{ V}$, $t_{\text{ox}}(\text{acc}) = 2.6 \text{ nm}$.

10.4 For a dual-gate CMOS, specifically, a n^+ poly on p-type substrate or p^+ poly on n-type substrate, please derive the following relationship (i.e., Eq.10.25):

$$t_{\text{ox}}^2(\text{inv}) = t_{\text{ox}}^2(\text{acc}) + \frac{2\epsilon_{\text{ox}}^2 V_G}{q\epsilon_{\text{Si}} N_{\text{poly}}}$$

where $t_{\text{ox}}(\text{inv})$ is the oxide thickness measured in the inversion region ($= \epsilon_{\text{ox}}/C'_{\text{inv}}$, where C'_{inv} is the gate capacitance per unit area in the inversion region); $t_{\text{ox}}(\text{acc})$ is equal to $\epsilon_{\text{ox}}/C'_{\text{acc}}$; ϵ_{ox} and ϵ_{Si} are the dielectric constant of SiO_2 and Si, respectively; V_G is the gate bias; and N_{poly} is the poly concentration. To derive the equation shown above, the following assumption was used:

$V_{FB} + \phi_s \approx 0$, where V_{FB} is the flat-band voltage and ϕ_s is the surface potential in the substrate in strong inversion.

- 10.5** Following Problem 10.4, please calculate the C_{inv}/C_{acc} ratio due to poly depletion for $t_{ox}(acc) = 7, 6, 5, 4, 3,$ and 2 nm (at $V_G = 2.5$ V, and ignore the tunneling leakage current for the thin oxides):
- (a) for $N_{poly} = 8.3 \times 10^{19}$;
 - (b) for $N_{poly} = 2.9 \times 10^{19}$.

- 10.6** For the gate sheet resistance, derive Eq.10.24:

$$R_{sh} = \frac{\tau_{delay} \cdot \delta}{a \cdot C'_{ox} \cdot (W_n^2 + W_p^2)}$$

- 10.7** The drain current of a long-channel MOSFET is given by

$$I_D = \begin{cases} \frac{W}{L_{eff}} \mu_n C'_{ox} \left(V_{GS} - V_T - \frac{(1 + \delta)V_{DS}}{2} \right) V_{DS} & \text{if } V_{DS} \leq V_{DSat} \\ \frac{W}{L_{eff}} \mu_n C'_{ox} \frac{(V_{GS} - V_T)^2}{2(1 + \delta)} & \text{if } V_{DS} > V_{DSat} \end{cases}$$

- (a) Show that g_m is given by $W/L_{eff} \mu_n C'_{ox} V_{DSat} (1 - \alpha)$.
- (b) Show that τ_1 given by Eq. 10.53 can be written as

$$\tau_1 = \frac{C_{dg} - C_{gd}}{g_m}$$

- (c) Show that $\omega_0 \cdot C_{ox} = g_m / (1 - \alpha)$.
- (d) Find g_d .
- (e) Find g_{d0} , which is the maximum value of g_d , or the g_d value as V_{DS} approaches zero.

- 10.8** The gate-to-bulk capacitance of the MOSFET is given by

$$C_{gb} = C_{ox} \left[\frac{\delta}{3(1 + \delta)} \frac{(1 - \alpha)^2}{(1 + \alpha)^2} \right].$$

- (a) From the fact that $C_{gg} = C_{gd} + C_{gs} + C_{gb}$, find an expression for C_{gs} .
- (b) Substitute the C_{gs} expression and the τ values given in this chapter into the R_{ch} expression. Show that as V_{DS} approaches 0

$$R_{ch} \rightarrow \frac{1}{6g_m} \frac{V_{DS}}{V_{DSat}}$$

- (c) Show that when the transistor is in saturation,

$$R_{ch} = \frac{1}{5g_m}$$

- (d) Assume δ appearing in Eq. 10.40 to be zero. When the transistor is in saturation, show that the channel thermal resistance associated with $\overline{i_g^2}$ is

$$R_{tg} = \frac{135C_{gg}^2}{16g_m C_{ox}^2} = \frac{15}{4g_m}$$

- (e) Assume that the gate leakage current is zero. When the transistor is in saturation, show that the channel thermal resistance associated with $\overline{i_g^2}$ is

$$\overline{i_g^2} = 4kT \frac{4g_m}{15} \left(\frac{\omega}{\omega_T} \right)^2 \Delta f$$

- (f) We shall write the gate current noise source as¹²³

$$\overline{i_g^2} = \beta 4kTR_{ch} g_m^2 \left(\frac{\omega}{\omega_T} \right)^2 \Delta f$$

Show that the β factor is equal to $\frac{4}{3}$. This ratio does not bear much physical significance; it is quoted for mnemonic purposes only.

- 10.9** Consider an MOS amplifier designed for class A operation. Its maximum operating current is $I_{max} = 40$ mA. The circuit is biased at $V_D = 3$ V, and $V_{knee} = 0.5$ V. The amplifier is untuned.

- (a) What is the value of ωt when $i_D(t)$ reaches its dc value of I_{dc} ?
- (b) When $i_D(t)$ is equal to I_{dc} , is $v_D(t)$ equal to V_D ?
- (c) The transistor is operated with the optimum load resistance. What is the load resistance value?
- (d) What is the output power?
- (e) What is the drain efficiency?

- 10.10** Consider an eight-finger MOSFET with $0.29 \times 32 \mu\text{m}^2$ per finger area [$R_{sh} = 5 \Omega/\text{square}$; $R_S = 0 \Omega$; $R_D = 0 \Omega$; oxide thickness = 70 Å; $\alpha = 0$ (device in saturation); $\delta = 0.2$; $g_d = 1.5$ mS, $g_m = 10$ mS]. The gate-to-drain and gate-to-source overlap capacitances is 20% of C_{ox} . Find the cutoff and the maximum oscillation frequencies.

- 10.11** A MOSFET has exactly the same total area as the transistor in Problem 10.10. However, this MOSFET has a wide emitter, consisting of only one finger. Find the cutoff and the maximum oscillation frequencies.

- 10.12** A constant- g_m MOSFET operating in saturation is characterized by

$$i_D(t) = \begin{cases} WC'_{ox}(V_{GS}(t) - V_T - V_{Dsat})v_{sat} & \text{when } v_{GS} \geq V_T \\ 0 & \text{when } v_{GS} < V_T. \end{cases}$$

$$v_{GS}(t) = V_{GS} + v_{gs}\cos(\omega t)$$

where v_{sat} is the saturation velocity. We assume that v_{DS} throughout the RF

cycle exceeds the saturation voltage V_{Dsat} , so that the output voltage v_{DS} does not enter into the drain current calculation.

- (a) Derive an expression for the conduction angle. How does it compare to the conduction angle of the linear g_m MOSFET examined in this chapter?
- (b) Show that the drain current expression can be written as

$$i_D(t) = \begin{cases} I_{\max} \frac{(\cos \omega t - \cos \theta_c)}{(1 - \cos \theta_c)} & \text{when } v_{GS} \geq V_T \\ 0 & \text{when } v_{GS} < V_T \end{cases}$$

- (c) Find I_{dc} and the Fourier coefficients of $i_D(t)$.
- (d) The transistor has a knee voltage of 0.5 V. It is biased for large-signal operation, with $V_D = 10$ V and $I_{\max} = 50$ mA. What are the optimum load resistance, output power, and power-added efficiency in class A operation?
- (e) What are the optimum load resistance, output power, and power-added efficiency in class B operation?

SYSTEM-ON-CHIP CONCEPTS

MARCEL J. M. PELGROM

Philips Research Laboratories,
Eindhoven, The Netherlands

11.1 INTRODUCTION

Advances in technology have generated an abundance of new consumer products. The evolution of a technology from a concept to large-scale consumer good production is illustrated in Figure 11.1. Many technologies have started as an idea, existing only in the minds of writers and their audience. Then technological breakthrough enables the transition into a small-series product. The invention of the transistor allowed a (first) realization of many ideas: from big computing machines to portable audio equipment. These products moved from ideas in the mind of a few to status symbols for many.

Further evolution of technology allowed cost reduction to a level where the original function no longer was the main attraction, (see Table 11.1). As a fashion carrier, the electronic product becomes part of styling and culture such as watches, radios, and mobile phones. A key enabler for this evolution is the refinement of the silicon processing and the introduction of the integrated circuit concept whose evolutionary behavior is perhaps best illustrated by the well-known Moore's law.

Now much electronic equipment is on the edge of the last phase. Further improvements in the technological process will result in such a low price that many products are discarded at the moment convenient to the consumer. Examples can be found in the area of gadgets, cameras, wish cards, and so on. Many more electronic products will move from their present position of fashion to being disposables; such as communication products, amusement items, and games.

The main force behind this process will be the very rapid evolution of microelectronics and its technology. Feature-size reduction allows the compression of many electronic functions into inexpensive, small equipment. Rather complex systems are realized with a small space, power, and cost budgets. The ability to

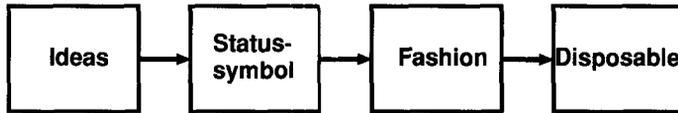


Figure 11.1 Technologies move products from ideas to disposables.

TABLE 11.1 Characteristics of Products during Development of a Technology

Characteristic	Status Symbol	Fashion	Disposable
Excitement	Progress	Features/look	Ease of use
Price	High	Price/performance	Lowest
Quality	No issue	High	Acceptable
Availability	Scarce	Time to market	Everywhere
Advertisement	Free	Right focus	Low profile
Selling theme	Technical	Appearance	Necessity
Life	Long	Fashion period	Short
Service	Full	Marginal	None

realize high-performance signal processing at a moderate cost allows the use of physical media to their limit. Bandwidth limitations in transmission systems are bypassed by smart data reduction schemes, multipoint mobile telephony is feasible by sophisticated time- and code-division multiple access techniques (TDMA and CDMA), and optical and magnetic data storage is driven to the physical limits by advanced servo systems and data coding schemes.

System on chip is the general term that reflects the present status of this progress in microelectronics. System-on-chip applications allow many products to move along the evolutionary line shown in Figure 11.1. It aims at the integration of the complete functionality of regular consumer systems into a very limited amount of hardware, preferably one-chip (see Figure 11.2). Within the various definitions that are used for “system-on-chips” a few main criteria can be distinguished:

- A major part of system functionality is concentrated in one die or one package. For smaller systems this can already be realized to a great extent,

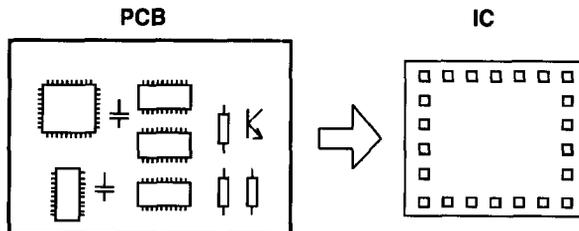


Figure 11.2 Functions shift from PCB into a single chip.

such as radios, and CD players. In other systems the input, power, or output interfaces are rather resistant to integration, such as all cathode-ray-tube (CRT)-related applications, or in applications that require distributed functionality such as automobile electronics.

- There should be a mix of analog, digital, and memory blocks. A part of the technical magic in the term “system on chip” relates to the additional opportunities or difficulties that result when these various functional blocks are put together.

This chapter discusses the underlying techniques. Basically, the device and physics aspects are discussed, while the system aspects are treated by example. High-level aspects, including hardware/software tradeoff, communication structure, mapping of applications on programmable blocks, and test, are discussed in specialized literature.^{1–3}

The next section describes the global outline of the systems-on-chip concept and discuss a number of tradeoffs. Then, the technological requirements for the addition of analog modules in digital CMOS chips are described. The next section focuses on the analog interface, and we discuss a number of basic building blocks. Finally we will describe a number of examples of systems-on-chip.

11.2 EMBEDDED MODULES ON AN IC

11.2.1 Functionality

The clear requirements for an economical, high-performance digital CMOS technology have resulted in a broadly accepted roadmap as formulated by the Semiconductor Industry Association.^{4,5} The *SIA Roadmap* is generally accepted as the dominant direction for CMOS digital technology. It predicts the expected IC products and required material, process, and equipment needs. The roadmap is based on the extrapolation of the reduction of feature size (Moore’s law) and expected computer components.

Much more functionality in addition to the present integration density is expected. On the basis of these possibilities, existing and new electronic consumer systems are evolving at a pace of one generation per year. Table 11.2 shows a number of applications and some characteristic numbers. The resolution (N) refers to the number of binary bits that corresponds to a signal sample, while the sample rate (f_s) indicates the succession speed of the signal samples. The bandwidth together with the resolution reflects the information density of the application.

This list of applications suggests a wide variety of functions that have to be performed. Many applications, however, use only a few basic mathematical functions. If the incoming data at a moment in time $t = nT_s$ are denoted as $X(n)$, which is a sequence of signals samples spaced at sample intervals $T_s = 1/f_s$, and $Y(n)$ is the result value of operations on this sequence, then a number of operations can be formulated.⁶

TABLE 11.2 Various Systems that Are or Will Be Available as One-Chip Solutions

Application	Resolution Bits (N)	Sample Rate f_s (Ms/s) ^a	Bandwidth BW (MHz)	Remarks
Signal processing				
Enhanced TV	9	13.5	5	50 mW power
Digital TV	9	32	15	50 mW power
Camcorder	10	13.5–20	2	10–30 mW
Video PC cards	8–9	32	5	—
PC monitor	10	100–400	40	Output drive
Medical imaging	10–12	40	20	300 mW
Oscilloscopy	8	100–500	250	
Signal transmission				
Teletext	7	27	6.75	Data
DVB 64-QAM, VSB	10	40	4–12	Low jitter
DVB QPSK	6–7	70	20	Low jitter
GSM, digital tele- communication	12	5–20	2	Carrier: 70 MHz
PDA ^b	8–10	10–20	3	10–20 mW
IF conversion AM, FM, TV	10–12	40–100	10.7–38.9	IM3 = 80 dB
Signal storage				
Disk drive	6–8	800	500	
Optical	6–8	250	125	
Solid state	8–64	100–200	—	

^aMillion samples per second.^bPersonal digital assistant.

- Delay:

$$Y(n) = X(n - i) \quad (11.1)$$

where the delay period is given by iT_s .

- Transversal filtering:

$$Y(n) = \sum_{i=0}^{i=k} a_i X(n - i) \quad (11.2)$$

Filtering is a weighted summation of previous incoming data.

- Discrete Fourier transform:

$$Y(m) = \sum_{i=0}^{i=M-1} X(i) e^{-j2\pi mi/M} \quad (11.3)$$

Fourier transformation is a special form of summation of data sequences that results in a mapping of the data in the frequency domain.

- Maximum or minimum:

$$Y(n) = \max, \min_{i=0}^{i=k} X(n - i) \quad (11.4)$$

The maximum or minimum in an interval of $0..k$ samples is obtained.

- Autocorrelation:

$$Y(n) = \frac{1}{M} \sum_{i=0}^{i=M-n} X(i)X(i+n) \quad (11.5)$$

- Comparison:

$$Y(i) = \text{if}(\text{condition} = \text{true})\text{then } X_1(i) \text{ else } X_2(i) \quad (11.6)$$

Many types of signal operations in systems for consumer applications can be mathematically described by these equations. Streams of incoming signal samples undergo a sample-by-sample operation and stream-type operations are again applied to the results. In most consumer systems, there are only a limited number of “irregular” operations in the data flow: operations that require rearrangement of the operation mode. User intervention is less frequent in signal processing systems than in regular software.

This form of data handling, large amounts of identical operations with a little variability, is a major characteristic of signal processing. Nevertheless, complex functions can be composed from Eqs. 11.1–11.6 such as data (de)compression, error (de)coding, signal filtering, two-dimensional (inverse) Fourier transforms. A cascade of these complex functions completed with a controller then define the systems in the above Table 11.2.

11.2.2 Heterogeneous ICs

A digital implementation of these basic functions requires the following three categories of circuits:

- *Conditioning Circuits.* These capture the analog signal from the physical world (an antenna, cable, sensor, etc.) and transform it into bits. After processing, these bits are converted to physical output quantities such as voltage for driving displays or energy for transmitters.
- *Arithmetical Circuits.* Digital circuits to perform the basic operations such as adding, multiplying, counting, and comparing. On a higher layer of hierarchy these elements are combined to function-specific ASICs or more general-purpose microcomputers.
- *Memory Circuits.* Read-only memories (ROM) and random access memories (RAM) for delaying incoming data or storing instructions or fixed data.

Figure 11.3 shows the typical setup for realizing signal processing in electronic equipment. In this heterogeneous IC, dedicated analog circuits (sensors, tuners, etc.)

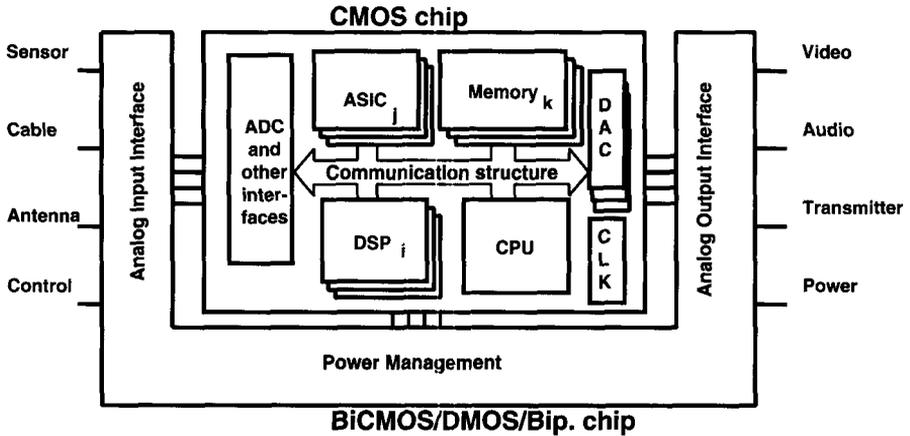


Figure 11.3 General setup of an electronic system. The digital CMOS chip performs signal processing and the analog shell interfaces to input source, power, and outputs.

provide signals to the processing core. The results of the processing are fed to analog output circuits (power amplifiers, display drivers, etc.). Signal processing is done in the inner digital CMOS core. All blocks in Figure 11.3 exchange data over a communication structure. This structure is called “bus,” which was formally a bunch of wires to pass on clock synchronous data. In modern designs, a bus is a complex data handling device, which facilitates high level protocols. The main task in mastering the design of systems-on-chip is handling the complexity of the system. The aspects related to high-level architecture design (communication setup, tradeoffs between hardware and software, etc.) are covered by various studies.^{7,8}

The advances in digital CMOS technology have been focussed on an enhancement of the performance of the digital blocks.⁹ The resulting processing power can be used to expand development in two directions: to increase the performance of a single CPU/memory combination in a single high performance unit or to combine multiple processing units, each tailored to a specific task. For general purpose tasks the wide programmability possibilities of the powerful single-processor approach can be exploited. For more specific tasks, like the processing requirements of a consumer system, the advantages of a heterogeneous IC are clear. Specific hardware executes a regular task much more efficiently than a general system, there is no need for a large-scale operating system, and better cost/performance ratios can be achieved. Nevertheless, the partitioning between hardware and software tasks for the CPU and the DSP is an ongoing debate.

11.2.3 Digital Blocks

Various forms of digital processing modules are used in the setup of Figure 11.3:

- In *application-specific integrated circuits* (ASICs) a function is directly translated in CMOS digital library cells. A description in a high-level

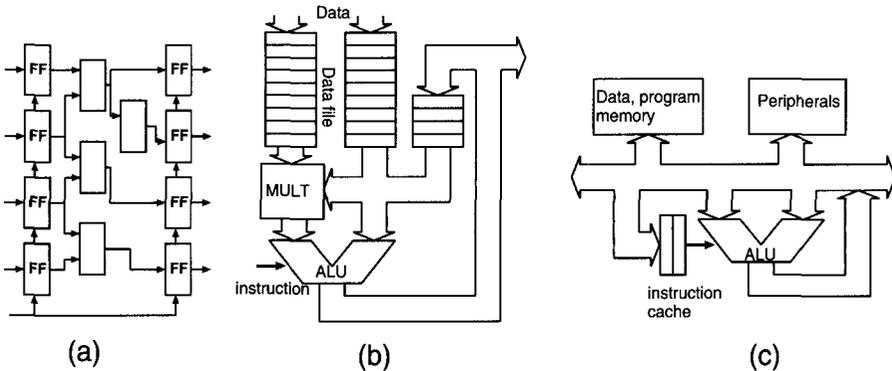


Figure 11.4 Examples of ASIC (a), DSP (b) and CPU (c) architectures.

description language (VHDL) is optimized and mapped onto a number of library cells. The design basically consists of a proper description and the layout route, which is mostly a mechanical process. Programmability of this form of circuit implementation is rather low. Figure 11.4a shows an ASIC example taken from an adder tree.

- The second digital building block is the *digital signal processor* (DSP). This block is structured to handle the typical data manipulations as given in the previous paragraph. There is a strong relation between the required function and the implementation in the DSP. However, some form of flexibility is available, for instance, coefficients can be changed, and delays are variable. Signal busses are separated from instruction busses and often the proper organization of data transport on the busses determines the performance of the DSP. In many system-on-chip applications the DSPs perform the bulk of the data processing. DSPs can be designed optimally for the major data flow: parallel for high-speed or serial for low-frequency operation. Figure 11.4b shows a DSP architecture for operation on two data streams. The data is buffered in a data stack and can be processed by the multiplier or the arithmetic logic unit (ALU). Results can be fed back into the data stream.
- The *central processor unit* (CPU) is a general purpose instruction execution unit. The software provides detailed instructions for every operation. These instructions are fetched from a memory and then executed in an arithmetic logic unit. A CPU can handle the most versatile tasks, albeit at the cost of operation time. The tasks of CPUs in a system on chip configuration are related to irregular functions such as the user interface, control communication, and low data rate signal processing. Figure 11.4c shows a basic CPU. In this example data and instructions pass over the same bus to the shared memory.

Figure 11.5 compares the performance per unit power of three types of processing elements as a function of process generation. The general purpose CPU mostly lags

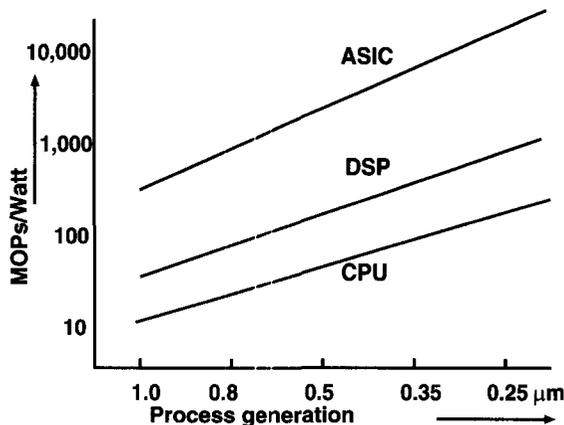


Figure 11.5 Comparison of executable operations for DSPs, CPUs and ASICs, based on examples in literature.

TABLE 11.3 Characteristics of Digital Processing Elements

	ASIC	DSP	CPU
Function	Hard-wired	Weakly programmable	Programmable
Area efficiency	High	Medium	Low
Flexibility	Nil	Low	High
Use	One function	Similar functions	Many functions
Processing	Optimum	Highly parallel	Sequential

behind the DSP by an order of magnitude, while the ASIC that is tailored completely to the function performs best. Table 11.3 lists some characteristic properties of the three processing elements. Most basic operations can be carried out in an ASIC in a single clock cycle, the data width can be optimized locally and wiring can be reduced to a minimum. In a DSP architecture data has to be loaded from the memory units before the operation can be activated, which results in either fewer operations per time unit or, by using parallelism, more energy per operation. In a CPU, often additional time is lost in address computation and bus access. Modern architectures use cache memories and long-instruction-width operation to reduce these effects on operating speed. Experience shows that identical operations in an ASIC and in software on a CPU differ by two orders of magnitude in energy efficiency.*

Many heterogeneous chips have one CPU on board for general control of the functions and for tasks that allow sharing the CPU's hardware. DSPs are loaded with coefficients by the CPU and rearranged if system parameters require it.

* Compare, for instance the hardware implementation of a video decoder¹⁰ with an equivalent software implementation requiring wattage (Watts of power).

TABLE 11.4 Layers of Hierarchy in an Heterogeneous IC^a

	Time	Data Type	Example
Heterogeneous systems	$10^{-1}, 10^{-5}$	Protocol	PDA, GSM phone
Processor	$10^{-4}, 10^{-7}$	Array, list	CPU core
Building block	$10^{-6}, 10^{-8}$	composite signal Integer	SDRAM, ADC ALU block
Cell	$10^{-8}, 10^{-9}$	baseband signal Bit (boolean)	LNA, ROM block Standard cell
Device	$10^{-9}, 10^{-10}$	small signal mV, μ A	diff. pair, RAM cell Transistor wire, antenna

^aAnalog and digital hardware share the same physics to realize an application.

The arithmetic units are complemented with memory blocks. The choice whether a memory block is on or off chip is determined by the balance between the need for many read/write cycles (memory bandwidth) and the more efficient implementation of memories in specialized technologies.¹¹ In some cases this dilemma is solved by implementing the logic blocks in a memory process, called *embedded logic* (see also the discussion in Section 11.6.4).

Table 11.4 analyzes the heterogeneous IC on various levels of hierarchy and indicates complexity and time span. On the device level all building blocks use the standard MOS transistor. However, on the cell level, various elements are combined into building blocks. Various building blocks are used to implement a basic function of the total system.

An analog shell is always needed for conditioning the analog input signals to meet the requirements of the CMOS processing chip and to provide basic conditions for digital operation: clock reference or clock retrieval, power conversion and signal conversion, and conditioning. The conversion and conditioning (filtering, correct dc level, etc.) of signals before the digital blocks can interpret them, is an essential task. This requirement arises both on the input and on the output side of the system chip.

The addition of analog blocks to form a system-on-chip structure in digital CMOS technology sets some additional requirements. In the next section we discuss these requirements.

11.3 TECHNOLOGY FOR MIXED-SIGNAL CIRCUITS

The first-order effect of further integration of functions is the increased power consumption on a die. On a system level power consumption in digital integration is addressed by tackling the major power sinks in the digital circuits. To reduce the power consumption, we reduce clock load, reduce activity in circuits that are temporarily irrelevant, use (partly) asynchronous circuits, optimize algorithms for low-power operation, and so on.

The technological needs of the CMOS digital building blocks are determined by the need for fast switching and low power. This results in a demand for short gate lengths combined with small interconnect capacitors. The low area requirements push toward 4–6 layers of interconnect. Low standby currents require relatively high threshold voltages, ranging from 0.5 V in ASICs to 1 V in memories. The speed in logic circuitry is, in first order, determined by the difference between power supply and threshold voltage $V_{dd} - V_T$. So a compromise on the threshold voltage has to be found. A threshold voltage of approximately $0.2 \times V_{dd}$ is often seen. An additional way to maintain the speed without power loss is to multiplex critical paths.

Once the system aspects of an IC have become clear and the system designer has fixed several boundary conditions, some important options are still open that will influence the performance of analog interfaces. Some choices relevant for digital circuits are also in line with the needs of analog interface circuitry such as the need for accurate models.^{12,13} In other aspects analog designs pose more or opposing demands on the technology, including power supply voltage, characterization, and passive components. This section discusses the choices in technology for the analog interface. The focus will be on CMOS technology, but many arguments will also hold equally for other technologies. The technological items that are important for the designer of analog interfaces, are signal swing, feature size (speed), process options, and tolerances.¹⁴

11.3.1 Signal Swing

The performance of many analog circuits is limited by signal swing. Figure 11.6 shows the expected power supply voltage according to the SIA (bold line) for desktop and battery applications.⁴ The available signal swing (dotted line) is derived by taking 90% from the nominal power supply (minimum power supply) and subtracting a maximum threshold voltage (0.5 V) and 0.2 V gate bias. The graph clearly shows that this form of signal handling has reached its limits. Besides a low

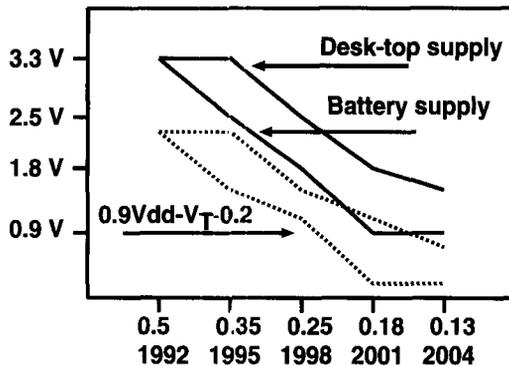


Figure 11.6 Evolution of the power supply voltage as a function of process generation. The signal swing (dotted line) is based on a minimum power supply minus a threshold and a bias voltage.

V_T option or a local V_{DD} boost, the most in suitable form for signals on mixed-signal chips is the differential mode. This form of signal is less sensitive to substrate noise and inherently cancels even harmonic distortion, but requires differential inputs and outputs, two-pin analog I/Os. Despite these measures, signal swing is expected to decrease from 2 to some $0.5\text{--}0.75 V_{\text{peak-peak}}$ in minimum power supply applications. This decreased swing leads to reduced performance as the thermal noise level remains constant at constant power consumption, whereas the digital induced noise level will increase because of extended complexity. Further limits with respect to the reduced signal swing are discussed in Section 11.4.

11.3.2 Feature Size

In digital CMOS technology the drive for smaller feature sizes is, of course, a dominant factor. For low-power operation the development in capacitances is of particular importance, because low capacitance values lower the required power. For digital power consumption the general law in CMOS:

$$\text{Power} = f_s N_i (C_{\text{load}} + C_{\text{gate}}) V_{DD}^2 + f_s V_{DD} I_{sc} + P_{\text{leak}} \quad (11.7)$$

The power is composed of the capacitor charging component and (usually of minor importance): a short-circuit component I_{sc} that occurs when both the NMOST and PMOST are conducting, and leakage component consisting of residual current in a switched-off transistor and junction leakage. N_i is the number of transitions per operation (activity), which in a worst case, may exceed 1 due to, for instance, carry mechanisms. C_{load} represents the node capacitance, which is strongly affected by the technology and the feature size. The close relation between the power consumption and the supply voltage V_{DD} explains the drive for lower voltages (see Fig. 11.6). In analog circuits, feature-size reduction promises a higher cutoff frequency for the same current. However, short-channel drain feedback at minimum gate lengths will reduce the DC gain and thereby limit the use of minimum gate lengths.

Figure 11.7a shows the evolution of gate and diffusion capacitors. These capacitors dominate the internal performance of analog building blocks, whereas in digital building blocks the wiring is of prime importance. The capacitance of a gate $1\ \mu\text{m}$ wide and of minimum gate length shows that the reduction in gate length is largely compensated by the thinner gate oxide. The diffusion capacitance is determined, at zero junction voltage, by a $1\text{-}\mu\text{m}$ -wide slice of diffusion that allows one contact hole to be placed using minimum-dimension design rules. Incorporated are $1\ \mu\text{m}$ of gate edge and LOCOS edge capacitors. The improvement of the p diffusion in the 0.35- and $0.6\text{-}\mu\text{m}$ processes is due to the use of lower doped twin wells. The n diffusions, however, increase in capacitance.

The ratio of transconductance and load capacitance results in a cutoff frequency ($f_i = g_m/2\pi C$). Figure 11.7b compares the cutoff frequencies for diffusion, gate, and $10\text{-}\mu\text{m}$ densely spaced interconnect. The speed performance determined by diffusion capacitance shows only little improvement. Note that these cutoff frequencies were determined at $V_{GT} \approx 0.2\text{--}0.4\ \text{V}$ used by analog designers, whereas

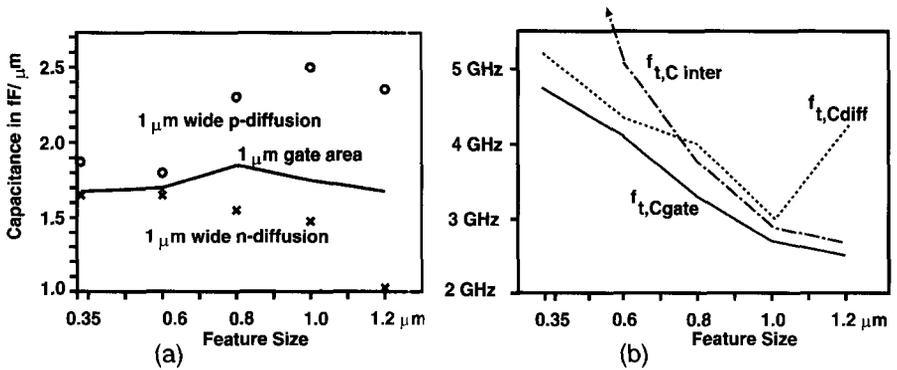


Figure 11.7 Evolution of gate and diffusion capacitance (a) and the cutoff frequency f_i (b) for a minimum gate length NMOS transistor biased at 5 μ A current/ μ m gate width.

technologists usually use much higher V_{GT} (>2 V) to show maximum cutoff frequency of their technology. Such high V_{GT} values are used only in specific circuits such as RF circuits and CMOS low noise amplifiers (LNA).

The improvement in the speed for interconnects is notable, representing an increase of about 30% for every generation. The so-called interconnect crisis is therefore more of a design automation (placement) issue than a technological one. To a lesser extent the speed improvement also holds for minimum-length gate capacitance. If, however, dc gain must be kept constant to maintain a circuit's accuracy specification, the gate length cannot be scaled with technology and the gate capacitance cutoff frequency will not improve significantly. These observations must be taken into account when comparing digital and analog speed performance.

11.3.3 Process Options

Most manufacturers offer two or more generations of a process: a mature process with a large feature size, an advanced process recently released for production, and experimental processes. However, a design team is rarely able to choose the main line of technology to be used to design an integrated circuit. Issues like specific demands, availability, experience, CAD, library issues, and cost dominate the choice of the main process line. One level below that choice there is much more freedom. Baseline processes allow a variety of process options in wiring layers, in additional elements (capacitors, resistors, fuses), and in device parameters. Process options are often used to circumvent the inherent drawbacks of baseline digital CMOS processing. Options, moreover, allow designers to map the function better on the process that allow a reduction in the amount of power consumed. Options are defined as everything other than that offered by a baseline process and can be subdivided into four categories:

- *No Extra Masks, No Process Adaptations.* Analog characterization and monitoring of specific process parameters is most important. Examples are

transistor noise, matching, temperature, and voltage dependencies of passive and active components. Proper characterization (of, e.g., matching) immediately allows a design for optimum performance. The reduction in process spreads also falls in this category (see Section 11.3.4);

- *No Extra Masks, Minor Process Adaptations.* Redefinition of transistor threshold voltages is crucial for low-voltage operation. Low threshold voltages enlarge the available signal swing when the transistor is used as input device or as a switch. Analog designers also want to use structures that are normally only parasitic devices in a digital circuit in the baseline process. This does not introduce new mask steps, but may require another combination of masks. Examples are vertical and lateral PNP bipolar transistors, interconnect used as resistors and interconnect stacks used as capacitors;
- *Extra Masks, Minor Process Adaptations.* Creation of devices with differing parameters, especially a second threshold transistor in order to optimize performance for analog circuits. Definition of resistors in a silicide process fit into this category;
- *Extra Masks, Major Process Adaptations.* The addition of new elements may necessitate new mask levels. This is certainly the most critical category of process options because it affects the costs. Examples are a second gate oxide for high analog voltages and a second poly layer. A double-poly capacitor has a decade less parasitics than a stacked-layer capacitor, so the same function is realized more efficiently.

The lists of examples in these categories can, of course, be expanded. The first category is needed in all industrial analog design. The second category is also very often needed for A/D designs. The use of the third and particularly the fourth category (adding masks) is disputable. Here the analog designer's wishes will inflict a cost penalty upon the digital part of the design. This will rarely be acceptable for a commercially viable application.

11.3.4 Tolerances

The absolute tolerances of the components, the power supply variation, and the temperature span in which a circuit must operate strongly influence the performance. Standard specifications for the power supply tolerances are $\pm 5\%$, $\pm 10\%$, and $\pm 20\%$. Similar variations are observed in the spreads of process capacitances and MOST (MOS transistor) currents. The temperature range may vary from $0\text{--}70^\circ\text{C}$ to -40°C to 140°C . The circuit nevertheless has to function and meet the specifications. Although these three factors influence all aspects of the IC design, here the attention is on the effect on speed performance versus power.

One of the critical points in a design is speed optimization. Suppose that a MOS transistor has to (dis) charge a load capacitor C_{load} within a fraction $\alpha \ll 1$ of the clock period. The time constant is given by the transistor's on resistance and the load

capacitor:

$$R_{on}C_{load} \approx \frac{C_{load}}{(W/L)C_{ox}\mu(V_{DD} - V_T)} = \alpha T_{clock} \tag{11.8}$$

where C_{ox} is the gate oxide capacitance per unit area and W, L are the gate width and length, which leads to the following relation for the minimum gate area needed in the nominal case:

$$C_{gate,nom} = WLC_{ox} = \frac{C_{load}L^2}{\alpha T_{clock}\mu(V_{DD} - V_T)} \tag{11.9}$$

If not restricted by other demands, the designer will minimize the load capacitance and transistor gate length and maximize the available time and drive voltage. The speed optimization of this component becomes critical when worst-case figures replace the ideal-case figures:

$$C_{gate} > \frac{C_{load,max}L_{max}^2}{\alpha T_{clock,min}\mu_{rain}(V_{DD,min} - V_{T,max})} \tag{11.10}$$

where L_{max} is the maximum value of the minimum gate length. A worst-case inspection of the required overdesign factor adds up the tolerances on the load capacitor, the gate length, the clock skew, or the clock duty-cycle variation, the mobility as a function of temperature and the power supply. Table 11.5 compares large, medium, and small tolerances with respect to nominal conditions for power, temperature, and device parameters.

Because the power per transition is the well-known relation $power = (C_{load} + C_{gate})V_{DD}^2$, it is obvious that the increase in C_{gate} to a 2–7 times larger value due to tolerances is a major power problem in both analog and digital designs. This argument plays a role in the specification of both the system and the IC.

TABLE 11.5 Increase in Gate Capacitance of Driving Transistors Due to Parameter Tolerances

Parameter Variation	Nominal	Small	Medium	Large
C_{load} wiring/diffusion	0%	+ 10%	+ 20%	+ 30%
L minimum length	0%	+ 10%	+ 20%	+ 30%
T_{clock} clock skew	0%	- 10%	- 20%	- 30%
Temperature T	25°C	85°C	120°C	140°C
μ mobility $\propto T^{-2}$	0%	- 30%	- 42%	- 48%
$V_{DD}-V_T$ gate drive	0%	- 5%	- 10%	- 15%
Effect on C_{gate}	1 ×	2.2 ×	4.1 ×	7.1 ×
	$C_{gate,nom}$	$C_{gate,nom}$	$C_{gate,nom}$	$C_{gate,nom}$

11.4 TECHNOLOGY LIMITS

11.4.1 Systematic and Random Mismatch

Many analog circuits rely on the property of technology that identical layout structures are also identical in an electrical sense. The electrical quantities voltage, charge, or current are derived from basic components like resistors, capacitors, transistors, and units of time. The equality of the electrical quantities depends strongly on the reproducibility of the basic components. Deviations in the reproducibility of the basic components are described by “systematic mismatch” and “random mismatch”: unwanted differences in the effective value of equally designed components.^{15–21} *Systematic mismatch* (offset) is defined as the nonzero mean value of the difference in value between many pairs of components, while *random matching* refers to the stochastic spread σ (see Fig. 11.8). Systematic and random mismatch in the basic components is caused by deviations in the fabrication process, deviations in the electrical conditions during use, or unequal timing moments. *Mismatch* will show as a Gaussian distribution if the source of mismatch is composed of many mutually independent events such as implantation of ions.

For electrical matching of the basic components, the voltages on all the elements must all be the same (equally affected by voltage drops in power lines, leakage currents in diodes, etc.), the substrate coupling must be the same, parasitical components must be matched, and so on. In extended analyses also electrically derived effects must be considered, such as, heat gradients due to dissipation, aging of components, and threshold drift due to hot-carrier effects. Especially in high-frequency A/D conversion, the “conversion units” are based on the reproduction of physical basic components, while low-bandwidth converters rely strongly on timing accuracy. Examples of systematic mismatch in basic components due to lithographical and chemical effects in the fabrication process are

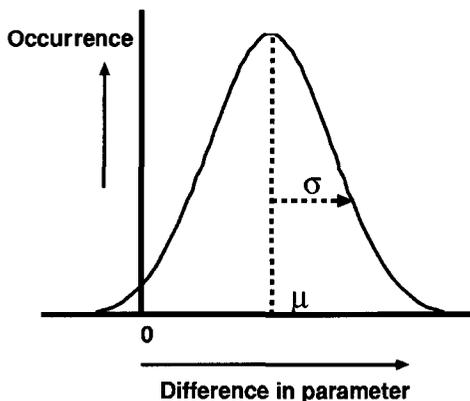


Figure 11.8 Systematic mismatch μ and random mismatch σ of differences in parameters of two identically designed components.

- The proximity effect. The line width is enlarged by diffused light from neighboring structures. The standard procedure is to arrange dummy structures at distances of up to 20–40 μm .
- Gradients in doping, resistivity, and layer thickness. Although structures tend to decrease in dimensions, situations may occur in which equality is required over a distance in the order of millimeters. In CMOS thresholds may deviate up to 5–20 mV over this distance, and resistivity gradients may reach the percentage level. Common centroid structures are used to reduce these effects;
- Patterns in one layer may affect the patterning in other layers. During the spinning of the resist fluid, resist may accumulate at altitude differences (resulting from previously formed structures) on a partly processed wafer. This effect results in circular gradients and is, in the stage of final product, not recognized as a systematic offset.
- Matched devices should not be covered by other structures.²²
- During the ion-implantation steps, the implantation beam is tilted about 5–8° to prevent too deep penetration of the ions into the lattice; this effect is called “channeling.” As a result of this nonperpendicular implantation, source and drain diffusions may be asymmetrical; one may extend further underneath the gate than the other. To prevent inequalities in currents or overlap capacitors, the directions in which the MOS currents flow must be chosen to run parallel, and not rotated or anti-parallel.

Systematic deviations will have to be identified and measures for overcoming them will have to be found in extensive study of the fabrication process. Table 11.6 gives a short guideline for proper design of equal structures.

11.4.2 Component Matching

Many forms of systematic mismatch can be circumvented at the expense of design effort. Random component mismatch is more difficult to cope with through design.

TABLE 11.6 Guidelines for Layout to Prevent Systematic Mismatch

-
1. Matched devices are of the same type, size, and shape.
 2. Operating voltages, currents, and temperatures are as similar as possible.
 3. Currents in matched devices run in parallel, not perpendicular or antiparallel.
 4. Use cross-coupling only if there is evidence for a gradient (in doping, temperature, etc.). Often poorly laid-out cross-coupled structures create more problems than they solve.
 5. Do not overlay matched structures with other materials such as wiring.
 6. Put dummy structures up to 20 μm of the matched devices.
 7. Use star-connected wiring for power and timing sensitive signals.
 8. Keep track of the current paths, from the point where the current flows into the IC to the external ground. Do not interconnect the power wiring of different blocks.
 9. Check voltage drops in power lines and connecting vias.
 10. Stay 200 μm away from chip edges.
-

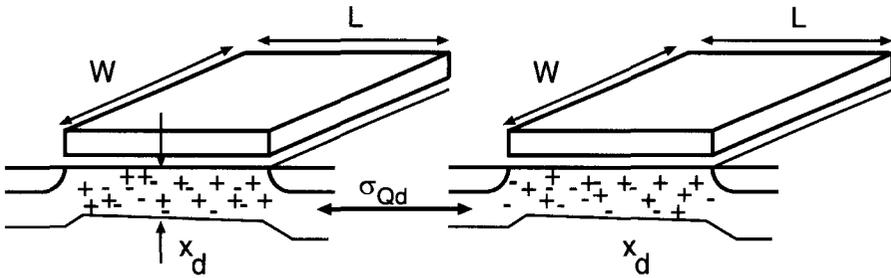


Figure 11.9 Variation in fixed charges in the depletion region of two MOS transistors determine a major part of the threshold (mis)match.

The basic problem in random offset concerns the local variations that occur in the fabrication process. Granularity occurs on linear structures (resistors, capacitors) in the form of edge roughness of polysilicon lines. In MOS circuit design the effect of two-dimensional local variations is the dominant cause of random offset in transistors. The basic principle behind random offset in the threshold definition of MOS transistors is the number of fixed-charged atoms in the active and depletion region (see Fig. 11.9). These charged atoms (dopants, dislocations, oxide charge, etc.) are implanted, diffused or generated during the manufacturing process, but not in an atom-by-atom controlled manner. The average value is controlled by implantation dope levels or average substrate dopes. The actual number of carriers in a particular depletion region may well differ from this average. In this analysis we assume that the presence of dopants is governed by a Poisson process; the presence or absence of an individual atom is independent of the presence of other charges. During the operation of a MOS transistor with well-controlled voltages, the balance of all charges and potentials will determine the channel charge in that transistor. If that channel charge varies from one transistor to another due to a varying number of depletion or implanted charges, the threshold voltage will vary accordingly. The threshold voltage is given by

$$V_T - V_{FB} = \frac{Q_B}{C_{ox}} = \frac{qN_x x_d}{C_{ox}} = \frac{\sqrt{2q\epsilon N_x \phi_b}}{C_{ox}} \tag{11.11}$$

where ϵ is the permittivity, N_x the dopant concentration, and ϕ_b the built-in potential; x_d is the depletion width: $\sqrt{2\epsilon\phi_b/qN_x}$. If the depletion area of a transistor (see Fig. 11.9) is defined by a width W , length L , and a depletion region depth x_d , then the volume of the depletion region is (in first order): $W \times L \times x_d$. Different impurities are active in this region: $N_x \approx 10^{16} - 10^{17} \text{ cm}^{-3}$. N_x contains acceptor and donor atoms from the intrinsic substrate dope, the well, threshold adjust, and punchthrough implantations.* In the variance analysis it is important to note that the total number

*For ease of understanding only a uniformly distributed dopant is used here; more complicated distributions must be numerically evaluated.

of charged atoms must be considered, not the net resulting charge. The standard deviation of this amount of charge in a Poisson process is now approximated by

$$\sigma_{WLxN} = \sqrt{WLx_dN_x} \tag{11.12}$$

The threshold variance can be derived from Eq. (11.11) by considering that

$$\sigma_{\text{single},VT} = \sigma_{WLxN} \frac{\delta(V_T)}{\delta(WLx_dN_x)} \tag{11.13}$$

As matching usually occurs between pairs of transistors, the variance of the difference between two transistors is

$$\sigma_{VT} = \sqrt{2}\sigma_{\text{single},VT} = \frac{qt_{ox}\sqrt{2N_x x_d}}{\epsilon_{ox}\sqrt{WL}} = \frac{A_{VT}}{\sqrt{WL}} \propto \frac{t_{ox}\sqrt{N_x}}{\sqrt{WL}} \tag{11.14}$$

The linear relation between σ_{VT} and $1/\sqrt{\text{area}}$ is well known.¹⁸ Figure 11.10a shows the measured dependence for σ_{VT} versus $1/\sqrt{\text{area}}$.

In order to test the hypothesis that depletion charge is the dominant factor in threshold matching, Table 11.7 compares the A_{VT} coefficients as measured and as calculated using Eq. 11.14. The quantity $N_x x_d$ was derived in process simulation that was tuned with accurate C/V measurements.^{14,23} In the case of three out of the four coefficients the fit is good; the deviation of the 0.6- μm NMOST was caused in this experiment by annealing problems. The large PMOST coefficient of the 0.8- μm PMOSTs is caused by the compensation that occurs between the N- and PMOST threshold adjust and n-well implants. The total dopant quantity $N_x = (N_a + N_d)$ is

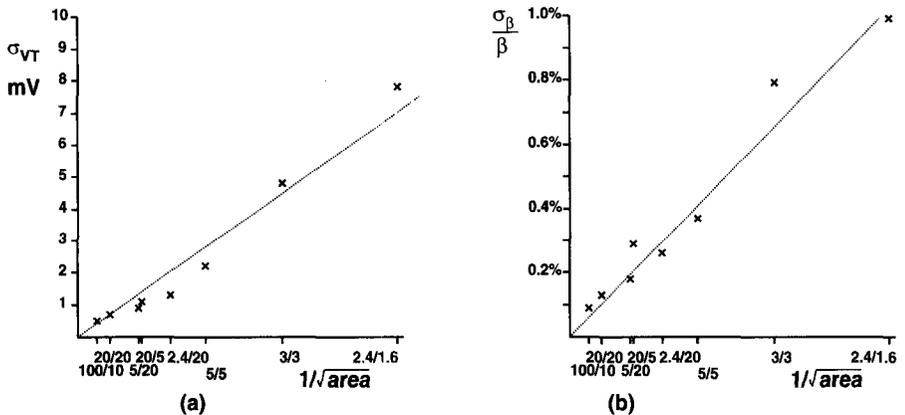


Figure 11.10 The standard deviation of the NMOST threshold (a) and the relative current factor (b) versus the inverse square root of the area, for a 17.5-nm gate oxide process.

TABLE 11.7 Comparison of Measured and Calculated Threshold Mismatch Coefficients

	A_{VT} Measured (mV/ μm)	A_{VT} Calculated (mV/ μm)
0.8- μm data		
NMOST	10.7	10.6
PMOST	18.0	18.6
0.6- μm data		
NMOST	11.0	7.4
PMOST	8.5	8.6

relevant for matching, while the net value ($N_a - N_d$) determines the threshold. In the 0.6 μm PMOST a twin-well construction with a single-well implant was used.

Figure 11.10*b* shows the matching of the current factor. The proportionality factor for the current factor $\beta = \mu C_{\text{ox}} W/L$ is approximated by

$$\frac{\sigma_{\beta}}{\beta} = \frac{A_{\beta}}{\sqrt{WL}} \quad (11.15)$$

The relative matching of the current factor is also proportional to the inverse square root of the area. It is assumed that the matching of the current factor is determined by local variations of the mobility.¹⁸

Low Voltage and Matching

The analysis on the origins of MOS transistor matching in the previous section allows us to analyze the development of matching in various processes. Figure 11.11 shows the development of power supply voltage and matching coefficient over several process generations. A 1/1 transistor was chosen, which corresponds to the numerical value of the factor A_{VT} in the above analysis. During the process development from 1- μm processes to 0.1- μm processes, a slight matching improvement was observed. Equation 11.14 predicts a reduction in the matching factor proportional with the gate oxide thickness, which is confirmed by measurements^{14,24} in Figure 11.11.

The power supply remained fixed at 5 V for many process generations, which led to signal swings of between 1 and 2.5 V. The part of analog CMOS performance that relied on the ratio between signal and component matching (high-speed converters) could improve in previous process generations.

At the level of 0.6- μm CMOS processes, the maximum electrical fields in intrinsic transistors were reached, both the vertical gateoxide field and the lateral field governing the charge transport. For this reason, and to reduce power consumption, efforts started to concentrate on lowering the power supply voltage. On the other hand, the need to tailor the internal fields in the transistor to avoid hot-carrier degradation, has led to more and higher implantation doses. As can be

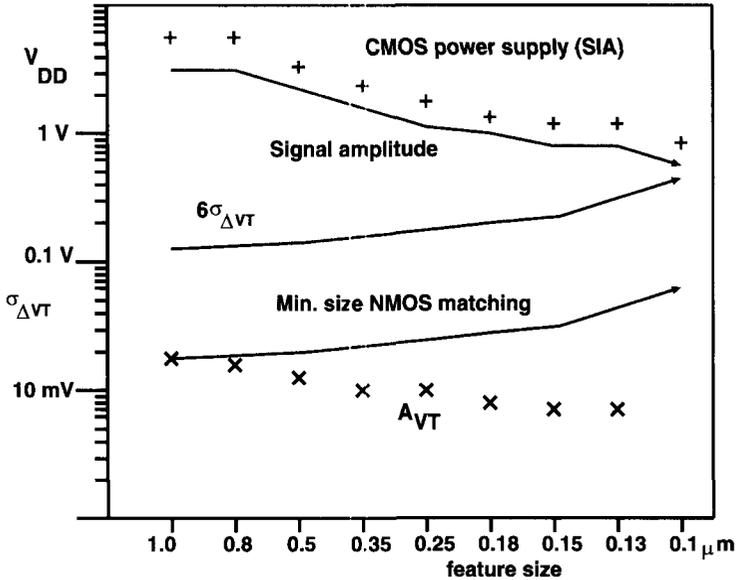


Figure 11.11 Development of power supply voltage and the measured NMOS threshold matching factor A_{VT} through various process generations. The matching coefficient for the two most advanced processes is based on extrapolation. The 6σ curve indicates that even for digital circuits matching can become a limiting factor.

expected from the theoretical background, $A_{VT} \propto \sqrt[4]{N_x}$, the threshold matching factor A_{VT} tends to increase in deep submicrometer processes, becoming especially pronounced when minimum dimensions are used. The shrinking of analog blocks in submicrometer processes is far from trivial.

The expected reduction in the signal-to-matching coefficient ratio in submicrometer CMOS will necessitate changes in the system or technology. In order to perform high-quality data conversion, data converters may shift from digital CMOS to BiCMOS processes. Another possibility would be to use analog options in the technology (see Section 11.3).

Limits of Power and Accuracy

One of the main questions in analog design is the ultimate limit to signal handling. There is no mathematical evidence showing that zero-power mapping of analog values on a discrete amplitude scale would not be possible.

In physics, however, a lower limit can be derived from quantum-mechanical effects, concerning the minimum number of electrons required to represent a bit. Another limit is based on the assumption for oversampling A/D converters that the thermal noise on the input impedance or transconductance is dominant.²⁵⁻²⁷ This approach results in an energy limit based on the product of SNR and thermal kT noise. These limits are, however, four to five decades removed from practical realizations. This is partly due to the fact that much “overhead” power has to be incorporated in real designs. Transconductances in MOS need large amounts of

current, parasitics have to be overcome and safety margins for all kinds of unexpected variations have to be accounted for.

The starting point for a limit based on matching performance is the observation that component-random variations determine relevant specifications in circuits in which the signal passes through different parts of the circuitry. This may happen in multiplexed circuits or in circuits in which the signal path is level-dependent, such as in A/D converters and sense amplifiers of memories. The component variations between the various paths will result in unwanted (spurious) signals such as fractions of sample rates, and fixed pattern noise. These component variations decrease when the area of a MOS transistor is increased. On the other hand, the loading of the signal will also increase when gate areas increase:

$$\begin{array}{cc} \text{Capacitive load} & \text{Threshold variance} \\ C_{\text{gate}} = WLC_{\text{ox}} & \sigma_{VT} = \frac{A_{VT}}{\sqrt{WL}} \end{array} \quad (11.16)$$

The voltage uncertainty on the gate capacitance can be described as an energy term:^{28,29}

$$E_{\sigma VT} = C_{\text{gate}} \sigma_{VT}^2 = C_{\text{ox}} A_{VT}^2 = 4.5 \times 10^{-19} \text{ joules} \quad (11.17)$$

which is independent of the transistor size and corresponds to about 100 kT at room temperature. This energy can be seen as the energy required to toggle a latch pair of transistors in meta-stable condition into a desired state with a 1σ certainty. In circuits with parallel data paths, unwanted signal resulting from component mismatch may hence dominate over more general noise mechanisms as kT noise in the voltage domain.

Local Threshold Variation

In digital design the effects of reduced power supply and increased minimum transistor mismatch become visible. Matching is mostly discussed as a local threshold variation,²⁴ but the same mechanisms as in analog matching are addressed. Basically the threshold variation affects digital circuits through the reduction in safety margin between a “0” and a “1”. When the threshold of a minimum size transistor varies with $\sigma = 30$ mV (see Fig. 11.11), then an additional uncertainty in the threshold voltage of $7\text{--}10\sigma = 200\text{--}300$ mV can occur in a circuit on a multi-million-transistor chip. With a power supply voltage of about 1 V, this would result in the need to enlarge the digital transistors. This sketched energy limit will influence choices for deep-submicron digital design.

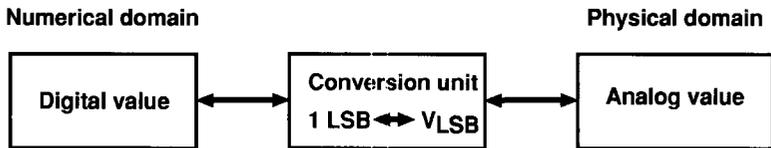
11.5 ANALOG INTERFACES

11.5.1 Analog-to-Digital Conversion: Basic Terminology

Analog-to-digital converters and digital-to-analog converters (A/D and D/A) form the link between the world of physical quantities and the abstract realm of bits and

TABLE 11.8 Key Functions in A/D and D/A Conversion

Analog-to-Digital	Digital-to-Analog
Reference to a conversion unit	Reference from a conversion unit
Amplitude discretization	Amplitude restoration
Time discretization	Holding

**Figure 11.12** A conversion unit is needed for analog-to-digital and digital-to-analog conversion.

numbers. In the process of conversion, a number of key functions can be distinguished³⁰ (see Table 11.8). A/D and D/A conversion both require referencing a conversion unit. The digital value is the ratio of the physical signal value and the conversion units (Fig. 11.12). In many practical systems a reference voltage is used as a conversion unit, which is subdivided by means of resistors, capacitors, or transistors. Current, charge, and time units can also be used. In A/D terms, the reference value, and consequently the maximum input signal, of an N -bit converter is subdivided into 2^N least significant bits (LSBs), where V_{LSB} is the physical value corresponding to one LSB. During operation of an A/D or D/A converter the conversion unit is subdivided, copied, or multiplied, which is subject to various deviations (see Section 11.4). During the conversion from analog to digital, the quantization in both amplitude and time is the dominant signal-disturbing mechanism. The quantization error in an A/D conversion with a small number of quantization levels results in a signal with odd harmonics. When the signal is quantized with a larger ($N > 6$) number of quantization levels, the resulting error is approximated as a uniformly distributed error, with an average value of

$$E_{\text{err}} = \frac{V_{\text{LSB}}^2}{12} \quad (11.18)$$

This “quantization noise” results in a white noise spectrum and leads to a maximum obtainable signal-to-noise ratio (SNR) for a full-scale signal:

$$\text{SNR} = 1.76 + 6.02 * N \quad \text{decibels} \quad (11.19)$$

All converters suffer from more errors than this quantization error alone. In order to

characterize the converter, the effective number of bits (ENOB) is calculated by reversing the above formula and reformulating the SNR:

$$\text{ENOB} = \frac{\text{SINAD} - 1.76}{6.02} \tag{11.20}$$

where SINAD stands for signal-to-noise-and-distortion, the ratio of the signal power to all the unwanted components, such as quantization noise, thermal noise, and distortion.

A well-known component is the total harmonic distortion (THD). The THD is the ratio between the signal and its harmonics. Usually the first 5 or 10 harmonics are counted as THD, while higher-order components and folded products are counted as SINAD contributions. The spurious free dynamic range is the distance in dB between the signal and the largest single unwanted component (see Fig. 11.13).

The dynamic range (DR) is not equivalent to SNR or SINAD as it represents the ratio of the full-scale input signal and the noise floor at a small signal input. The difference between DR and SNR is clearly present in, for example, range-switching configurations.

Two other important specifications are integral and differential nonlinearity (INL, DNL). Two succeeding digital codes should be spaced in the physical domain at a distance of $1 V_{\text{LSB}}$, a deviation from this measure is the differential nonlinearity curve:

$$\text{DNL} = \frac{V_{j+1} - V_j}{V_{\text{LSB}}} - 1 \quad \text{for all } 0 < j < 2^N - 1 \tag{11.21}$$

where V_j is the physical value that corresponds to the digital code j .

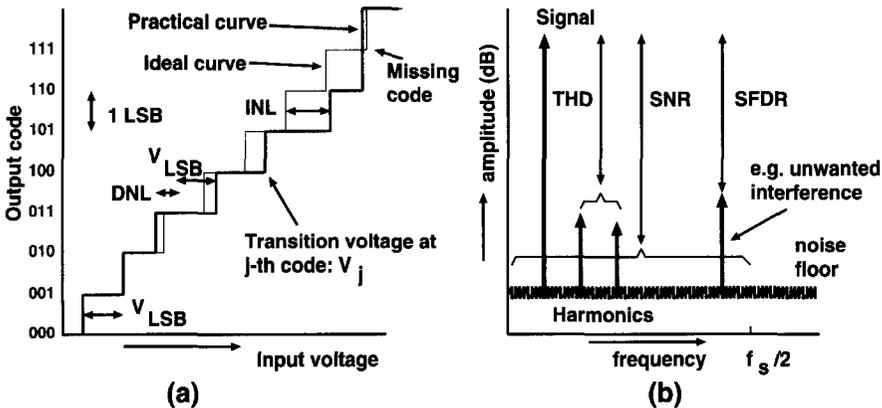


Figure 11.13 Static conversion errors can be seen in an input–output transfer curve (a). Note that the DNL and INL are expressed as a portion of V_{LSB} . The dynamic errors are shown in a frequency plot (b).

Integral non-linearity is defined as the deviation of the real conversion value (measured in LSBs) from the ideal conversion value:

$$\text{INL} = \frac{V_j - V_0}{V_{\text{LSB}}} - j \quad \text{for all } 0 < j < 2^N \quad (11.22)$$

where V_0 is the physical value at minimum code. For A/D and D/A converters, the DNL and INL are specified as a graph for all codes or, more condensed, as a number that represents the maximum value in the entire range. INL and DNL are parameters that are mostly measured at near-dc conditions. They highlight the single strongest deviation of an ideal transfer curve. The ENOB curve as a function of input or sampling frequency represents an average measure, but reflects more accurately the behavior of a converter in dynamic circumstances. In the frequency domain, the bandwidth (BW) that can be ideally converted is limited to half of the sampling rate f_s (Nyquist criterion). Note that the conversion bandwidth usually, but not necessarily, starts at zero Hz (baseband). During reproduction of the signal in the D/A converter, the signal is held at the value of its last sample moment. This operation results in a characteristic signal attenuation and phase delay, which is given by:

$$\frac{\sin(\pi f_{\text{signal}}/f_s)}{\pi f_{\text{signal}}/f_s} e^{-j\pi f_{\text{signal}}/f_s} \quad (11.23)$$

Often this effect is named “ $\sin(x)/x$ ” and the phase term is left out. Note that this attenuation is an average value over all possible phases of the signal to the sampling frequency. In the case of a signal that is locked in some way to the sample frequency (fixed phase), any attenuation can occur between 0 (for sampling at the zero-crossings of a signal frequency at half the sampling rate) and 1 (for sampling the extreme values of the same signal).

Figure 11.14 shows characteristics for three types of A/D converters used in video signal processing [charge-coupled device (CCD) interface for camcorder, baseband

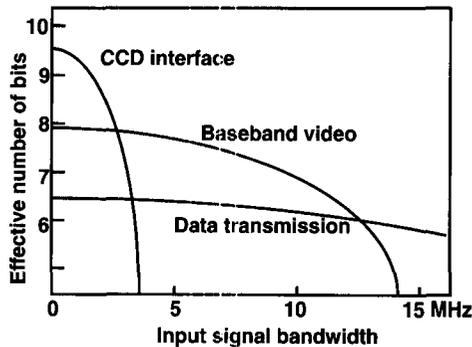


Figure 11.14 Analog-to-digital converters for CCD interfaces, baseband video, and data transmission applications.

video, and data transmission]. The plot shows the relation between the required effective number of bits and the signal frequency. The CCD interface A/D converters require a relatively high degree of accuracy at low frequencies to meet the CCD dynamic range requirement. Because these devices are used in portable cameras, the emphasis is on low-power and low-voltage operation.³¹ Data transmission applications (Teletext and digital satellite transmission) need 6–7 effective bits, but require proper conversion at high signal frequencies.

Between these two extremes is the baseband video converter, used in the signal paths of high-end television sets. Its requirements are as follows: good dc linearity (DNL = 0.5 LSB) to prevent visibility of quantization errors in large areas of slowly varying intensity), 7.5-bit resolution at the color carrier frequency (3.57–4.43 MHz) for color decoding, and a minimum 7-bit performance up to 10–12 MHz signal frequencies for terrestrial and cable digital TV transmission. Although all three converters are used for video signal processing, the specifications are hardly interchangeable.

Several suboptimizations exist for each of these three converters, which have resulted in a wide range of power-performance combinations. Some simple benchmarking can, however, be derived from the combination of bandwidth and accuracy requirements* (see also Fig. 11.20):

$$\frac{\text{Power}}{2^{\text{ENOB}} \times 2 \times \text{BW}} \quad (11.24)$$

where the ENOB number is valid throughout the entire bandwidth. For converters which keep their performance up to the Nyquist criterion, BW can be substituted by half of the sampling rate f_s .

This figure of merit still depends on technology, architecture, and other specifications. Figure 11.15 compares some recent AD converters. The differences are attributable to architecture (number of comparators), technology, noise, or limited matching (see, e.g., Section 11.4).

11.5.2 Conversion Architectures

D/A Converters

In a digital-to-analog converter the digital number is transformed into a physical quantity by means of a conversion unit. A conversion voltage or current is subdivided by passive or active elements, such as resistors, CMOS transistors, or capacitors, or it is subdivided in time. A few comments on each of these elements are

- *Resistor.* The relative accuracy is in the order of 10^{-3} – 10^{-4} , the absolute value will suffer from (large) process variations and temperature. The resistor value is determined by the successive loading. Capacitive coupling

* Other benchmarks use dc accuracy and sampling rate, ignoring the signal performance.²⁵

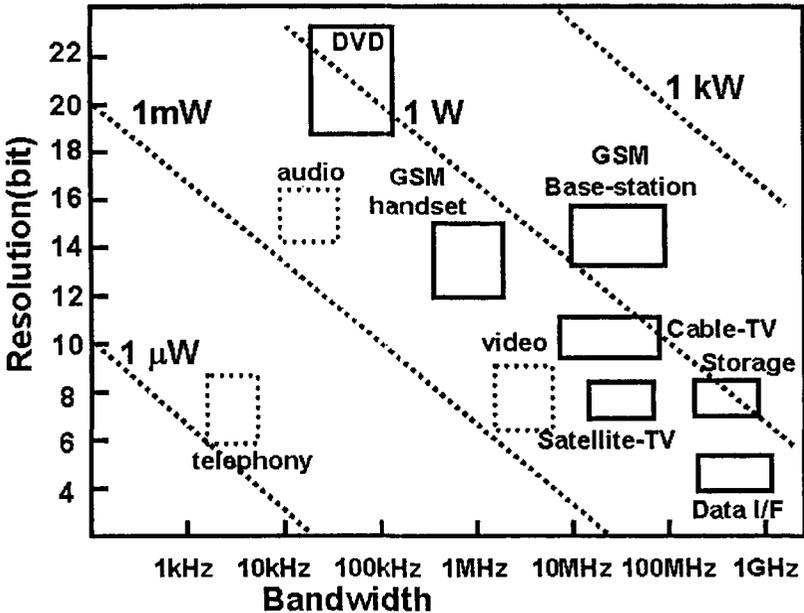


Figure 11.15 Power efficiency of (mostly) ISSCC-published A/D converters. An improvement over the years of conversion efficiency is clearly observed.

to the substrate may lead to noise pickup. Constant current is always needed.

- *Capacitor.* The relative accuracy is in the order of 10^{-3} – 10^{-4} , the absolute value is usually well-defined in a double-poly process. An opamp configuration is needed to manipulate charges. Minimum size is determined by parasitics or a kT/C noise floor. Often seen as low-power solution, but requires large peak currents during charge transfers. Sensitive to different parasitic couplings.³²
- *Transistor.* Relative accuracy is in the order of 10^{-3} ; the absolute value is sensitive to temperature and process spread. Mostly used as current source or current divider. Back-gate modulation and $1/f$ noise must be considered.
- *Time.* With more or less fixed variation (30–100 ps_{rms}), the best accuracy is at low signal bandwidths; converters based on time accuracy convert bandwidths that are several orders of magnitude lower than the sampling frequency. Interference enters via clock buffers.

Two representations are used for the conversion itself: unary and binary. In a unary format 2^N copies of the conversion unit of an LSB are present and the conversion is performed by selecting the proper number of units. Examples are resistor strings and

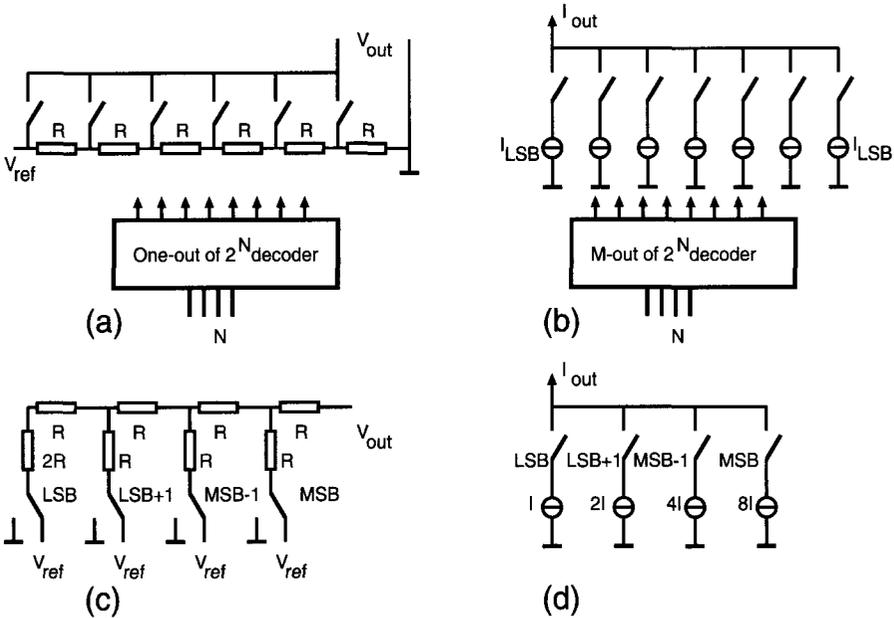


Figure 11.16 Four basic forms of D/A conversion: unary voltage (a) and current (b) D/A conversion and binary voltage (c) and current (d) conversion.

parallel current sources (see Fig. 11.16a,b). There are similar techniques with capacitor arrays and timing (counting D/A).

The alternative is the direct use of the binary information. Instead of 2^N copies, there are N physical values that correspond to the binary powers of the digital signal. Straightforward use of those binary numbers, in case of a direct binary representation, results in the required conversion. Figure 11.16c,d shows voltage and current schemes for binary D/A conversion. All the schemes involve the problem of accurately reproducing the unit value. Several improvements have been proposed.³³⁻³⁵ Often the binary representation offers the lowest area use. Most of these schemes do not directly affect the converter's overall power consumption. In signal quality the difference occurs at transitions at which many bits flip (01111→10000). While the unary organized D/A adds another unit, the binary converter changes from one group of elements to a completely different one. Imperfections in the different groups of elements directly result in errors at those code transitions in the form of differential nonlinearity. Most D/A schemes can be classified along these lines, though there are a few deviating forms such as ternary coding (+ 1,0, - 1), which is sometimes used in combination with sign/magnitude representation.

In the case of converters with a high resolution, the problem with these schemes is the large number of units involved or the wide range of binary values. Subranging is generally applied to circumvent these problems. A converter of N -bit resolution is subdivided into a cascade of two subconverters of M and $N - M$ bits. Subranging can

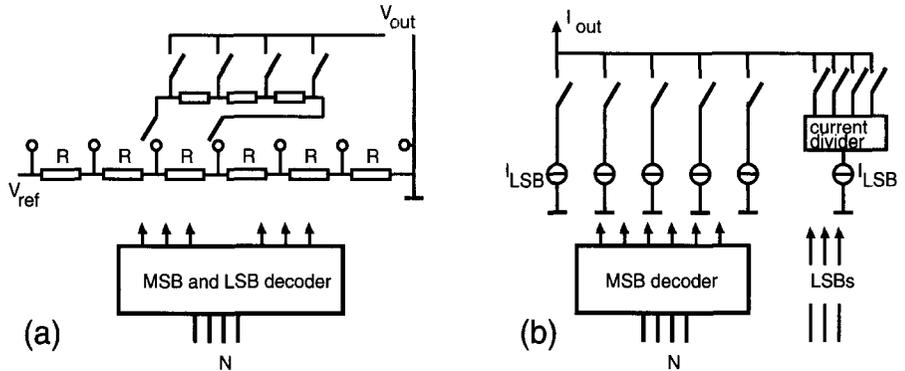


Figure 11.17 Basic circuits for voltage (a) and current (b) subranging D/A converters.

be extended to an N -bit converter that is split into N subconverters of 1 bit. Figure 11.17 shows some subranging schemes. The reduction in area is compensated with additional control or buffering circuits and for $N > 8$ (CMOS) the deviations become so large that some form of overlap on the edges of the subranges must be provided.

Oversampling

In Nyquist A/D and D/A converters, where the sampling rate is slightly higher than twice the bandwidth of interest, there is a direct match between the required accuracy and the number of levels in the converter. If the sampling rate is much higher than the Nyquist criterion requires, an exchange between time and accuracy is possible using “oversampled converters.” In an examination of the fundamental properties of A/D and D/A conversion, Section 11.5.1 discussed quantization errors. An exchange can then be made between SNR and the sample rate. If, for a given bandwidth, the sample rate of an N -bit converter is increased by a factor of 4, the noise energy will spread over a frequency band that is 4 times as large and the amplitude of the noise in that bandwidth will be reduced by a factor of 2. An appropriate decimation filter limiting the bandwidth to the desired frequency range, increases the effective resolution by 1 bit. The general relation between an increase in resolution and an increase in sample rate is

$$\Delta N = \frac{1}{2} \log_2 \frac{f_s}{2BW} \quad (11.25)$$

This solution allows to use converters with less resolution, but accuracy (e.g., comparator transition voltages) will still be required at the final resolution depth. Another disadvantage lies in the white-noise assumption for the quantization. In the case of low signal amplitudes or nearly dc, quantization becomes harmonic distortion and oversampling will only work in the presence of dither signals or noise fed back from the converter (delta modulation).

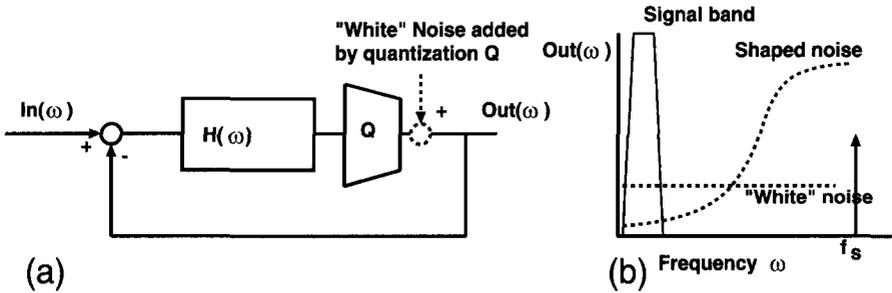


Figure 11.18 Sigma delta ($\Sigma\Delta$) modulation schematic (a) and resulting noise spectrum (b).

A second step to improve accuracy at the cost of sampling rate is to apply feedback techniques. These methods allow shifting the quantization energy out of desired signal frequency ranges. Figure 11.18 shows the basic principle of $\Sigma\Delta$ modulation.^{36–38} The feedback loop, with an integration (Σ) and a comparator function (Δ), shapes the noise energy (symbolized by “white” noise entering into the dashed summing point) from a flat spectrum into a shape that is the complement of the loop filter $H(\omega)$. When the quantizer’s effective gain is modeled as A_q , the input–output relation becomes

$$Out(\omega) = \frac{A_q H(\omega)}{1 + A_q H(\omega)} In(\omega) + \frac{1}{1 + A_q H(\omega)} Noise \quad (11.26)$$

At high values of $A_q H(\omega)$ the quantization noise is suppressed and the output signal equals the input signal. This type of noise shaping can be used in analog-to-digital, digital-to-analog, and digital-to-digital conversions of signals. A prerequisite is a sufficiently high oversample factor f_s over the signal bandwidth. This kind of conversion allows the use of very small quantizers (e.g., a few switches for a D/A) and is therefore very power and area efficient.³⁸

A/D Converters

Every A/D converter consists of an inherent D/A function in combination with a comparison. The number of comparators in an architecture is a dominant parameter for the power budget. Control and references are usually of minor importance. The mutual equality of comparators is very important because their input-referred errors add up to the errors of the attached D/A converter. The following section, “Effect of Matching on Comparators,” discusses fundamental problems of comparator mismatch.

An A/D comparator needs a minimum overdrive voltage $V_{overdrive}$ to achieve a stable output level within the allowed decision time T_d . It is, however, possible to define an input difference voltage range ΔV in which the comparator cannot decide in time. The bit error rate (BER) is a measure of that range and is closely related to the fundamental decision problem in metastable elements.

$$BER = \text{probability}(|V_{overdrive}| < \Delta V) \approx \frac{2\Delta V}{V_{LSB}} \quad (11.27)$$

The minimum overdrive ΔV can easily be derived from an exponential growth of the comparator decision levels toward the power supply V_{DD} :

$$\text{BER} = \frac{V_{DD} e^{-T_d/\tau}}{V_{LSB}} \approx 2^N e^{-T_d/\tau} \quad (11.28)$$

if an input signal swing equal to the power supply voltage is assumed. In CMOS τ is the time constant formed by the parasitic and gate capacitances and the achievable transconductance. A typical example with 8 bits is 5 fF total capacitance for 1- μm gate width, 5 $\mu\text{A/V}$ transconductance, and $T_d = 20$ ns, which results in a BER of 10^{-7} . This BER can be improved to better than 10^{-10} by more current in the latch transistors. In addition to the mere improvement of the latch speed, we can take measures in the decoding scheme to avoid serious code errors due to a metastable state.

A full-flash converter has one comparator for every possible reference value, so $2^N - 1$ comparators are needed. Full-flash converters are very power-hungry, but have the advantage of single clock-delay conversion. A variant of a full-flash converter is the folding A/D converter^{39,40} in which preprocessing stage “folds” the input signal, which reduces the number of comparators. Proper design of the preprocessing, combining high speed and high yield, is the critical issue. More recent techniques use interpolation for further optimization.^{40,41} An important question with respect to these converters is whether to use a sample-and-hold (S/H) circuit in the converter’s input stage. Experience shows that a high-speed S/H of full-signal and bandwidth performance requires 10–30% of the total A/D power budget. The advantage of using a S/H is that signal propagation errors in the analog preprocessing are reduced and that architectures can be used that make multiple use of the input signal.

As in the case of D/A converters, much use is made of subranging for A/D conversion (see Fig. 11.19). In 2 to N stages, smaller flash converters convert part of the signal and the remainder is amplified and passed to a next section. The amplification (A in Fig. 11.19b) is crucial as it reduces the effect of errors in the succeeding stages. Subranging requires more time for the signal to propagate. Additional hardware is required for proper high-speed operation; overranging circuits allow signals, that have been misinterpreted in prior sections, to be converted. Subranging converters require multiple access to the original or intermediate signals; these converters consequently are equipped with S/H circuits between each section. Nevertheless, the smaller number of comparators leads to efficient converters.

In an extreme case, N sections convert one bit in each section. This architecture is referred to as “pipelined.”^{43,44} Although a single comparator is used per section, equality problems between sections and between the comparator and the subtract circuit have to be solved.* In pipelined converters the signal latency can be as high as

* Of course, many variants exist: with different D/A structures, combination of techniques, using differential or residue signals, and so on. The reader is referred to the appropriate textbooks.³⁹

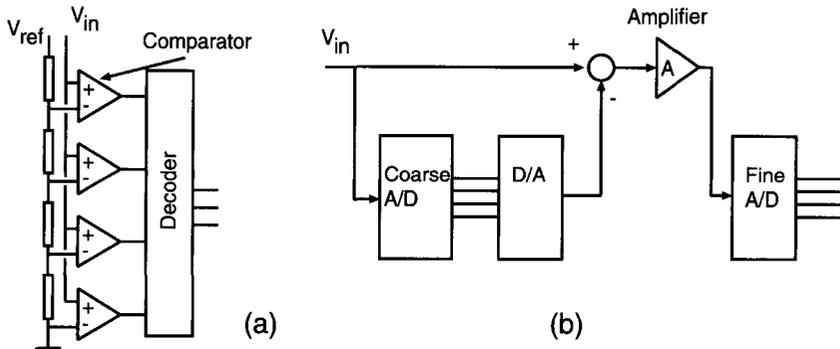


Figure 11.19 Full-flash converter (a) used in the fine and coarse A/Ds of the subranging converter (b).

20 clock cycles. This delay must be taken into account if the converter is part of a feedback loop (e.g., gain control). If more than 7-bit-accuracy ADCs in CMOS are needed, all these techniques require a form of random offset cancellation. A number of basic techniques can be used:

- “One-time” trimming is used in converters for professional applications. In a special layer on top of the IC, resistors are formed that can be trimmed by laser cutting. Different sources of mismatch can be removed through proper application. Once it has been trimmed, the A/D converter can run without timing restrictions. A disadvantage is that testing is more expensive: the converter is characterized in a first run, then trimming applied, and the converter is remeasured. There are other forms of one-time offset cancellation besides resistor trimming, for example, by means of fuses or PROM.
- The signal can be amplified with respect to the random offset. This can be done only through some form of amplitude adaptation. One form is the aforementioned “folding” technique. Another form is, in fact, subranging. In both techniques a well-known part of the reference voltage is subtracted from the input signal. This subtraction has to be done accurately; special attention is needed at range transitions.
- The offset can be measured and stored. A popular technique uses a combination of capacitors and switches at the input of the comparator. The disadvantage of these types of offset cancellation is the time required to measure and store the offset. A more fundamental problem is that the offset is never fully canceled. In the first place there is a gain-bandwidth limitation of the cancellation loop and, second, the circuit in which the offset is canceled differs in some way from the circuit that is used to convert the signal; for instance, MOSTs are switched ON or OFF. The offset-storage capacitor must be large to ensure a low mismatch voltage, but must at the same time be small to reduce settling and power problems. The switched load, related to the input of the converter, often causes excessive power in the driving circuit.

Single-comparator architectures do not suffer from comparator matching problems. Successive approximation converters do a binary search through the reference range and require as many clock pulses as output bits.

Another form of single comparator A/D conversion involves the use of a single comparator in a $\Sigma\Delta$ approach (see Fig. 11.18). The oversample ratio and the filter characteristics determine the bandwidth and resolution. Figure 11.15 shows how single comparator designs (the 13- and 16-bit designs) exploit this advantage by using less power per conversion.

Effect of Matching on Comparators

In A/D converters mismatch manifests itself as noise but is also observable in other characteristics. The most critical is differential nonlinearity because this effect relates to the maximum error in the parallel structure, while noise is only a power average value over all the error terms.

The charge standard deviation, Eq. (11.12), is linked to the A/D converter's DNL performance by deriving it from a minimum LSB size for the input signal. In a parallel structure many comparators will be involved in the decision process. A safety margin must be employed so that an LSB change in the input signal is always detected. This safety margin is defined via a $N(0,1)$ normal distribution as α . A sufficiently low error probability is obtained at $\alpha = 7$ to 10. Therefore, the resulting minimum LSB size for safe operation of a comparator is

$$Q_{\text{LSB}} = \alpha \times \sigma_{Q,d} = \alpha q \sqrt{WLx_d N_a} \quad (11.29)$$

The minimum charge for an LSB combined with the signal speed results in a current. It is assumed that no slewing of the comparator stage is allowed and the current is delivered in class A operation. In converter terminology: the DNL performance must be reached at a signal frequency f_{signal} . The minimum amount of time required for an input signal to change V_{LSB} is t_{LSB} :

$$t_{\text{LSB}} = \frac{V_{\text{LSB}}}{\delta V_{\text{signal}} / \delta t} = \frac{1}{2^N \pi f_{\text{signal}}} \quad (11.30)$$

In the context of this analysis we assume a time-continuous signal, a similar analysis holds for sample-and-hold signals. The current needed for supplying Q_{LSB} at the steepest point of an input signal at a frequency f_{signal} is

$$i_{\text{max}} = \frac{Q_{\text{LSB}}}{t_{\text{LSB}}} = 2^N \pi f_{\text{signal}} \alpha q \sqrt{WLx_d N_a} \quad (11.31)$$

When this current is calculated for an 8-bit resolution converter, and a 10 MHz signal frequency ($N = 8$, $f_{\text{signal}} = 10^7$ Hz, $\alpha = 10$, $N_a = 10^{16} \text{ cm}^{-3}$, $W = L = x_d = 1 \mu\text{m}$), the result is $i_{\text{max}} = 1.3 \mu\text{A}$ per transistor. This formula can be linked to the power per resolution and bandwidth of Figure 11.15 by multiplying with V_{DD} and

rearrangement of terms. The proportionality constant is then

$$\frac{\text{Power}}{2^N 2BW} = \frac{\pi \alpha q}{2} \sqrt{WLx_d N_a} \quad (11.32)$$

which amounts to about 10^{-3} pJ/conversion per relevant transistor. On the assumption of 4–6 matching critical transistors per comparator and 60–100 comparators per high-speed A/D converter, this (very rough) approximation results in 0.5 pJ/conv, which is about one or two orders removed from the realized converters.

Figure 11.20 shows a graphical representation of the current consumptions of some comparators in recent A/D converters. The line represents a comparator with a sampling rate of 30 Ms/s, with four critical transistors. Closer examination of this discrepancy shows the following contributing factors:

- It was assumed that only the intrinsic transistor was charged. In the design the parasitic capacitors require roughly the same amounts of charge.
- More effects than the depletion charge alone contribute to the uncertainty of an LSB such as W , L dependencies, and mobility variations.
- A/D converters are designed to operate under worst-case circumstances; temperature, power supply and process variations require an overdesign of about 2–3 times.
- In the analysis, minimum power consumption was realized by using minimum transistors. The consequence is that the LSB size is rather large,

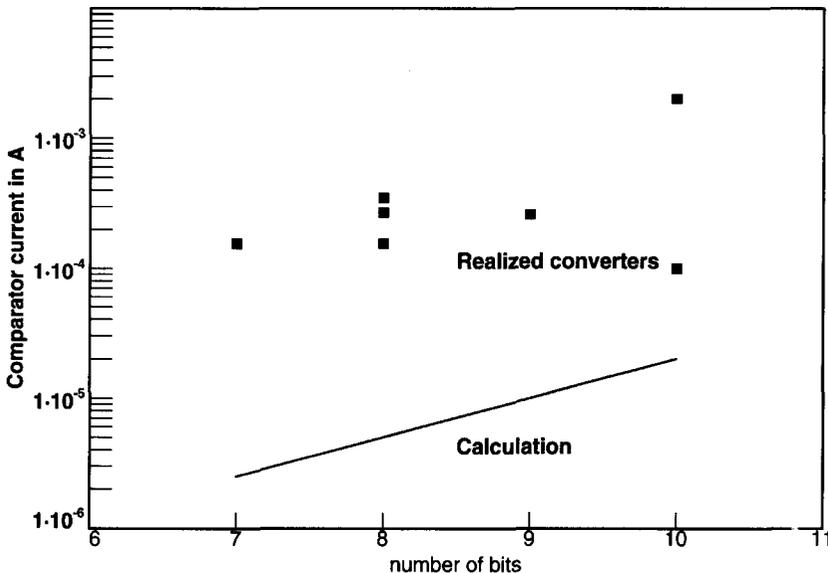


Figure 11.20 A comparison of power consumptions with the derived approximations.

$Q_{\text{LSB}}/WLC_{\text{ox}}$, and the corresponding input signal has an amplitude that probably exceeds practical power supply voltages. Because of this, additional current paths are required in a comparator.

Basically, offset compensation enables reductions in component variations.^{31,42,45,46} In practice this technique does not cancel the mismatch in the final decision element (the latch), but it reduces the random offset in the preamplifiers. In other words, energy must be supplied during mismatch cancellation. Another point is the additional random mismatch that is introduced into the signal capacitors by the switches. The ratio of the switch gate-capacitance and the signal-storage capacitors is limited because of the fast settling requirement, which limits the mismatch reduction capabilities. Real improvement may come from better intrinsic components, for example, by means of trimming of sensitive stages or new silicon active elements.

In our analysis we found that the basic problem of getting accuracy at high speeds was located in the comparators. This implies that N comparators for an N -bit digital word is the minimum configuration for low-power A/D converters. Most designs employ 10 times more comparators. In the architecture the number of contributing transistors must be minimized. Potentially this can be realized by multiplexing and pipelining. Another approach is to avoid the problems that parallel circuits introduce by going to single comparator oversampling. Then the tradeoff with digital filtering becomes important again.

11.6 INTEGRATION OF BLOCKS

The tradeoff on the system level demands the optimization of the entire signal chain, not only optimizing the different sub blocks. An important aspect of this optimization is flexibility. Unlike specification tolerances, flexibility demands the adaptation of the system to predetermined (large) system-parameter shifts. Some of these shifts are necessitated by the wish to serve several product lines of a manufacturer, such as compatibility with several power supply voltages or various output formats. In other cases, the flexibility is implemented in the system itself, because the system is multistandard, comprising several transmission standards, interfaces with different sources, and so on. Flexibility mostly implies more hardware and more power than an optimized solution. Sometimes the increase in power can be reduced by designing a part in such a way that losses are minimized; for example by bonding or external setting such as control of current. A proper balance has to be found in designing flexibility in a system with minimum power or component overheads.

In the digital domain, flexibility can be implemented by means of shifting toward more programmable building blocks. As discussed in Section 11.2, ASIC implementations can be replaced by DSP or full CPU-type implementation. The loss of efficiency has to be balanced with the required flexibility.

11.6.1 Signal Processing Strategy

In all consumer-oriented systems signal processing is the dominant function. The main parameters of analog signal processing on a system level are dynamic range, signal-to-noise ratio, and bandwidth. These quantities translate into resolution and sampling-rate requirements in the digital domain.

Sampling Rate and Bandwidth

On the system level the sampling rate of an A/D converter is locked to or derived from the system clock. The choice of clock (and sampling) frequency used on the system level is important with respect to specification and power. The main criterion for the sampling frequency is given by the Nyquist theorem ($f_s > 2BW$). For high bandwidth circuits, the sampling frequency is, in practice, 20–50% higher than the minimum rate required by the Nyquist theorem. The main reasons for this are in the tradeoff between analog signal properties and the consequences for digital signal processing where a higher clock rate has disadvantages (see Table 11.9). Power and area in the digital domain are balanced by signal quality and filter complexity in the analog domain.

In the case of low-signal bandwidths, the ratio of the sample rate and the bandwidth is chosen so that it optimizes the entire conversion chain. Figure 11.21 shows how a tradeoff can be found between the alias filter order and the sample rate/bandwidth ratio. A higher ratio of the sample rate and signal bandwidth allows a lower-order alias filter.

Another example of a very specific sampling frequency will be given in the paragraph “Sampling of Modulated Signals.”

Sample Rate and Accuracy

In the processing strategy it is very important to choose the appropriate conversion position in order to minimize the power required. There will usually be some freedom in choosing the position of the data converter within the signal chain. As A/D* conversion is quantized in the amplitude and in time domains, tradeoffs can be made between these two domains. In the amplitude domain the optimum SNR and the dynamic range have to be determined. In cases in which a lower SNR can be used

TABLE 11.9 Balancing Digital Drawbacks and Analog Benefits of a High Sampling Rate

Digital Drawbacks	Analog Advantages
More dynamic power consumption	Improved SNR due to oversampling
Longer digital filter structures	Alias filtering is simpler
Digital timing is more critical	Less steep voltage steps (slewing)
Larger storage units for fixed time delays	Less signal loss due to $\sin(x)/x$

* The arguments presented here apply to A/D conversion, but apply mostly for D/A conversion, too.

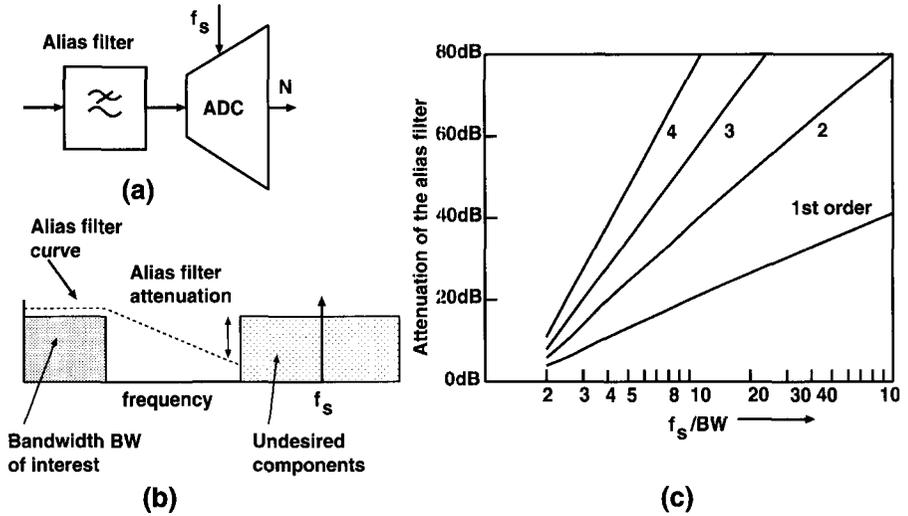


Figure 11.21 An alias filter before an ADC (a) will suppress undesired components (b). For baseband conversion the attenuation is a function of filter order and the ratio between bandwidth and sampling rate.

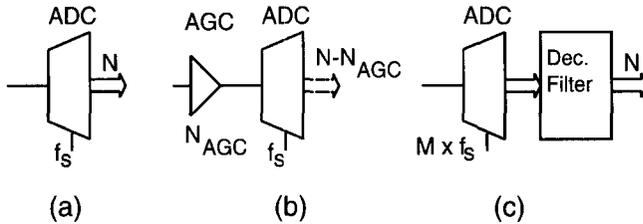


Figure 11.22 Standard A/D converter (a), A/D converter with gain control (b) and A/D with oversampling and decimation filter (c).

with respect to the dynamic range, the preferred system solution is a gain-controlled amplifier followed by a minimum SNR A/D converter, see Figure 11.22. Gain control requires some form of signal analysis, which is usually an inexpensive function in the digital domain. Another option in this case is a companding A/D converter, that employs different LSB step sizes in the signal range.

A certain degree of freedom will also exist if the sample rate of the A/D converter is much larger than the required bandwidth. An exchange can then be made between SNR and sample rate, as indicated in Eq. 11.25. In specific system architectures the advantages and disadvantages of these solutions have to be investigated. Generally a reduction in A/D resolution outweighs the costs of gain control or of a digital filter.

Sampling of Modulated Signals

Because many systems are used in some form of communication equipment, incoming signals may be modulated on a carrier frequency. These input signals

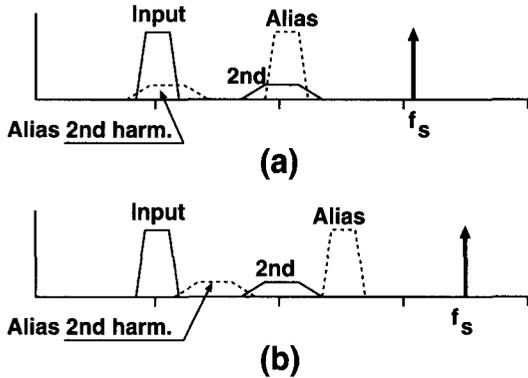


Figure 11.23 Spectrum of a sampled modulation signal with distortion. Sample rate equals 3.1 times the modulation frequency (a), sample rate equals 3.5 times the modulation frequency (b).

containing carrier-modulated components imply additional difficulties. Aliases of the harmonic distortion of the signals will fold back during the sampling process in the converter. If the sampling rate is close to an integer multiple of that carrier frequency, the harmonics will interfere with the original signal (see Fig. 11.23). This occurs in many transmission systems such as in PAL video in which the color modulation frequency is 4.433 MHz while the preferred sampling rate is 13.5 Ms/s. A sample rate close to an integer of the modulation frequency may necessitate additional suppression of the amplitude of the distortion component. Suppressing the distortion component can be achieved at the expense of more power in the converter.

11.6.2 Implementing System Functions

Proper signal analysis allows the functional use of the conversion properties. Examples are:

- The inherent sampling performs demodulation, which can be implemented in several A/D architectures. Particularly suitable are structures with high-performance sample-and-hold circuits.
- The inverse process is equally applicable. Use the upper bands generated by the D/A function for direct digital synthesis (DDS) of radiofrequency signals.
- Multiple input signals can be multiplexed on the S/H input of an A/D converter.
- Alias filtering can be combined with system-required filtering.
- A local sample rate increase can be used to relax the requirements in other parts of the system.

Conversion of Modulated Signals

Figure 11.24 shows an example of the use of the S/H function for realizing the down-mixing of modulated signals. The information content of the input signal is rather band-limited, but it is modulated on a relatively high carrier frequency. In this case it is power efficient to implement the conversion starting with a sample-and-hold circuit. The sampling function is used to modulate the signal band to a much lower frequency. In this example, downmodulation is performed at around twice the clock frequency, making the conversion task for the A/D converter core simpler.

Oversampled D/A Converter

As an example of the analog digital tradeoff, we will analyze an output stage for video-like specifications. Figure 11.25 shows a practical example of local oversampling. In Figure 11.25*a,b* a standard D/A conversion is succeeded by an off-chip filter. Due to the rather low ratio of sample rate and high signal frequency, large transient steps will occur at the on-chip buffer and the output pin. This will usually cause slewing and distortion. The driver has to be designed with additional bandwidth, and the input stages will need large bias currents. This setup, moreover, has a relatively poor HF performance due to $\sin(x)/x$ signal loss. The passive filter requires some three to seven poles and is expensive to produce, especially if $\sin(x)/x$ compensation is also needed.

Figure 11.25*c,d* shows an integrated circuit solution using oversampling. The sample rate is doubled locally and the odd alias terms in the frequency spectrum are removed by digital filtering. In this circuit large transients in the output are more than halved in amplitude, and relatively simple noncritical postfiltering (first-order) is sufficient to restore the analog signal. The inherent $\sin(x)/x$ is reduced by an order of magnitude, so no compensation is required. Figure 11.26 shows a chip photograph of a digital oversample filter succeeded by a D/A converter.

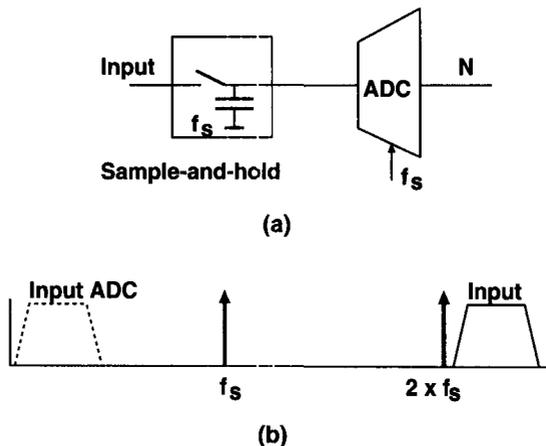


Figure 11.24 Converter arrangement for $1/2$ conversion (a); the sample-and-hold function performs downmodulation of the input signal (b).

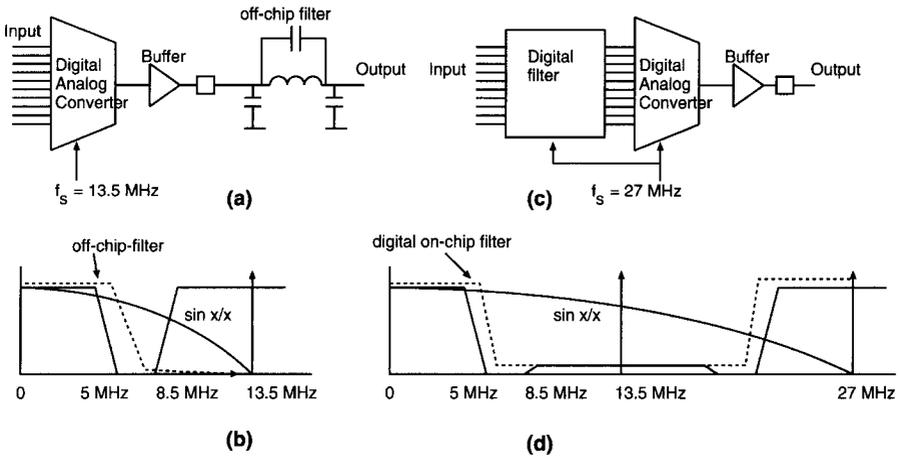


Figure 11.25 D/A converter with output filtered by external filter (a,b) and locally oversampled D/A driven by digital prefilter (c,d).

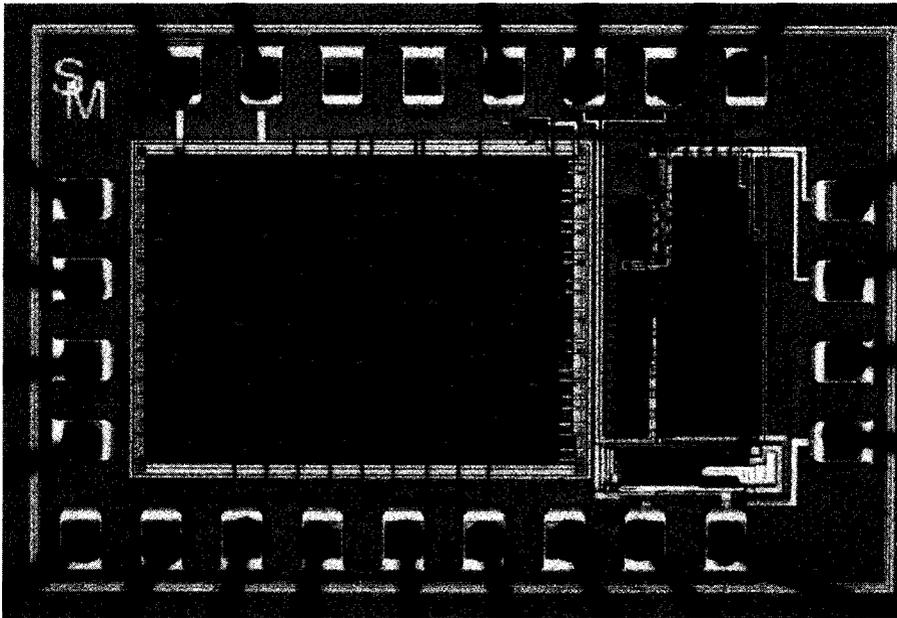


Figure 11.26 Chip photograph of an 8-bit D/A (right) and digital oversample simple, designed by S. Menten and M. Pelgrom.

From a power point of view a tradeoff must be made between the additional power and area for the filter and the quality loss and additional power in the buffer (see Table 11.10). In this section we discussed several system choices that influence the power budget of analog-to-digital conversion. Thorough signal and system analysis allows proper choices in the setup of a system chip.

TABLE 11.10 Comparison of the Two D/A Circuits in Figure 11.25

Direct D/A Conversion	Oversampled D/A
Off-chip filter needed	$\sim 1 \text{ mm}^2$ CMOS
10-mA current in driver	5 mA current in driver
—	Power for digital filter
$\sin(x)/x$ loss = 4 dB	$\sin(x)/x$ loss = 0.5 dB
	$2 \times$ clock needed

11.6.3 Interference

The many different blocks on a system IC often have more physical coupling paths than desired. Capacitive, resistive, and inductive coupling between blocks may cause interference in other parts of the chip. The effects can be performance degradation or incidental loss of functionality if digital signals are disturbed. Interference can be analyzed by looking at the interferer, the transmission medium or coupling mechanism, and the receiving (disturbed) circuit.

Figure 11.27 shows the currents in the two transition states of a basic inverter. The switching of states in digital circuitry (here represented by an inverter) causes the charge and discharge of the capacitors connected to the output nodes of the inverter. The currents involved are supplied or sunk into the power supply terminals. The supply lines show resistance and inductance via the coils in the bondwires, so the current spikes translate to voltage variations in the power lines. If a charge of $C_{\text{load}}V_{dd}$ is moved to or from the inverter in a time period ΔT , then the associated current spike is given by

$$\Delta I_{dd} = \frac{C_{\text{load}}V_{dd}}{\Delta T} \quad (11.33)$$

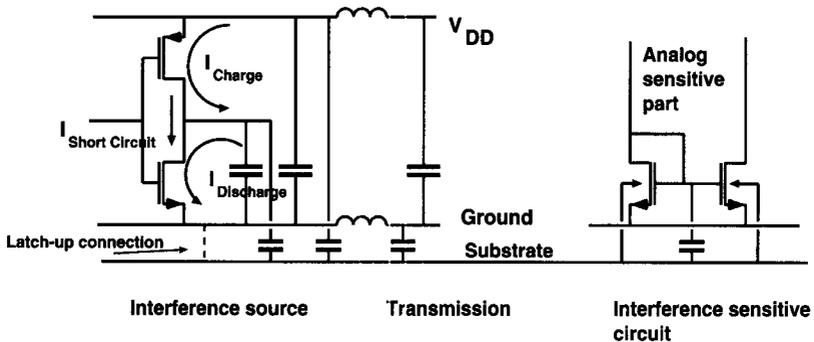


Figure 11.27 Interference schematic. Interference is generated by the currents in a digital inverter. The substrate is the medium that couples the interference into the current mirror of an analog circuit.

A simple first-order clock buffer driving 10 pF in 0.5 ns will cause current spikes of 60 mA. These current spikes, in turn, will lead to strong power supply variations because the voltage drop over the bondwire inductance is given by

$$\Delta V_{dd} = L_{\text{bond}} \frac{\delta I_{dd}}{\delta T} \quad (11.34)$$

In this example the voltage dip would amount roughly 0.5 V (with $L_{\text{bond}} = 5$ nH). By itself, this voltage dip will lead to reduced performance of the digital blocks; however, there are several mechanism that make the energy associated with these transitions spreads out to other circuits. Examples are

- Digital noise couples into the substrate and the digital and analog circuit both share a common substrate. The digital circuits are capacitively coupled to the substrate; or even worse, the digital ground line is sometimes directly connected to the substrate to reduce latchup problems. Here it is important to distinguish between low-ohmic substrates (used in most CMOS technologies) and high-ohmic substrates (in BiCMOS and SOI). In high-ohmic substrates the interference-sensitive circuits can be protected by guard rings that provide a locally clean substrate because an effective voltage drop of the interfering signal over the substrate resistance can be achieved. However, with low-ohmic substrates such a connection will cause a low-ohmic path from digital circuits into the sensitive regions.
- Connections between various power supply paths, directly, via bonding or PCB. If ground lines, or power supply lines, of heavy switching circuits and sensitive circuits are coupled, the latter circuits will suffer from interference. This effect is often recognized in analog versus digital circuits, however, at low power, voltage and interference from one digital cell (e.g., output buffers) can couple to another digital cell (e.g., RAM) and can cause a malfunction.
- A less trivial coupling is through timing signals. Any signal transition that passes a gate suffers a delay. This delay is a strong function of the actual power supply voltage, so the signal transition will be shifted in time depending on the supply variations. In the analog domain, this will result in inaccurate sampling and signal quality loss; in the digital domain, skew problems can lead to timing problems.

When a team designs a system chip, the issue of interference should be a day 1 priority, because it affects all parts of the design. Several measures can be taken to reduce interference problems:

- To reduce the interference generation, current spikes can be reduced by local capacitances and well-to-substrate capacitors. Danger of ringing exists if the loop formed by the capacitor and the wiring inductance is not sufficiently damped. Of course, avoid every form of overdimensioning of, for example, clock and output buffers. In addition, slew-rate-controlled buffers can help to

reduce spikes. Avoid direct connections of the digital cells to the substrate. Separate substrate wiring will cost a few percent of space, but is better for reduced interference. On the PCB level take care to reduce reflections that force protection supplies to sink excess charge.

- In the transmission path of interferer – interference pickup, avoid joint power wiring. Consider proper placement of bondpaths. Digital inputs and outputs should be separated by ground-connected bondpaths from analog inputs or outputs.
- At the interference-sensitive part a differential designed circuit may reduce the impact of interference. A major problem is the input path, which, for economical or compatibility reasons, is mostly single-sided. Referencing the input circuit to a supply voltage that is equally interfered is a workable solution.
- In mixed analog/digital CMOS circuits it is certainly not advisable to connect the analog power supply everywhere to the substrate. Part of the digital current will flow through the analog supply wiring. Local optimization may lead to exceptions, for example, to reduce interference picked up in current mirrors.
- Many analog/digital system chips are referenced to a single clock edge, analog processing is best carried out on the nonactive clock edge.
- Do not route analog timing signals via digital blocks. Derive analog clock signals directly from the source and buffer incoming digital signals in buffers and latches connected to clean supplies.

11.6.4 Partitioning

Digital/Memory Choices

Many system-on-chip concepts combine digital hardware, memories, and analog interface blocks. Despite a large number of similarities, there are major differences between the process optimizations for each of these categories. Optimized DRAM cells require trench or stacked capacitor structures, efficient (EEP)ROM memories need special technology options, and analog circuits require some modifications to the standard digital process.

In the case of memory structures the decision to include the memory on the CMOS chip or to use an external memory with higher packing density depends on the size of the memory and the access that is required. For memory sizes lower than a few megabytes, the standard CMOS packing density is sufficient for economic use. Larger memories require the use of specialized processes.

Figure 11.28 compares the density of DRAMs in standard CMOS generations and in specialized processes. The memory density in stand-alone specialized processes exceeds the standard CMOS density by about an order of magnitude. This observation, resulting in a choice for external memories, causes a dilemma considering that large DSPs or CPUs require many gigabits of data transfer to and from the memory. This leads to various alternative scenarios such as

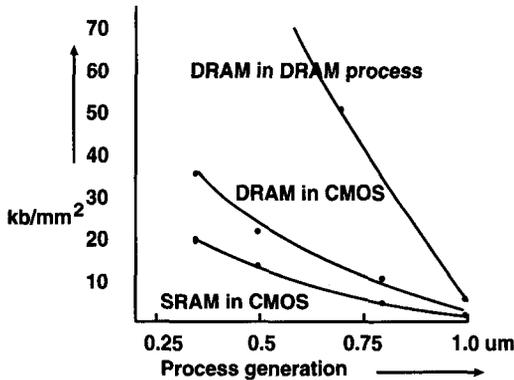


Figure 11.28 Memory density in standard CMOS and in DRAM processes. The data are estimated from ISSCC88-97.

- Extend the standard process with an efficient DRAM option. However, this scenario results in many technological problems and is a very costly solution.
- Transfer the digital circuitry into a DRAM process. Due to the highly specialized DRAM processes the only examples use processes that are one or two generations behind state of the art.
- Modify the processor architecture in such a way that the memory interaction takes place on two levels: relatively high interaction with a smaller memory in the processor chip, and less interaction with the bulk memory. This solution has not yet proved to have general applicability.
- Use special bonding techniques (flip chip) that allow to connect memory dies directly on top of the processor die.

Analog/Digital Partitioning

The tradeoff for analog circuits is determined by the performance that is required. In Section 11.4 several considerations on the expected performance issues of analog circuits in advanced CMOS processes were discussed. In Table 11.11 these items are

TABLE 11.11 Summary of CMOS Technology Scaling on Analog Performance

Analog Parameter	Effect in Advanced Process
Transconductance g_m	Limited by velocity saturation
Output impedance R_{out}	Limited by static feedback
Load capacitance C_{load}	Diffusion capacitance rises
Cutoff frequency	Marginal increase
Accuracy	Slight improvement for equal size
Signal swing	Reduces with power supply
Noise	Surface PMOST more noisy than buried
Passive elements	Raise costs

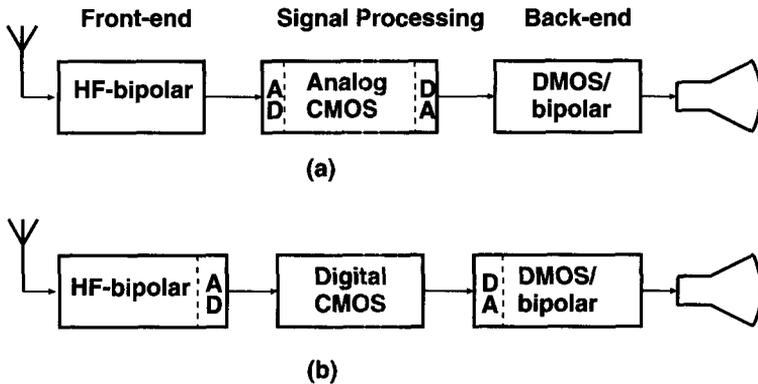


Figure 11.29 Alternative chip partitionings, using CMOS technology with analog extensions (a) or without extensions (b).

summarized.¹⁴ Digitally optimized CMOS is becoming less suited for today's forms of analog signal processing. However, as systems evolve, the requirements on analog signal processing (filtering, AGC, ADC) tend to increase. In addition, the continuous demand for higher-quality drives the required analog performance to the top. These two tendencies are clearly conflicting and the question arises what ways exist to optimize the overall system.

Figure 11.29a,b shows different partitionings. The upper scenario (a) allows some adaptations to the digital CMOS process. Looking at the number of additional masks several levels can be considered:

- Do not use any masks and maintain digital CMOS compatibility as far as possible. A lot of analog performance degradation can be circumvented if the digital CMOS process is optimized with the analog needs in mind. The matching behavior has already been discussed. Some minor process modifications can allow decent passive components. A major item is analog characterization in which capacitors, resistors, and MOS transistors should be specified as closely as possible.
- Another approach in circumventing reduced analog performance is to implement a form of correction or calibration. In some circuits the analog properties can be improved by a modest intervention (e.g., bandgap spread can be halved with 1-bit information). Such forms of digital correction (via DSP or PROM codes) or other means of trimming are feasible, but have impact on existing test equipment and packaging. Despite these measures, some deteriorating facts remain. Lowering the signal level while keeping the dynamic range or signal-to-noise ratio in place will simply mean (quadratically) lower impedance levels. The straightforward consequence is an increased power consumption. In many cases lowest overall analog power consumption is achieved at the largest voltage swing.

- A further step is to add one or two masks. The masks could be used for double-poly capacitors or resistor layers. Three or four masks would allow a form of BiCMOS. However, introduction of this CMOS adaptation is commercially risky. Such a process will always lag most recent CMOS developments. Moreover, lower yield and higher cost per square millimeter will occur. The process is neither optimum for analog nor digital.

In the other scenario of Figure 11.29*b* digital CMOS is expected to evolve to a digital-only monoculture. Analog performance in CMOS is extended as far as the pure digital technology allows and present designs are replaced by improved designs. However, more demanding analog functions will require moving the analog parts off the CMOS chip and moving them to the chips in the “shell”; see Figure 11.3. The different system/chip partitioning resulting from this approach leads to many chip-to-chip connections with potentially more interference. By itself, this may pose a large problem in situations such as an *A/D* converter with 10 output pads placed on a chip that has to accept inputs on a 10- μ V level. For other functions the need for one-package solutions may translate in multichip package (MCP) technology. Advantages in this scenario can be found in the separated design flow, test, CAD, and ease of shrink in various technology generations. The digital chip may shrink with every generation of CMOS, while the analog chips will be transferred through the generations although at a slower pace.

Moving from one to the other scenario has tremendous consequences for system chip sets, as any shift of functionality always affects at least two chips. Choices on partitioning and the location of the analog/digital interface are influenced by the realization possibilities of the technologies.

Multichip Packaging

The dilemma that exists in the two combinations of digital CMOS with high-density memory and digital CMOS with high-performance analog can be partly circumvented by choosing an alternative packaging technology. Multichip packaging allows combining two or more dies in one package. In Figure 11.30 three dies form a system-in-package. The analog die performs the interface functions with the outside world and takes care of clock generation, data conversion, and power interfacing. The digital processor has been designed in a very advanced CMOS process, and the memory is processed in standard DRAM. Many forms of multidie packaging exist, here are a few examples:

A multidie package is a more or less standard single-die package in which two dies are fixed with a number of mutual connections. This very cheap solution of combining dies is used in various products.

In a chip-on-chip package one die is placed upside down on the other die, this form of connecting two dies together is used in DSP-memory situations.

In a multichip module a special substrate is used to attach the dies and a few passive components. In passive integration the silicon substrate contains some (passive) elements and wiring layers.

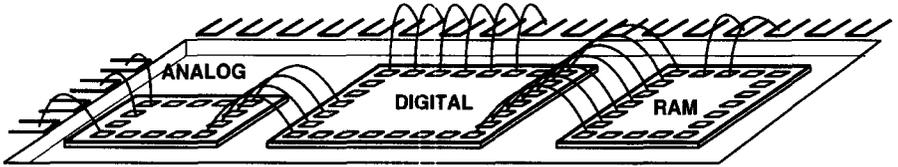


Figure 11.30 Partitioning within one package.

As these packaging technologies become economically more and more attractive, the system-on-chip issue may shift to system-in-package.

11.7 SYSTEM-ON-CHIP CONCEPT

11.7.1 One-Chip Television Chip

System-on-chip projects are mostly associated with large digital processors combined with some peripherals. These projects will certainly come true one day, today however, many high-volume systems on chip are realized in the analog domain. The reason is that traditional systems, such as standard television and radio, do not require the specific qualities of the digital domain such as storage and computing power. The first example of a system-on-chip device is a 30-mm² BiCMOS chip that performs the entire video decomposition after the tuner (see Figs. 11.31 and 11.32).

This chip is fed with a 38-MHz intermediate-frequency signal from the tuner and performs all functions needed for video decomposition. These functions are detection, sync separation, audio separation, video identification, color demodulation, and matrix and various forms of geometry correction. The signal processing is in the analog domain; filtering is performed by gyrator structures or $g_m C$ filters.

Although the function is completely determined by the construction of the circuit several control facilities are present. The circuit detects the video signal and DACs control various internal settings. A control bus allow a change in the functionality. About a quarter of the chip is digital.

11.7.2 Digital Video Front-End Chip

Figures 11.33 and 11.34 show a digital video signal processor. From a functionality point of view there is an overlap with the previous example. This chip accepts various forms of baseband video signal and decomposes them into the required output streams. The chip consists of an analog input processing part and a digital video signal processor. In the analog part the two larger blocks represent the A/D converters, additional circuitry enables gain control, multiplexing, and filtering. The digital part is optimized for PC use and allows various pixel configurations and several adaptations to the PC interface.

This IC is not fully equivalent to the IC discussed previously; therefore, a comparison of chip area is not very meaningful. It is, however, possible to look at the

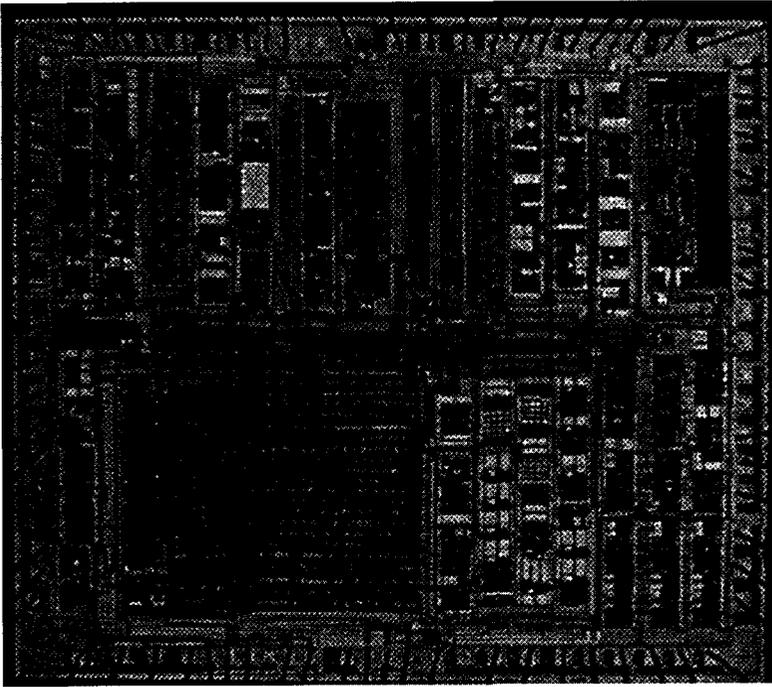


Figure 11.31 One-chip television IC in 1- μm BiCMOS technology. (Courtesy Philips Semiconductor Nijmegen, Hans Menting.)

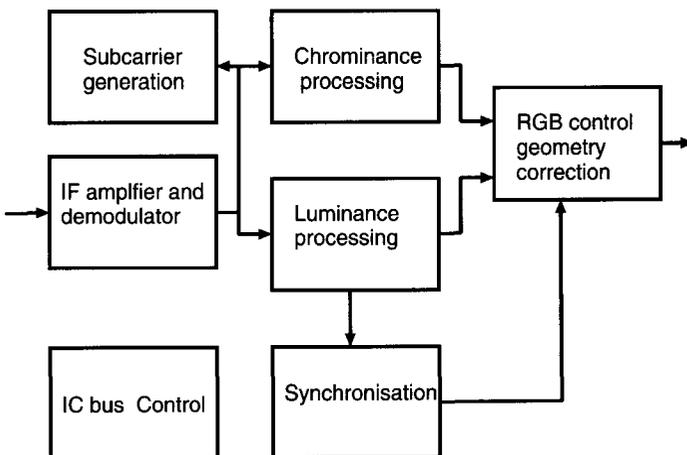


Figure 11.32 Block diagram of one-chip television IC.

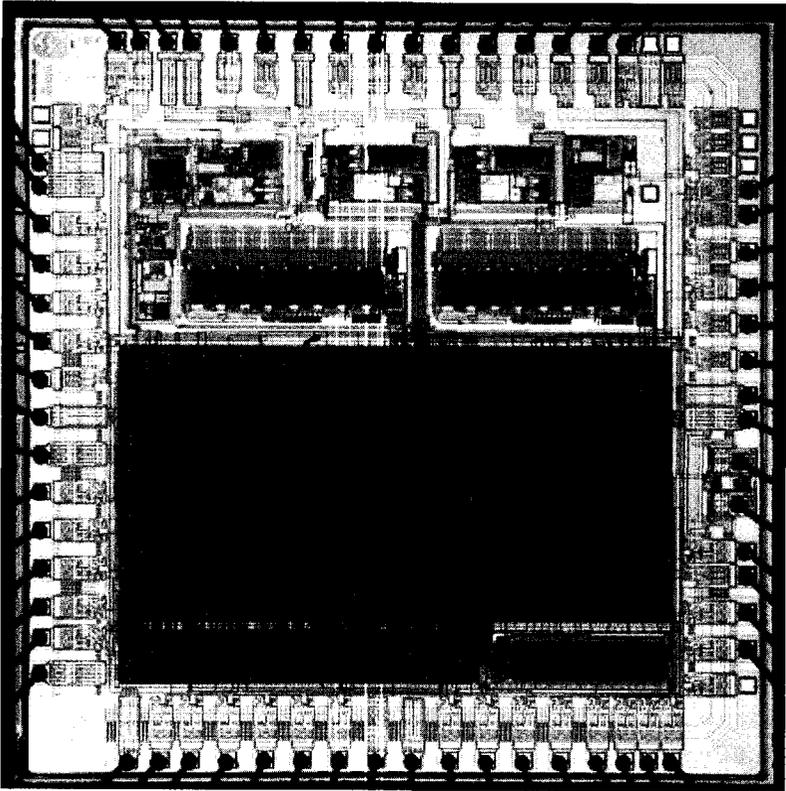


Figure 11.33 Digital one-chip television IC in 0.5- μm CMOS technology. (Courtesy Philips Semiconductor Hamburg, Robert Meyer.)

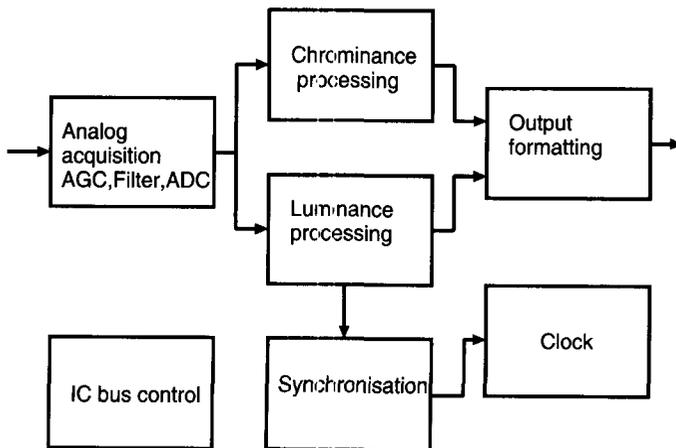


Figure 11.34 Block diagram of one-chip television IC.

TABLE 11.12 Area Comparison of Various Functions in Video Processing

	Digital		
	Analog 1- μm BiCMOS (mm^2)	1- μm CMOS (mm^2)	0.35- μm CMOS (mm^2)
Baseband circuit	0.8	3.3	0.8
Filter circuit	0.6	16	2.3
Timing circuit	1.6	5.6	0.6
Delay circuit	10	15.4	2.1
A/D conversion	N.A.	10	4

implementation of parts that are more or less equivalent in function. Table 11.12 indicates that a comparison between the various analog and digital implementation technologies is far from trivial. In technologies with equal linewidths the BiCMOS technology clearly is superior to the CMOS counterpart. At a difference of approximately three generations (halving the line width) the area difference becomes less. For filter circuits, in this case bandpass filters, analog implementation is still superior. Circuits with more digital functionality are comparable in area.

11.7.3 One-Chip Oscilloscope

Digital signal processing is especially advantageous where large amounts of data have to be stored or various forms of “non-linear” signal processing are required. Figures 11.35 and 11.36 show a circuit intended for a one-chip oscilloscope channel. In the digital domain sampling of data is inevitable and consequently aliasing of signals occurs. In addition to that, the waveform is displayed on a rastered display, causing a second source of aliasing. In this chip⁴⁷ methods are applied to get around the display artifacts by stochastic mapping of pixels and intensity scaling. Samples are converted to the digital domain and then stored, and their position relative to the display columns is calculated and allocated via a stochastic algorithm. Based on the number of sample points in a column, a 2-bit intensity level is determined. Because of this strong nonlinear operation, rarely occurring events can be made visible next to strong repetitive events. Finally, the 2-bit code is mapped on a display store and read out. This chip clearly illustrates the advantages of digital processing perfect storage and nonlinear signal operations.

11.8 SUMMARY AND FUTURE TRENDS

System-on-chip ICs are a natural step in the application of silicon technology for low-cost consumer applications. The combination of various forms of digital signal processing, memories, and analog interfaces enables implementation of most tasks in modern systems in one package. In this chapter we discussed various

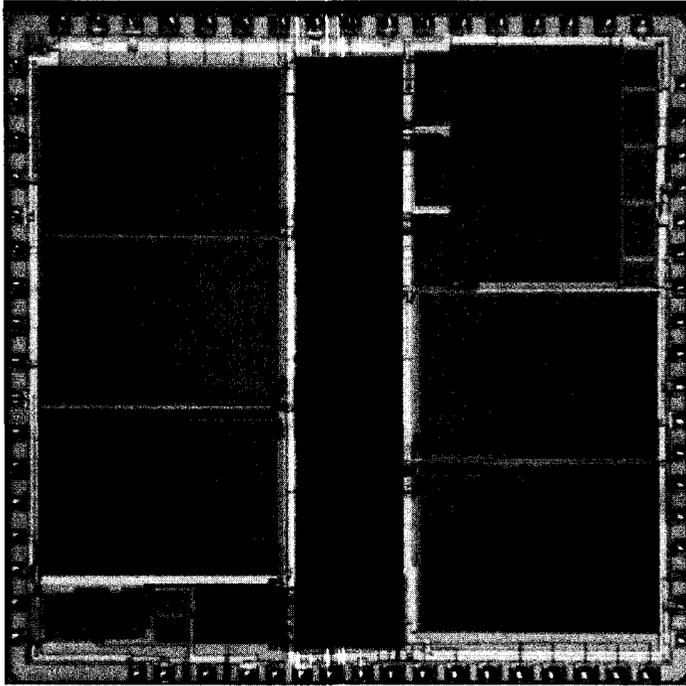


Figure 11.35 Digital one-chip oscilloscope IC in 0.5- μm CMOS technology. The processing core is shown in the middle with memory blocks on both sides. The analog part is shown on the lower left. (Courtesy Philips Research, Maarten Vertregt.)

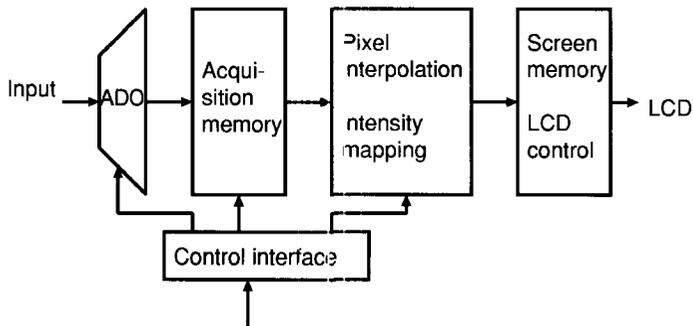


Figure 11.36 Block diagram of the one-chip oscilloscope IC.

consequences of the combination of analog and digital blocks in one CMOS technology. Technology tradeoffs, especially with respect to matching or threshold voltage variation, have been presented as well as a basic understanding of the functionality in analog-to-digital conversion.

As CMOS feature sizes decrease and power supply voltages reduce, it becomes more difficult to integrate all desired functionality into one die. Interfacing toward

the physical environment at a required quality level will become more difficult, and a different partitioning scheme between analog technology and digital technology may well be ahead.

ACKNOWLEDGMENT

Many colleagues have contributed to the contents of this chapter with discussions and input. In particular the author would like to mention the members of the “mixed-signal circuits and systems” group of Philips Research Laboratories in the Netherlands and colleagues in the Philips Semiconductor product division.

REFERENCES

1. M. Nakamura, “Challenges in Semiconductor Technology for Multi-Megabit Network Services,” *Int. Solid-State Circuits Conf. 98*, San Francisco, 1998, Vol. 41, p. 16.
2. J. Danneels, “GSM and Beyond—the Future of Access Network,” *Int. Solid-State Circuits Conf. 98*, San Francisco, 1998, Vol. 41, p. 22.
3. L. S. Milor, “A Tutorial Introduction to Research on Analog and Mixed-Signal Circuit Testing,” *IEEE Trans. Circ. Syst.-II CAS-45*, 1389 (1998).
4. Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors, Technology Needs*, Nov. 1997.
5. H. Komiya, “Future Technological and Economic Prospects for VLSI,” *Int. Solid-State Circuits Conf. 93*, San Francisco, 1993, p. 16.
6. A. W. M. v.d. Eenden and N. A. M. Verhoeckx, *Discrete Time Signal Processing: an Introduction*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
7. J. Borel, “Technologies for Multimedia Systems on a Chip,” *Int. Solid-State Circuits Conf. 97*, San Francisco, 1997, Vol. 40, p. 18.
8. R. W. Brodersen, “The Network Computer and its Future,” *Int. Solid-State Circuits Conf. 97*, San Francisco, 1997, Vol. 40, p. 32.
9. H. Veendrick, *Deep Submicron CMOS ICs*, Kluwer, Deventer, 1998.
10. L. Bolcioni et al. “A 1V 350 μ W Voice-controlled H.263 Video Decoder for Portable Applications,” *Int. Solid-State Circuits Conf.* San Francisco, 1998, Vol. 40, p. 112.
11. T. Sunaga, H. Miyatake, K. Kitamura, K. Kasuya, T. Saitoh, M. Tanaka, N. Tanigaki, Y. Mori, and N. Yamasahi, “DRAM Macros for ASIC Chips,” *IEEE J. Solid-State Circ. SC-30*, 1006 (1995).
12. S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1981.
13. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modeling for Circuit Design*, Springer-Verlag, Vienna/New York, 1990.
14. M. J. M. Pelgrom and M. Vertregt, “CMOS Technology For Mixed Signal ICs,” *Solid-State Electron.*, 967 (1997).
15. J. B. Shyu, G. C. Temes, and K. Yao, “Random Errors in MOS Capacitors,” *IEEE J. Solid-State Circ. SC-17*, 1070 (1982).

16. J. B. Shyu, G. C. Temes, and F. Krummenacher, "Random Error Effects in Matched MOS Capacitors and Current Sources," *IEEE J. Solid-State Circ.* **SC-19**, 948 (1984).
17. K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design," *IEEE J. Solid-State Circ.* **SC-21**, 1057 (1986).
18. M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching Properties of MOS Transistors," *IEEE J. Solid-State Circ.* **SC-24**, 1433 (1989).
19. M. J. M. Pelgrom, M. Vertregt, and H. Tuinhout, "Matching of MOS Transistors," MEAD course material, 1998.
20. Y. Tsividis, *Mixed Analog-Digital VLSI, Devices and Technology: An Introduction*, McGraw-Hill, New York, 1996.
21. M. J. M. Pelgrom, H. Tuinhout, and M. Vertregt, "Transistor Matching in Analog CMOS applications," *Int. Electron Devices Meeting*, San Francisco, Dec. 1998.
22. H. Tuinhout, M. J. M. Pelgrom, R. Penning de Vries, and M. Vertregt, "Effects of Metal Coverage on MOSFET Matching," *Int. Electron Devices Meeting*, San Francisco, Dec. 1996.
23. M. J. v. Dort and D. B. M. Klaassen, "Circuit Sensitivity Analysis in Terms of Process Parameters," *Int. Electron Devices Meeting*, Washington, DC, 1995, p. 37.3.1.
24. T. Mizuno, J. Okamura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuation Due To Statistical Variation of Channel Dopant Number in MOSFETs," *IEEE Trans. Electron Devices* **ED-41**, 2216 (1994).
25. E. Dijkstra et al., "Low Power Oversampled A/D Converters," in R. J. v.d. Plassche, ed., *Advances in Analog Circuit Design*, Kluwer, 1995, p. 89.
26. E. Vittoz, "Low Power Low-Voltage Limitations and Prospects in Analog Design," in R. J. v.d. Plassche, ed., *Advances in Analog Circuit Design*, Kluwer, 1995, p. 3.
27. E.J. Swanson, "Analog VLSI Data Converters — The First 10 Years," *Proc. ESSCIRC* **95**, 25 (1995).
28. M. J. M. Pelgrom, "Low-power High-speed A/D Conversion," *ESSCIRC94, Low-Power Workshop*, Ulm, Sept. 23, 1994.
29. P. Kinget and M. Steyaert, "Impact of Transistor Mismatch on the Speed Accuracy Power Trade-off," *CICC96*, San Diego, May 1996.
30. M. J. M. Pelgrom, "Low-Power CMOS Data Conversion," in E. Sanchez-Sinencio and A. Andreou, eds., *Low-Voltage Low-Power Integrated Circuits and Systems*, IEEE Press, New York, 1998.
31. K. Kusumoto et al., "A 10 b 20 MHz 3C mW Pipelined Interpolating CMOS ADC," *Int. Solid-State Circuits Conf.*, San Francisco, 1993, p. 62.
32. M. J. M. Pelgrom and M. Roorda, "An Algorithmic 15 bit CMOS Digital-to-Analog Converter," *IEEE J. Solid-State Circ.* **SC-23**, 1402 (1988).
33. J. R. Naylor, "A Complete High-Speed Voltage Output 16-bit Monolithic DAC," *IEEE J. Solid-State Circ.* **SC-18**, 729 (Dec. 1983).
34. D. W. J. Groeneveld, H. J. Schouwenaars, H. A. H. Termeer, and C. A. A. Bastiaansen, "A Self-Calibration Technique for Monolithic High-resolution D/A Converters," *IEEE J. Solid-State Circ.* **SC-24**, 1517 (1989).
35. M. J. M. Pelgrom, "A 10 b 50 MHz CMOS D/A Converter with 75 Ω Buffer," *IEEE J. Solid-State Circ.* **SC-25**, 1347 (1990).

36. J. C. Candy and G. C. Temes, eds., *Oversampling Delta-Sigma Data Converters: Theory, Design and Simulation*, IEEE, New York, 1992.
37. P. J. A. Naus and E. C. Dijkmans, "Multi-bit Oversampled $\Sigma\Delta$ A/D Converters as Front-end for CD Players," *IEEE J. Solid-State Circ.* **SC-26**, 905 (1991).
38. E. J. van der Zwan and E. C. Dijkmans, "A 0.2 mW CMOS $\Sigma\Delta$ Modulator for Speech Coding with 80 dB Dynamic Range," *IEEE J. Solid-State Circ.* **SC-31**, 1873 (Dec. 1996).
39. R. van de Plassche, "Integrated Analog-to-Digital and Digital-to-Analog Converters," Kluwer, The Netherlands, 1994.
40. B. Nauta and A. G. W. Venes, "A 70 Ms/s 110 mW 8-b CMOS Folding and Interpolating A/D Converter," *IEEE J. Solid-State Circ.* **SC-30**, 1302 (Dec. 1995).
41. K. Bult et al., "A 170 mW 10b 50 Ms/s CMOS ADC in 1 mm²," *Int. Solid-State Circuits Conf.* San Francisco, (1997), p. 136.
42. A. G. F. Dingwall and V. Zazzu, "An 8-MHz CMOS Subranging 8-bit A/D Converter," *IEEE J. Solid-State Circ.* **SC-20**, 1138 (Dec. 1985).
43. A. N. Karanicolas, H.-S. Lee, and K. L. Barcrania, "A 15-b 1-Msample/s Digitally Self-Calibrated Pipeline ADC," *IEEE J. Solid-State Circ.* **SC-28**, (Dec. 1993).
44. T. Cho and P. R. Gray, "A 10 bit, 20 MS/sec, 35 mW Pipeline ADC in 1.2 Micron CMOS," *IEEE J. Solid-State Circ.* **SC-30** (April 1995).
45. N. Fukushima et al., "A CMOS 40 MHz 8 b 105 mW Two-Step ADC," *Int. Solid-State Circuits Conf.* San Francisco, 1989, Vol. 32, p. 14.
46. M. J. M. Pelgrom, A. C. v. Rens, M. Vertregt, and M. B. Dijkstra, "A25-Ms/s 8-bit CMOS A/D Converter for Embedded Application," *IEEE J. Solid-State Circ.* **SC-29**, 879 (1994).
47. M. Vertregt et al., "A 0.4 W Mixed-Signal Digital Storage Oscilloscope Processor with Moire Prevention, Embedded 393 Kbit RAM and 50 MS/s 8 b ADC," *Int. Solid-State Circuits Conf.*, San Francisco, 1988, Vol. 41, p. 114.
48. A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Tokyo, 1965.

PROBLEMS

- 11.1 How much power will a digital circuit with 1000 nodes require for running at a data rate of 10 MHz, 3.3 V, and 0.1 pF load per node? Assume that there is no correlation between consecutive data samples.
- 11.2 Find the variance of the threshold voltage of a 10/0.5 transistor in a 7-nm gate oxide process.
- 11.3 The basic matching relation for threshold voltages ($1/\sqrt{WL}$) requires enlarging devices for better voltage matching. Now consider a MOS switch connected to a storage capacitor. Why should this switch be as small as possible for low variation of the stored charge?
- 11.4 Calculate from the basic current law

$$I_{DD} = \frac{W\beta}{2L}(V_{GS} - V_T)^2$$

the crossover point where errors in the current factor become more important than errors in the threshold voltage.

- 11.5 Is it possible to have a DNL that exceeds the INL by a factor of 2?
- 11.6 Is it possible to reduce the DNL figure of an ADC (analog/digital converter) by using oversampling as indicated by Eq. 11.25?
- 11.7 A signal frequency band between 33 and 38 MHz is sampled by an ADC running at 30 Ms/s (million samples per second). What is the lowest frequency band that will occur, and what happens if the sampling rate is lowered to 25 Ms/s?
- 11.8 Assume that the maximum current matching error in Figure 11.16 between unit current sources is 10%. What is the optimum choice between segmentation and binary implementation if monotonicity is required?
- 11.9 The effective density of on-chip DRAM memory is approximately $8 \mu\text{m}^2$ per bit, while external memories are available at $1 \text{ Mb}/1.6 \text{ mm}^2$. However, the interfacing to external memories requires an additional 7 mm^2 of area. For which memory size is an external memory more attractive areawise?
- 11.10 Which functional units (e.g., RF section) require additional technological features?

List of Symbols

Symbol	Description	Unit
a	Lattice constant	Å
\mathcal{B}	Magnetic induction	Wb/m ²
c	Speed of light in vacuum	cm/s
C	Capacitance	F
\mathcal{D}	Electric displacement	C/cm ²
D	Diffusion coefficient	cm ² /s
E	Energy	eV
E_c	Bottom of conduction band	eV
E_F	Fermi energy level	eV
E_g	Energy bandgap	eV
E_v	Top of valence band	eV
\mathcal{E}	Electric field	V/cm
\mathcal{E}_c	Critical field	V/cm
\mathcal{E}_m	Maximum field	V/cm
f	Frequency	Hz (cps)
$F(E)$	Fermi–Dirac distribution function	
h	Planck constant	J·s
$h\nu$	Photon energy	eV
I	Current	A
I_c	Collector current	A
J	Current density	A/cm ²
J_t	Threshold current density	A/cm ²
k	Boltzmann constant	J/K
kT	Thermal energy	eV
L	Length	cm or μm
m_0	Electron rest mass	kg
m^*	Effective mass	kg
n	Density of free electrons	cm ⁻³
n_i	Intrinsic density	cm ⁻³
N	Doping concentration	cm ⁻³
N_A	Acceptor impurity density	cm ⁻³
N_C	Effective density of states in conduction band	cm ⁻³
N_D	Donor impurity density	cm ⁻³
N_V	Effective density of states in valence band	cm ⁻³

Symbol	Description	Unit
p	Density of free holes	cm^{-3}
P	Pressure	Pa
q	Magnitude of electronic charge	C
Q_{it}	Interface trapped charge	charges/cm ²
R	Resistance	Ω
t	Time	s
T	Absolute temperature	K
v	Carrier velocity	cm/s
v_s	Saturation velocity	cm/s
v_{th}	Thermal velocity	cm/s
V	Voltage	V
V_{bi}	Built-in potential	V
V_{EB}	Emitter-base voltage	V
V_B	Breakdown voltage	V
W	Thickness	cm or μm
W_B	Base thickness	cm or μm
x	x direction	
∇	Differential operator	
∇T	Temperature gradient	K/cm
ϵ_0	Permittivity in vacuura	F/cm
ϵ_s	Semiconductor permittivity	F/cm
ϵ_i	Insulator permittivity	F/cm
ϵ_s/ϵ_0 or ϵ_i/ϵ_0	Dielectric constant	
τ	Lifetime or decay time	s
θ	Angle	rad
λ	Wavelength	μm or nm
ν	Frequency of light	Hz
μ_0	Permeability in vacuum	H/cm
μ_n	Electron mobility	$\text{cm}^2/(\text{V}\cdot\text{s})$
μ_p	Hole mobility	$\text{cm}^2/(\text{V}\cdot\text{s})$
ρ	Resistivity	$\Omega\cdot\text{cm}$
ω	Angular frequency ($2\pi f$ or $2\pi\nu$)	Hz
Ω	Ohm	Ω

International System of Units (SI Units)

Quantity	Unit	Symbol	Dimensions
Length ^a	meter	m	
Mass	kilogram	kg	
Time	second	s	
Temperature	kelvin	K	
Current	ampere	A	
Light intensity	candela	Cd	
Angle	radian	rad	
Frequency	hertz	Hz	1/s
Force	newton	N	(kg · m)/s ²
Pressure	pascal	Pa	N/m ²
Energy ^a	joule	J	N · m
Power	watt	W	J/s
Electric charge	coulomb	C	A · s
Potential	volt	V	J/C
Conductance	siemens	S	A/V
Resistance	ohm	Ω	V/A
Capacitance	farad	F	C/V
Magnetic flux	weber	Wb	V · s
Magnetic induction	tesla	T	Wb/m ²
Inductance	henry	H	Wb/A
Light flux	lumen	Lm	Cd · rad

^a It is more common in the semiconductor field to use cm for length and eV for energy (1 cm = 10⁻² m, 1 eV = 1.6 × 10⁻¹⁹ J).

Unit Prefixes^a

Multiple	Prefix	Symbol
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^9	giga	G
10^6	mega	M
10^3	kilo	k
10^2	hecto	h
10	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

^aAdopted by International Committee on Weights and Measures. (Compound prefixes should not be used; e.g., not $\mu\mu$ but p.)

Greek Alphabet

Letter	Lowercase	Uppercase
Alpha	α	A
Beta	β	B
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ϵ	E
Zeta	ζ	Z
Eta	η	H
Theta	θ	Θ
Iota	ι	I
Kappa	κ	K
Lambda	λ	Λ
Mu	μ	M
Nu	ν	N
Xi	ξ	Ξ
Omicron	\omicron	O
Pi	π	Π
Rho	ρ	P
Sigma	σ	Σ
Tau	τ	T
Upsilon	υ	Υ
Phi	ϕ	Φ
Chi	χ	X
Psi	ψ	Ψ
Omega	ω	Ω

Physical Constants

Quantity	Symbol	Value
Angstrom unit	\AA	$1 \text{\AA} = 10^{-1} \text{ nm} = 10^{-4} \text{ }\mu\text{m} = 10^{-8} \text{ cm} = 10^{-10} \text{ m}$
Avogadro constant	N_{av}	6.02214×10^{23}
Bohr radius	a_{B}	0.52917\AA
Boltzmann constant	k	$1.38066 \times 10^{-23} \text{ J/K}(R/N_{\text{av}})$
Elementary charge	q	$1.60218 \times 10^{-19} \text{ C}$
Electron rest mass	m_0	$0.91094 \times 10^{-30} \text{ kg}$
Electron volt	eV	$1 \text{ eV} = 1.60218 \times 10^{-19} \text{ J} = 23.053 \text{ kcal/mol}$
Gas constant	R	$1.98719 \text{ cal/mol} \cdot \text{K}$
Permeability in vacuum	μ_0	$1.25664 \times 10^{-8} \text{ H/cm}(4\pi \times 10^{-9})$
Permittivity in vacuum	ϵ_0	$8.85418 \times 10^{-14} \text{ F/cm}(1/\mu_0 c^2)$
Planck constant	h	$6.62607 \times 10^{-34} \text{ J} \cdot \text{s}$
Reduced Planck constant	\hbar	$1.05457 \times 10^{-34} \text{ J} \cdot \text{s}(h/2\pi)$
Proton rest mass	M_p	$1.67262 \times 10^{-27} \text{ kg}$
Speed of light in vacuum	c	$2.99792 \times 10^{10} \text{ cm/s}$
Standard atmosphere		$1.01325 \times 10^5 \text{ Pa}$
Thermal voltage at 300 K	kT/q	0.025852 V
Wavelength of 1-eV quantum	λ	$1.23984 \text{ }\mu\text{m}$

Properties of Si at 300 K

Properties	Si
Atoms/cm ³	5.02×10^{22}
Atomic weight	28.09
Breakdown (V/cm)	$\sim 3 \times 10^5$
Crystal structure	Diamond
Density (g/cm ³)	2.329
Dielectric constant	11.9
Effective density of states in conduction band, N_c (cm ⁻³)	2.86×10^{19}
Effective density of states in valence band, N_v (cm ⁻³)	2.66×10^{19}
Effective mass, m^*/m_0	
Electrons	$m_{l}^* = 0.92$
	$m_{t}^* = 0.20$
Holes	$m_{lh}^* = 0.15$
	$m_{hh}^* = 0.54$
Electron affinity, χ (V)	4.05
Energy gap (eV) at 300 K	1.124
Index of refraction	3.42
Intrinsic carrier concentration (cm ⁻³)	9.65×10^9
Intrinsic Debye length (μm)	41
Intrinsic resistivity ($\Omega \cdot \text{cm}$)	3.3×10^5
Lattice constant (\AA)	5.43102
Linear coefficient of thermal expansion, $\Delta L/L\Delta T$ ($^{\circ}\text{C}^{-1}$)	2.59×10^{-6}
Melting point ($^{\circ}\text{C}$)	1412
Minority-carrier lifetime (s)	3×10^{-2}
Mobility (drift) [$\text{cm}^2/(\text{V} \cdot \text{s})$]	
μ_n (electrons)	1450
μ_p (holes)	505
Optical-phonon energy (eV)	0.063
Phonon mean free	
Path λ_0 (\AA)	76 (electrons)
	55 (holes)
Specific heat [$\text{J}/(\text{g} \cdot ^{\circ}\text{C})$]	0.7
Thermal conductivity at 300 K [$\text{W}/(\text{cm} \cdot \text{K})$]	1.31
Thermal diffusivity (cm ² /s)	0.9
Vapor pressure (Pa)	1 at 1650 $^{\circ}\text{C}$
	10^{-6} at 900 $^{\circ}\text{C}$

- Access gates, gated-access silicon memory cells, 444–446
- Accumulation-mode (AM) SOI MOSFETs:
 - characteristics of, 230–232
 - subthreshold slope, 237
 - threshold voltage, 234
- Accuracy:
 - block integration, system-on-chip concepts, sampling rate and, 665–666
 - system-on-chip technology limits, component matching, 650–651
- AC models, lifetime estimation, 301–302
- Acoustic deformation potential (ADP), intervalley scattering, 182
- Acoustic phonon scattering:
 - equations and calculations, 177–181
 - MOSFET transport properties, 103–104
 - mobility models, 108
- Active region:
 - bipolar transistors, 20
 - inner-box shaped transistor, static characteristics, 31–35
- Additive noise, voltage-controlled oscillators (VCOs), 524–525
- Address transition detection (ATD) circuit, SRAM (static random access memory) architecture, 369–370
- Admittance matrix, bipolar transistors, 23
- Airy functions, MOSFET inversion layers, 2DEG quantization, Boltzmann transport equation (BTE) simulations, 203–210
- Alias filtering, block integration, system-on-chip concepts, system function implementation, 667–670
- Alloy annealing, hot-carrier effects (HCE), back-end processing, 321–322
- Alpha particles:
 - DRAM (dynamic random access memory) soft errors, 344–345
 - SRAM (static random access memory) soft errors, 365–366
- Alternate metal ground (AMG) technique,
 - floating-gate memory arrays, UV EPROMS, 395–399
- Amplituded-modulated signals, voltage-controlled oscillators (VCOs), 524–525
- Analog circuits:
 - CMOS technology:
 - integration levels, 477–479
 - low-noise amplifiers, 493–508
 - highly integrated transceivers, 504–508
 - mixers and frequency translation, 508–520
 - highly integrated transceivers, 513–520
 - mixer fundamentals, 508–513
 - silicon-on-insulator (SOI) devices, 255–256
 - system-on-chip concepts:
 - analog-to-digital conversion terminology, 651–655
 - conversion architectures, 655–664
 - A/D converters, 659–662
 - comparator matching, 662–664
 - D/A converters, 655–659
- Analog/digital partitioning, block integration, system-on-chip concepts, 673–675
- Analog-to-digital (A/D) conversion, system-on-chip concepts:
 - conversion architecture, 659–662
 - terminology, 651–655
- Ansatz construction, Boltzmann transport equation (BTE), 189–192
- Antenna ratios, hot-carrier devices:
 - plasma damage, 314–317
 - resist damage, 316
- Anti reflecting (AR) material, silicon-on-insulator (SOI) devices, laser recrystallization, 224
- Application-specific integrated circuits (ASICs), digital blocks, embedded modules on IC, 636–639
- Area-enhancement factor (AEF), DRAM (dynamic random access memory) circuits, memory cell scaling, 338–339
- Arithmetical circuits, heterogeneous ICs, system-on-chip concepts, 635–636

- Arrhenius equation, silicon nitride memory, decay rate, 435–438
- Auger recombination, inner-box shaped transistor, 28–29
- Autocorrelation, embedded modules on IC, system-on-chip concepts, 635
- Avalanche injection, floating-gate memory physics, charge transfer, 379–381
- Back-end processing, hot-carrier effects (HCE), 321–322
- Back-gate insulators, advanced MOSFET structures, 133–135
- Balanced mixer, CMOS technology, radiofrequency (RF) circuit design, 511–513
- Balance equations, Boltzmann transport equation (BTE), 195–200
- Bandgap energy, MOSFET models, device scaling, 86–88
- Band structure:
 - dispersion relationship, 152–160
 - equilibrium statistics, 163–170
 - floating-gate memory, 378
 - single-electron effect-mass equation, 150–152
- Band-to-band tunneling:
 - floating-gate memory physics, Fowler-Nordheim tunneling, 388–389
 - MOSFET parasitic effects, gate-induced drain leakage (GIDL), 122–124
- Band-to-band tunneling-induced substrate hot-electron (BBISHE) injection, MOSFET parasitic effects, gate-induced drain leakage (GIDL), 123–124
- Bandwidth (BW):
 - analog-to-digital (A/D) conversion, 653–655
 - block integration, system-on-chip concepts, signal processing and, 665
- Barium strontium titanate, DRAM (dynamic random access memory) circuits, memory cell scaling, 337–339
- Barkhausen criteria, voltage-controlled oscillators (VCOs), 523–525
 - highly integrated transceivers, 528–535
- Base sheet resistance, inner-box shaped transistor, static characteristics, 34–35
- BESOI (bond and etchback SOI):
 - smart power circuits, 259–261
 - techniques for, 227–228
- Bessel series expansion, digital applications of power transistors, cutoff and maximum oscillation frequencies, 586–597
- β ratio, SRAM (static random access memory) circuits, cell stability analysis, 364–365
- Biasing operations:
 - digital applications, large-signal power and efficiency, 598–607
 - field-enhanced tunnel injector flash (FETIF) EPROM, 411–412
 - floating-gate memory arrays, scaling trends, 429–430
 - NAND cell architecture, 416–417
 - T-cell Flash EEPROM, 407–410
- BiCMOS, bipolar transistors, 249–250
- Binary representation, system-on-chip concepts, D/A converters, 656–658
- Binary sensing, floating-gate memory cells, 391–393
- Bipolar junction transistor (BJT):
 - developmental history, 19
 - heterojunction bipolar transistor, 60–66
 - narrow-bandgap base, 63–65
 - pseudomorphic SiGe layers, 66
 - SiGe/Si material system, 65–66
 - wide-bandgap emitter, 62–63
 - history, 3–4, 19–20
 - inner-box shaped transistor, 25–51
 - high-frequency behavior, 35–39
 - static characteristics, 29–35
 - thermal effects, 39–51
 - heat flux and thermal diffusion, device performance, 43–45
 - one-dimensional heat transfer, analytical approach, 46–49
 - safe operation area (SOA), 49–51
 - second breakdown instabilities, 42–43
 - temperature current, 39–42
 - self-adjusted structures, 51–60
 - device and circuit results, 57–60
 - high current density operations, 55–57
 - polysilicon base and emitter contacts, 51–55
 - double-poly transistor with inside spacer, 53–55
 - structure and operating regimes, 20–25
- Bipolar-MOS “hybrid” devices, components of, 239–242
- Bipolar transistors, silicon-on-insulator (SOI) devices, 248–250
- Bit density, floating-gate memory arrays, UV EPROMS, 394–395
- Bit error rate (BER), system-on-chip concepts, analog-to-digital (A/D) conversion, 659–662
- Bit line pairs:
 - DRAM (dynamic random access memory) circuits:
 - charge share sensing, 340–341
 - memory array architectures, 346, 348–351

- sense amplifiers, 351–352
- field-enhanced tunnel injector flash (FETIF) EPROM, 412
- floating-gate memory arrays:
 - DINOR cell, 417–421
 - UV EPROMS, 400–402
- SRAM (static random access memory) circuits, sense operation, 362
- textured poly E²PROM cell, 406
- Bloch electrons:
 - band structure, dispersion relationship, 154–160
 - scattering theory, 173
 - semiclassical electron dynamics, 160–163
- Bloch-Floquet wavefunctions, semiclassical electron dynamics, 160–163
- Blocking dynamic range (BDR), CMOS low-noise amplifiers (LNA), 497–508
- Blocking signals, CMOS low-noise amplifiers (LNA), 497–508
- Block integration, system-on-chip concepts:
 - interference, 670–672
 - partitioning, 672–676
 - analog/digital partitioning, 673–675
 - digital/memory choices, 672–673
 - multichip packaging, 675–676
 - signal processing strategy, 665–667
 - modulated signal sampling, 666–667
 - sample rate and accuracy, 665–666
 - sampling rate and bandwidth, 665
 - system function implementation, 667–670
 - modulated signal conversion, 668
 - oversampled D/A converter, 668–669
- Body-centered cubic (BCC) reciprocal lattice, band structure, dispersion relationship, 159–160
- Body-effect coefficient:
 - bulk MOSFETs, 221–222
 - SOI MOSFET, 234–235
- Body factor, SOI MOSFETs, 235
- Boltzmann distribution, one-dimensional drain current MOSFET model, 75–76
- Boltzmann entropy, equilibrium statistics, Fermi-Dirac distribution, 166–170
- Boltzmann transport equation (BTE):
 - device simulation, 185–211
 - average quantities, 192–195
 - balance equations and method of moments, 195–200
 - carrier concentrations, 193
 - carrier kinetic energy, 193
 - current density, 194
 - ensemble relaxation times, 195
 - MOSFET simulations, 200–211
 - 2DEG quantization in inversion layers, 200–210
 - hydrodynamic studies, 210–211
 - semiclassical electron dynamics, 149–150
- Born-Von Karman periodic boundary condition, band structure, dispersion relationship, 157–160
- Boron compounds, high-speed digital applications:
 - boron penetration, 575–577
 - gate dielectrics for penetration suppression, 578–580
 - SCE-channel dopant N_{sub} engineering, 557–560
 - short-channel effects (SCE), junction depth processing, 563
- Bose-Einstein distribution:
 - equilibrium statistics, Fermi-Dirac distribution, 169–170
 - phonon scattering, 178–181
- Boundary conditions, first-order MOSFET models, one-dimensional Poisson equation, charge density, 77–79
- Bravais lattice, band structure:
 - dispersion relationship, 158–160
 - single-electron effect-mass equation, 150–152
- Breakdown voltages, inner-box shaped transistor, 27–29
- Brillouin zone:
 - band structure, dispersion relationship, 159–160
 - intervalley scattering, 181–182
 - semiclassical electron dynamics, 162–163
- Brooks-Herring model:
 - ionized impurity scattering, 174
 - relaxation time averaging, 175–176
- BSIM3v3 model, digital applications of power transistors, cutoff and maximum oscillation frequencies, 591–597
- Bulk-limited current-density expression, silicon nitride memory, 433–435
- Bulk MOSFET, structural evolution, 131–133
- Buried oxide layer (BOX), silicon-on-insulator (SOI) devices, SIMOX (separation by implanted oxygen), 225–227
- Byte-alterable E²PROMs, floating-gate memory arrays, 402–406
 - FLOTOX, lithographic definition in drain extension, 402–404
 - textured poly E²PROM cell, 404–406
- Capacitor, system-on-chip concepts, D/A converters, 656

- Capacitor area, DRAM (dynamic random access memory) circuits, memory cell scaling, 337–339
- Capacitor dielectric leakage, DRAM (dynamic random access memory) circuits, 343–344
- Capture cross-section:
 - charge-pumping detection, interface-state levels, 307
 - hot-carrier effect (HCE):
 - low gate voltage stresses, 287
 - oxide traps N_{ot} , 280
- Carrier-carrier scattering:
 - device simulation, 176–177
 - inner-box shaped transistor, temperature dependence, collector current, 40–42
- Carrier concentrations, Boltzmann transport equation (BTE), 193
- Carrier distribution, MOSFET models, device scaling, 86–88
- Carrier velocity, MOSFET transport properties, vs. electric field, 102–103
- Cascade voltage switch logic (CVSL):
 - CMOS digital switching, 489–492
 - CMOS low-noise amplifiers (LNA), highly integrated transceivers, 505–508
- Cauer filter, inner-box shaped transistor, one-dimensional heat transfer, 49
- CB (collector-base) junction:
 - bipolar transistors, 20–22
 - inner-box shaped transistor, 27–29
 - high-frequency behavior, 37–39
 - self-adjusted transistor structures, polysilicon base, 52–55
- CBiCMOS, bipolar transistors, 249–250
- Cell scaling, DRAM (dynamic random access memory) circuits, 335–339
- Cell stability analysis, SRAM (static random access memory) circuits, 364–365
- Central processing unit (CPU):
 - digital blocks, system-on-chip applications, 637–639
 - MOSFET technology and, 7
- Channel dopant N_{sub} engineering, high-speed digital applications, short-channel effects (SCE), 557–560
- Channel hot-electron (CHE) injection:
 - floating-gate memory arrays:
 - disturb failures, 424–426
 - scaling trends, 426–430
 - UV EPROMS, 397–399
 - floating-gate memory physics, charge transfer, avalanche injection, 381–383
 - triple-poly, virtual ground (TPVG) flash cell, 412–414
- Channeling, system-on-chip technology limits, systematic and random mismatch, 646
- Channel length modulation (CLM), MOSFET parasitic effects, output resistance, 128–130
- Characteristic length theory, MOSFETs, short-channel effects, 99–102
- Charge centroid, silicon nitride memory, tunneling-emission mechanisms, 433–435
- Charge characteristics, hot-carrier effect (HCE), oxide traps N_{ot} , 279
- Charge-coupled device (CCD), analog-to-digital (A/D) conversion, 654–655
- Charge density, first-order MOSFET models, one-dimensional Poisson equation, 76–79
- Charge-injection transistor (CHINT), ULSI applications, 8
- Charge-pumping techniques, hot-carrier effect (HCE):
 - damage localization, 307–308
 - interface-state capture cross section, 307
 - interface state detection, 303–304
 - intermediate gate voltage stresses, 281–283
 - oxide trap detection, 304–307
- Charge share sensing:
 - DRAM (dynamic random access memory) operation, 340–341
 - MOSFETs, short-channel effects, 89–97
- Charge transfer, floating-gate memory physics, 379–391
 - avalanche injection, 379–381
 - channel hot-electron injection, 381–383
 - Fowler-Nordheim tunneling, 385–389
 - hot-electron injection, 379
 - source-side electron injection, 383–384
 - substrate injection, 384–385
 - ultraviolet light erase, 389–391
- Chemical-mechanical polishing (CMP)
 - processes, MOSFETs, evolution of, historical development, 130–131
- Chemomechanical polishing, silicon-on-insulator (SOI) devices, UNIBOND material, 229–230
- Chip-on-chip packaging, partitioning, block integration, system-on-chip concepts, 675–676
- Class A operation, CMOS technology, power amplifiers, 536–538
- Class B operation, CMOS technology, power amplifiers, 537–539
- Class C operation, CMOS technology, power amplifiers, 539–540

- Clock-driven architecture, voltage-controlled oscillators (VCOs), highly integrated transceivers, 531–535
- Clock frequency:
 - block integration, system-on-chip concepts, signal processing and bandwidth, 665
 - technology trends in, 10
- CMOS technology:
 - analog and RF applications:
 - low-noise amplifiers, 493–508
 - highly integrated transceivers, 504–508
 - mixers and frequency translation, 508–520
 - highly integrated transceivers, 513–520
 - mixer fundamentals, 508–513
 - digital applications:
 - differential logic, 484–492
 - current mode logic (CML), 485–487
 - differential split-level logic (DSL) and cascode voltage switch logic (CVSL), 489–492
 - folded source-coupled logic (FSCL), 488–489
 - inverter dynamic characteristics, 481–484
 - inverter static characteristics, 479–481
 - future trends in, 540–541
 - integration levels, 477–478
 - operational amplifiers, low-voltage, low-power (LVLP) circuits, 258–259
 - power amplifiers, 535–540
 - fundamentals, 535–540
 - highly integrated transceivers, 540
 - voltage-controlled oscillators (VCOs):
 - fundamentals, 520–525
 - highly integrated transceivers, 525–535
- Collector current, inner-box shaped transistor, temperature dependence, 39–42
- Collector-substrate (CS) capacitances, self-adjusted transistor structures, polysilicon base, 52–55
- Column address strobe (CAS) signal:
 - DRAM (dynamic random access memory) circuits, 346–348
 - high-speed DRAM circuits:
 - extended data out (EDO) signal, 354
 - fast page mode (FPM), 354
 - synchronous DRAM (SDRAM), 355–356
- Common emitter equivalent circuit, inner-box shaped transistor, high-frequency behavior, 37–39
- Common source configuration:
 - CMOS low-noise amplifiers (LNA), common gate/source configuration, 502–508
 - digital applications of power transistors, cutoff and maximum oscillation frequencies, 587–597
- Comparator matching, system-on-chip concepts, analog-to-digital (A/D) conversion, 662–664
- Complementary MOSFET (CMOS) circuits:
 - clock frequency, 10
 - developmental history, 5–9, 19
 - MOSFETs, evolution of, historical development, 130–131
 - power dissipation, 11
- Component matching, system-on-chip technology limits, 646–651
 - local threshold variation, 651
 - low voltage, 649–650
 - power and accuracy limits, 650–651
- Computational power, increases in, 10–11
- Conditioning circuits, heterogeneous ICs, system-on-chip concepts, 635–636
- Conduction angle, CMOS technology, power amplifiers, 538–540
- Confined lateral selective epitaxy (CLSEG), silicon-on-insulator (SOI) devices, epitaxial lateral overgrowth (ELO) variations, 225
- Continuity equations:
 - Boltzmann transport equation (BTE), 187–188
 - balance equations, 196–200
 - inner-box shaped transistor, static characteristics, 30–35
- Continuous-wave (CW) lasers, silicon-on-insulator (SOI) devices, recrystallization, 223–224
- Conversion architectures, system-on-chip concepts, analog circuits, 655–664
 - A/D converters, 659–662
 - comparator matching, 662–664
 - D/A converters, 655–659
- Conwell-Weisskopf model, device simulation, 176
- Core processing units, self-adjusted transistor structures, device results, 59–60
- CoSi₂ salicide, high-speed digital applications, 571
- Cosmic rays:
 - DRAM (dynamic random access memory) soft errors, 344–345
 - SRAM (static random access memory) soft errors, 366–367
- Cost analysis, DRAM and SRAM chip manufacture, 372
- Coulomb blockade mechanism, quantum effect devices, 253–254

- Coulomb scattering, MOSFET transport
 - properties, 104
 - mobility models, 108
 - universal mobility, 106–107
- Coupled oscillators, voltage-controlled oscillators (VCOs), highly integrated transceivers, 529–535
- Coupling ratios, floating-gate memory physics, channel hot-electron (CHE) injection, 382–383
- Critical current density, inner-box shaped transistor, static characteristics, 33–35
- Critical dose, silicon-on-insulator (SOI) devices, SIMOX (separation by implanted oxygen), 226–227
- Crystal, band structure, dispersion relationship, 155–160
- Crystal momentum, semiclassical electron dynamics, 161–163
- Curie constant, ferroelectric memory, polarization and temperature dependence, 455
- Curie temperature, ferroelectric memory, materials and structure, 450–452
- Curie-Weiss behavior, ferroelectric memory, polarization and temperature dependence, 454–455
- Current continuity equations, MOSFET models, device scaling, 84–88
- Current-controlled oscillator (CCO), vs. voltage-controlled oscillators (VCOs), highly integrated transceivers, 533–535
- Current density, Boltzmann transport equation (BTE), 193
- Current dependency, inner-box shaped transistor:
 - high-frequency behavior, 38–39
 - static characteristics, 32–35
- Current mode logic (CML), CMOS digital switching, 485–487
- Cutoff oscillation frequencies:
 - digital applications, 585–597
 - mixed-signal circuits, feature size, system-on-chip concepts, 641–642
- Cutoff region, bipolar transistors, 20
- CVD TiSi_2 deposition, high-speed digital applications, raised source/drain structures, 574–575
- Czochralski growth, silicon-on-sapphire (SOS) material, 222–223
- Damage region:
 - charge-pumping detection, localization of, 307–308
 - length of, 311
- DCIV method, hot carrier measurement, 310
- De Broglie wavelength:
 - band structure:
 - dispersion relationship, 153–160
 - single-electron effect-mass equation, 151–152
 - Boltzmann transport equation (BTE), 214–216
 - ULSI simulations, 212–216
- Debye length:
 - first-order MOSFET models, one-dimensional Poisson equation, charge density, 78–79
 - ionized impurity scattering, 173–174
 - MOSFET models, device scaling, 85–88
- Decay rates, silicon nitride memory, 435–438
- Deep-depletion regions, MOSFET parasitic effects, gate-induced drain leakage (GIDL), 119, 121
- Deep-submicrometer CMOS transistor, high-speed digital applications, drive current models, 550–551
- Delay operations, embedded modules on IC, system-on-chip concepts, 634–635
- δ -function, scattering theory, 171–173
- Density of states (DOS):
 - Boltzmann transport equation (BTE), 192–193
 - equilibrium statistics, 163–165
 - effective mass calculations, 169–170
 - intervalley scattering, 181–182
 - phonon scattering, 180–181
- Depletion layers:
 - advanced MOSFET structures, 137–138
 - inner-box shaped transistor, 26–29
- Design rule, technology trends in, 9
- Destructive readout structure, ferroelectric memory architecture, 459–462
- Deuterium, hot-carrier effects (HCE), back-end processing, 321–322
- Device lifetime, MOSFET parasitic effects:
 - hot carriers, 112, 115
 - light doped drain (LDD) structure, 118–119
- Device modeling, inner-box shaped transistor, static characteristics, 29–35
- Device structures, MOSFETs, evolution of, 130–135
 - advanced MOSFET device structures, 133–135
 - historical development, 130–131
 - state-of-the-art bulk MOSFET, 131–133
- Diamond lattice materials, heterojunction bipolar transistor (HBT), 62–63
- Differential logic, CMOS technology, digital switching, 484–492
 - current mode logic (CML), 485–487
 - differential split-level logic (DSL) and cascode voltage switch logic (CVSL), 489–492

- folded source-coupled logic (FSCL), 488–489
- Differential mode equivalent circuit, voltage-controlled oscillators (VCOs), highly integrated transceivers, 527–535
- Differential nonlinearity (DNL):
 - analog-to-digital conversion, 653–655
 - system-on-chip concepts, analog-to-digital conversion, 662–664
- Differential split-level logic (DSL), CMOS digital switching, 489–492
- Diffusion capacitance, mixed-signal circuits, feature size, system-on-chip concepts, 641–642
- Diffusion current:
 - DRAM (dynamic random access memory) circuits, junction leakage, 342
 - high-speed digital applications, short-channel effects (SCE), junction depth processing, 562–563
- Digital applications:
 - CMOS technology:
 - differential logic, 484–492
 - current mode logic (CML), 485–487
 - differential split-level logic (DSL) and cascode voltage switch logic (CVSL), 489–492
 - folded source-coupled logic (FSCL), 488–489
 - integration levels, 477–479
 - inverter dynamic characteristics, 481–484
 - inverter static characteristics, 479–481
 - cutoff and maximum oscillation frequencies, 585–597
 - future technology trends, 616–618
 - heterogeneous ICs, system-on-chip concepts, 636
 - high-speed design issues:
 - channel length and threshold voltage, 552–553
 - deep-submicrometer CMOS transistor drive current, 550–551
 - device design parameters, 551–552
 - gate/active-region sheet resistances, 564–568
 - gate dielectrics, boron suppression, 578–580
 - gate length, power supply voltage and maximum off-state leakage current, 552
 - gate length critical dimension (CD) control, 553–554
 - gate oxide thickness, 554–555
 - gate-to-drain overlap capacitance (C_{GD}), 577–578
 - performance figure of merit (FOM), 548–550
 - poly depletion, 575–577
 - self-aligned silicides, 568–575
 - short-channel effects (SCE), 555–564
 - channel dopant N_{sub} engineering, 557–560
 - junction depth X_j processing techniques, 561–564
 - source/drain resistance (R_{SD}), junction depth X_j and, 560–561
 - large-signal power anad efficiency, 597–607
 - low voltage/low power considerations, 581–585
 - device design, 582–585
 - dual- V_T CMOS, 582–583
 - dynamic threshold voltage CMOS (DTMOS), 584
 - fully or partially depleted SOI CMOS, 585
 - low- V_T CMOS, adjustable substrate bias, 583–584
 - SOI on active substrate (SOIAS), 585
 - noise figure, 607–616
- Digital blocks, embedded modules on IC, system-on-chip concepts, 636–639
- Digital correction, block integration, system-on-chip concepts, analog/digital partitioning, 673–675
- Digital/memory choices, partitioning, block integration, system-on-chip concepts, 672–673
- Digital noise, block integration, system-on-chip concepts, interference, 671–672
- Digital signal processor (DSP), digital blocks, system-on-chip applications, 637–639
- Digital-to-analog (D/A) conversion, system-on-chip concepts:
 - conversion architectures, 655–659
 - terminology, 651–655
- Digital video front-end chip, system-on-chip concepts, 676, 678
- DINOR cell, floating-gate memory arrays, 417–421
- Direct digital synthesis (DDS), block integration, system-on-chip concepts, system function implementation, 667–670
- Direct lattice, band structure, dispersion relationship, 158–160
- Direct tunneling, plasma damage and, 317
- Dirichlet boundary conditions, inner-box shaped transistor:
 - heat flux and thermal diffusion, 43–45
 - one-dimensional heat transfer, 46–49
- Discrete Fourier transform (DFT):
 - band structure, dispersion relationship, 157–160

- Discrete Fourier transform (DFT) (*continued*)
 embedded modules on IC, system-on-chip concepts, 634–635
- Dispersion relationship, band structure, 152–160
- Disturb failures, floating-gate memory arrays, 424–426
- Divided word line (DWL) structure, SRAM (static random access memory) architecture, 368–369
- DMOS transistors, smart power circuits, 259–261
- Dopant control, advanced MOSFET structures, 133–135
- Doping profiles:
 state-of-the-art bulk MOSFET, 132–133
 system-on-chip technology limits, systematic and random mismatch, 646
- Double-balanced mixer design, CMOS technology, radiofrequency (RF) circuits, 511–513
- Double-gated (DG) FET, advanced MOSFET structures, 134–136
- Double heterojunction bipolar transistor (DHBT), narrow-bandgap base, 64–65
- Double-poly transistor, self-adjusted transistor structures, inside spacer, 53–55
- Double solid-phase epitaxy (DSPE) technique, silicon-on-sapphire (SOS) material, 223
- Drain capacitance:
 floating-gate memory arrays, NAND cell, 414–417
 SOI MOSFET, 230
- Drain current models:
 digital switching, CMOS technology, static inverter, 480–481
 first-order MOSFET models:
 one-dimensional drain current model, 75–76
 strong inversion approximation, 79–82
 high-speed digital applications, noise parameters, 609–616
 MOSFET parasitic effects:
 gate-induced drain leakage (GIDL), 119–124
 output resistance, 128–130
 MOSFETs, short-channel effects, 89–97
- Drain engineering:
 hot-carrier structure, 311–312
 MOSFET parasitic effects, hot carriers, 118–119
- Drain extension, byte-alterable E²PROMs, floating-gate tunnel oxide (FLOTOX), 402–406
- Drain-induced barrier lowering (DIBL):
 high-speed digital applications, short-channel effects (SCE), 556
- parasitic effects in MOSFETs, output resistance, 128–130
 short-channel effects in MOSFETs, 89–90
 quasi-two-dimensional analysis, 97–99
- DRAM (dynamic random access memory) circuits:
 cell structures, 333–339
 cell scaling, 335–339
 memory cell concept, 333–335
 chip structure and operation, 346–347
 ferroelectric memory architecture and, 458–463
 destructive readout structure, 460–462
 future trends in, 371–372
 high-speed architectures, 353–358
 extended data out (EDO), 354–355
 fast page mode (FPM), 354
 rambus DRAM (RDRAM), 355–358
 synchronous DRAM (SDRAM), 355
 synchronous link DRAM (SLDRAM), 358
 memory array, 346, 348–351
 MOSFET models, device scaling, 87–88
 operating principle, 340–346
 charge share sensing, 340–341
 refresh operations, 341–344
 capacitor dielectric leakage, 343
 junction leakage, 342
 parasitic leakage paths, 344
 pass transistor subthreshold leakage, 342–343
 soft errors, 344–346
 alpha particles, 344–345
 cosmic rays, 345–346
 operation principle, soft-error requirement, vs. SRAM, 367
 partitioning, block integration, system-on-chip concepts, digital/memory choices, 672–673
 role of, in computer hierarchy, 333–334
 sense amplifier operation, 351–352
 silicon-on-insulator (SOI) devices, 261–262
 technology trends in, 9–10
 word line boosting, 352–353
- Drift-diffusion equations:
 Boltzmann transport equation (BTE), balance equations, 196–200
 MOSFET transport properties, mobility models, 108–109
- Drifted Maxwellian model, hot carriers, MOSFET parasitic effects, energy distribution functions, 116–118
- Drift transistor, inner-box shaped transistor, 29
- Drift velocity, MOSFET transport properties, 109–110

- Drive currents, high-speed digital applications, deep-submicrometer CMOS transistor, 550–551
- Drude free-electron model, band structure:
 - dispersion relationship, 155–160
 - single-electron effect-mass equation, 151–152
- Dual-gate CMOS, high-speed digital applications, poly depletion, 575–577
- Dual-gated-access memory cell, structural characteristics, 444–446
- Dual-gate MOSFET, structural components, 242–248
- Dual- V_T CMOS, low voltage/low power considerations, 582–583
- Dynamic characteristics, CMOS inverter, digital switching, 481–484
- Dynamic range (DR):
 - analog-to-digital (A/D) conversion, 653–655
 - CMOS low-noise amplifiers (LNA), 498–508
- Dynamic threshold MOS transistor (DTMOS):
 - advanced MOSFET structures, 135
 - low voltage/low power considerations, 584
- Dynamic write inhibit, silicon nitride memory, 439–440

- Early effect, inner-box shaped transistor:
 - high-frequency behavior, 37–39
 - static characteristics, 34–35
- EB (emitter-base) junction:
 - bipolar transistors, 20–22
 - inner-box shaped transistor, 25–29
 - high-frequency behavior, 36–39
 - static characteristics, 35
- EEPROM cell:
 - flash EEPROM memories, 406–421
 - DINOR cell, 417–421
 - field-enhanced tunnel injector flash (FETIF) EPROM, 410–412
 - NAND cell, 414–417
 - T-cell flash EPROM, 407–410
 - triple-poly, virtual ground (TPVG) flash cell, 412–414
 - quantum effect devices, 253–254
- Effective channel length (L_{eff}), high-speed digital applications, 552–553
 - SCE-channel dopant N_{sub} engineering, 558–560
- Effective field, MOSFET transport properties, universal mobility, 106–107
- Effective number of bits (ENOB), analog-to-digital (A/D) conversion, 653
- Eigen energies, MOSFET inversion layers, Boltzmann transport equation (BTE) simulations, 204–210

- Einstein relation:
 - Boltzmann transport equation (BTE), 197–200
 - inner-box shaped transistor, temperature dependence, collector current, 40–42
 - one-dimensional drain current MOSFET model, 75–76
- Electromigration, inner-box shaped transistor, safe operation area (SOA), 49–51
- Electron approximation, band structure, single-electron effect-mass equation, 152
- Electron concentration, quantum effect devices, 251–254
- Electron shading effect, plasma damage, 316
- Electron traps:
 - hot-carrier effect (HCE):
 - high gate voltage stresses, 287–288
 - low gate voltage stresses, 285–286
 - p-MOS device, 289–292
 - hot-carrier oxidation, nitridation process, 319
 - lifetime estimation:
 - n-MOS hot-electron-generation, 296–297
 - n-MOS hot-hole-generation, 295–296
- Electrothermal feedback, inner-box shaped transistor, second breakdown thermal instabilities, 42–43
- Elevated source-drains, hot carrier structure, 312
- Embedded mixer design, CMOS radiofrequency mixers, highly integrated transceivers, 515–520
- Embedded modules, system-on-chip concepts, 633–639
 - digital blocks, 636–639
 - functionality, 633–635
 - heterogeneous ICs, 635–636
- Emission mechanisms, silicon nitride memory, 432–435
- Emitter contacts, self-adjusted transistor structures, polysilicon base, 51–55
- Emitter injection efficiency:
 - heterojunction bipolar transistor (HBT), 60–66
 - inner-box shaped transistor, static characteristics, 31–35
- Emitter-to-collector current, bipolar-MOS “hybrid” device, 240–242
- Endurance failures:
 - floating-gate memory, 423–424
 - silicon nitride memory, 438
- Energy-balance equation, Boltzmann transport equation (BTE), 196–200
- Energy distribution functions, MOSFET parasitic effects, hot carriers, 116–118
- Energy position, hot-carrier effect (HCE), oxide traps N_{ot} , 279

- Ensemble Monte Carlo (EMC) simulations:
 - device modeling, 184
 - heterojunction bipolar transistor (HBT), 184–185
- Ensemble relaxation times, Boltzmann transport equation (BTE), 194
- Epitaxial lateral overgrowth (ELO):
 - dual-gate MOSFET, 247–248
 - silicon-on-insulator (SOI) devices, 225
- EPROMs, floating-gate memory arrays:
 - flash EEPROM memories, 406–421
 - DINOR cell, 417–421
 - field-enhanced tunnel injector flash (FETIF) EPROM, 410–412
 - NAND cell, 414–417
 - T-cell flash EPROM, 407–410
 - triple-poly, virtual ground (TPVG) flash cell, 412–414
 - floating-gate memory arrays:
 - byte-alterable E²PROMs, 402–406
 - flash EEPROM memories, 406–421
 - UV EPROMs, 393–402
 - floating-gate memory physics:
 - channel hot-electron (CHE) injection, 382–383
 - substrate injection, 384–385
 - ultraviolet light erase, 390–391
- Equilibrium statistics:
 - device simulation, 163–170
 - Fermi-Dirac statistics as, 165–170
- Equipartition theorem, ULSI simulations, 212–216
- Erase operations:
 - EPROMs:
 - T-cell Flash EEPROM, 407–410
 - ultraviolet light erase, 389–391
 - field-enhanced tunnel injector flash (FETIF) EPROM, 410–412
 - floating-gate memory arrays:
 - DINOR cell, 418–421
 - NAND cell, 414–417
 - silicon nitride memory, write inhibit and repeated erase, 439–441
- Erratic erase, T-cell Flash EEPROM, 409–410
- Error correction circuits (ECC), DRAM and SRAM chips, 372
- Etchback, silicon-on-insulator (SOI) devices, 227–228
- Euler relation, equilibrium statistics, Fermi-Dirac distribution, 166–170
- Excitation mechanisms, floating-gate memory physics, ultraviolet light erase, 389–391
- Extended data out (EDO) signal, high-speed DRAM (dynamic random access memory) circuits, 354–355
- External driving potential, semiclassical electron dynamics, 161–163
- Extra masks:
 - major process adaptations, mixed-signal circuits, system-on-chip concepts, 643
 - minor process adaptations, mixed-signal circuits, system-on-chip concepts, 643
- Face-centered cubic (FCC) lattice, band structure:
 - dispersion relationship, 158–160
 - single-electron effect-mass equation, 150–152
- FAMOS (floating gate avalanche-injection metal oxide semiconductor), floating-gate memory physics, charge transfer, avalanche injection, 379–381
- Fanout (FO) values, high-speed digital applications, performance figure of merit (FOM), 548–550
- Fast page mode (FPM), high-speed DRAM (dynamic random access memory) circuits, 354
- Fatigue degradation, ferroelectric memory, 455–457
- Feature size, mixed-signal circuits, system-on-chip concepts, 641–642
- Fermi-Dirac distribution:
 - Boltzmann transport equation (BTE), 186–192
 - equilibrium statistics, 163–170
 - MOSFET models, device scaling, 86–88
 - MOSFET parasitic effects, gate capacitance degradation, 124–127
 - semiclassical electron dynamics, 149–150
- Fermi golden rule, scattering theory, 171–173
- Fermi potential:
 - quantum effect devices, 252–254
 - SOI MOSFET, threshold voltage, 232–234
 - SOI MOSFETS, 231–232
- Fermi-valence band difference, inner-box shaped transistor, heat flux and thermal diffusion, 45
- Fermi velocity, ULSI simulations, 213–216
- Ferroelectric memory:
 - cell structure, 458–463
 - destructive readout structure, 459–462
 - SRAM-type structure, 462
 - transistor structure, 462–463
 - fatigue and imprint, 455–457
 - hysteresis and retention, 452–453
 - materials and structure, 449–452
 - physics of, 449–458
 - polarization temperature dependence, 454–455
 - radiation hardness, 458

- reliability, 458
- research background, 448–449
- scaling, 458
- switching time and current, 453–454
- Ferroelectric nonvolatile memory, ULSI applications, 7
- “FET ring” topology, CMOS technology, radiofrequency (RF) circuits:
 - double-balanced mixer design, 512–513
 - highly integrated transceivers, 513–520
- Field-dependent mobility, Boltzmann transport equation (BTE), 199–200
- Field-enhanced tunnel injector flash (FETIF) EPROM:
 - disturb failures, 425–426
 - floating-gate memory arrays, 410–412
- First-order MOSFET models, 74–83
 - drain current:
 - one-dimensional model, 75–76
 - strong inversion approximation, 79–82
 - subthreshold drain current, 82–83
 - one-dimensional Poisson equation, charge density, 76–79
- Flash EEPROM memories:
 - endurance failures, 423–424
 - floating-gate memory arrays, 406–421
 - DINOR cell, 417–421
 - field-enhanced tunnel injector flash (FETIF) EPROM, 410–412
 - NAND cell, 414–417
 - T-cell flash EPROM, 407–410
 - triple-poly, virtual ground (TPVG) flash cell, 412–414
- Flicker noise characteristic, voltage-controlled oscillators (VCOs), 524–525
- Floating-gate memory, 378–393
 - arrays:
 - byte-alterable E²PROMs, 402–406
 - flash EEPROM memories, 406–421
 - future trends in, 426–430
 - reliability, 421–426
 - UV EPROMs, 393–402
 - future trends in, 426–430
 - hot-carrier effects, 308–309
 - physics of, 379–393
 - cell sensing, 391–393
 - binary sensing, 391–393
 - multilevel storage, 393
 - charge transfer, 379–391
 - avalanche injection, 379–381
 - channel hot-electron injection, 381–383
 - Fowler-Nordheim tunneling, 385–389
 - hot-electron injection, 379
 - source-side electron injection, 383–384
 - substrate injection, 384–385
 - ultraviolet light erase, 389–391
 - reliability of, 421–426
 - disturb failures, 424–426
 - endurance failures, 423–424
 - retention failures, 421–423
 - vs. silicon nitride memory, 431
- Floating-gate tunnel oxide (FLOTOX), byte-alterable E²PROMs, 402–406
- 1/f noise, high-speed digital applications, noise parameters, 609–616
- Folded bit line configuration, DRAM (dynamic random access memory) circuits, memory array architectures, 348–349
- Folded source-coupled logic (FSCL), CMOS digital switching, 488–489
- Forbidden gap, hot-carrier effect (HCE), interface states N_{it} , 278–279
- Foster filter, inner-box shaped transistor, one-dimensional heat transfer, 47–49
- Fourier transform:
 - band structure, dispersion relationship, 158–160
 - digital applications, large-signal power anad efficiency, 600–607
 - ionized impurity scattering, 174
 - scattering theory, 172–173
- Fowler-Nordheim (FN) tunneling:
 - DRAM (dynamic random access memory) circuits, capacitor dielectric leakage, 343–344
 - floating-gate memory arrays:
 - DINOR cell, 417–421
 - disturb failures, 425–426
 - endurance failures, 423–424
 - retention failure, 422–423
 - scaling trends, 427–430
 - floating-gate memory physics:
 - byte-alterable E²PROMs, 402–406
 - charge transfer, 379, 385–389
 - NAND cell, 414–417
 - high-speed digital applications, gate dielectrics for penetration suppression, 578–580
 - hot-carrier devices, plasma damage, 314–317
 - hot-carrier effect (HCE), oxide traps N_{ot} , 279
 - silicon nitride memory, 431
 - mechanisms of, 432–435
 - T-cell Flash EEPROM, 407–410
- Free-flight time, Monte Carlo simulations, 182–185
- Frenkel-Poole leakage mechanism, DRAM (dynamic random access memory) circuits, capacitor dielectric leakage, 343–344

- Frequency limits, self-adjusted transistor structures, 57–60
- Friis' law, CMOS low-noise amplifiers (LNA), 498–508
- Full-flash converters, system-on-chip concepts, analog-to-digital (A/D) conversion architectures, 660–662
- Fully depleted (FD) SOI CMOS, digital applications, low voltage/low power considerations, 585
- Fully depleted (FD) SOI MOSFETS:
 - body effect, 235
 - characteristics of, 230–232
 - low-voltage, low-power (LVLP) circuits, 256–259
 - output characteristics and transconductance, 235–237
 - subthreshold slope, 237
 - threshold voltage, 233–234
- Functionality, embedded modules on IC, system-on-chip concepts, 633–639
- Gamma functions, Boltzmann transport equation (BTE), 193
- Gate-all-around (GAA) MOSFET, dual-gate MOSFET as, 247–248
- Gate capacitance:
 - degradation, MOSFET parasitic effects, 124–127
 - inversion-layer capacitance, 125–126
 - polysilicon gate depletion, 127
 - high-speed digital applications, performance figure of merit (FOM), 548–550
 - mixed-signal circuits:
 - feature size, system-on-chip concepts, 641–642
 - tolerances, system-on-chip concepts, 643–644
- Gate coupling ratio, floating-gate memory arrays, UV EPROMS, 394
- Gate current, MOSFET parasitic effects:
 - gate-induced drain leakage (GIDL), 123–124
 - hot carriers, 111–112
- Gated-access silicon nitride memory cells, structural characteristics, 444–446
- Gated-diode technique, hot-carrier measurements, 309–310
- Gate dielectrics, high-speed digital applications, boron penetration suppression, 578–580
- Gate-induced drain leakage (GIDL), MOSFET parasitic effects, 119–124
- Gate length:
 - advanced MOSFET structures, 137
 - high-speed digital applications, performance figure of merit (FOM), 552
 - hot carrier structure, 311
- Gate length critical dimension (CD), high-speed digital applications, 552–553
- Gate overlap capacitance, hot-carrier measurements, 310
- Gate sheet resistance, high-speed digital applications, 564–568
- Gate-to-drain overlap capacitance (C_{GD}), high-speed digital applications, 577–578
- Gate tunneling current, MOSFET parasitic effects, 127–128
- Gate voltage:
 - charge-pumping detection, interface states N_{it} , 304–305
 - hot-carrier devices, plasma damage, 314–315
 - hot-carrier effect (HCE), stresses:
 - high voltage stressing, electron trapping, 287–288
 - intermediate stresses, 281–283
 - low stress, 283–287
 - p-MOS systems, 288–294
 - MOSFETs, short-channel effects, 89
 - quantum effect devices, 251–254
- Gaussian distribution:
 - MOSFET transport properties, mobility models, 108
 - system-on-chip technology limits, systematic and random mismatch, 645–646
- Gaussian distribution,:
 - first-order MOSFET models, one-dimensional Poisson equation, charge density, 78–79
- Generation-recombination equation:
 - DRAM (dynamic random access memory) circuits, junction leakage, 342
 - inner-box shaped transistor, static characteristics, 31–35
- Geometric factors, high-speed digital applications, gate sheet resistance, 564–568
- Germanium. *See also* Silicon/germanium compounds
 - heterojunction bipolar transistor (HBT), narrow-bandgap base, 63–65
- Gilbert cell topology, CMOS technology, radiofrequency (RF) circuits:
 - double-balanced mixer design, 512–513
 - highly integrated transceivers, 518–520
- Greek alphabet table, 691
- Green's functions:
 - device simulations, 215–216
 - inner-box shaped transistor, one-dimensional heat transfer, 46–49

- Gummel number, heterojunction bipolar transistor (HBT), 61–66
- Gummel-Poon model, inner-box shaped transistor, static characteristics, 31–35
- Half-pitch array, DRAM (dynamic random access memory) circuits, memory array architectures, 349–350
- Halo implants, high-speed digital applications, SCE-channel dopant N_{sub} engineering, 558
- Hamiltonian equations of motion, band structure, dispersion relationship, 154–160
- Hartley quadrature mixer architecture, CMOS technology, radiofrequency (RF) circuits, highly integrated transceivers, 518–520
- Hawkins expression, self-adjusted transistor structures, device results, 59–60
- Heat flux, inner-box shaped transistor:
 - one-dimensional heat transfer, 46–49
 - thermal diffusion and, 43–45
- Heaviside function, inner-box shaped transistor, one-dimensional heat transfer, 47–49
- Heisenberg uncertainty:
 - band structure, dispersion relationship, 154–160
 - Boltzmann transport equation (BTE), 214–216
 - scattering theory, 172–173
- Helmholtz free energy, equilibrium statistics, Fermi-Dirac distribution, 167–170
- Hemispherical grain (HSG), DRAM (dynamic random access memory) circuits, memory cell scaling, 338–339
- Heterogeneous ICs:
 - digital blocks, system-on-chip applications, 637–639
 - system-on-chip concepts, 635–636
- Heterojunction bipolar transistor (HBT), 60–66
 - developmental history, 4, 19
 - Monte Carlo (EMC) simulations, 184–185
 - narrow-bandgap base, 63–65
 - pseudomorphic SiGe layers, 66
 - SiGe/Si material system, 65–66
 - wide-bandgap emitter, 62–63
- HICUM model, inner-box shaped transistor, static characteristics, 31–35
- High current density operations, self-adjusted structures, 55–57
- High-frequency behavior, inner-box shaped transistor, 35–39
- High gate voltage stresses:
 - hot-carrier effect (HCE), 287–288
 - p-MOS, 292–293
 - lifetime estimation, p-MOS static hot-carrier damage modes, 301–302
- Highly integrated transceivers, CMOS technology:
 - low-noise amplifiers (LNA), 504–508
 - mixer design, 513–520
 - power amplifiers, 540
 - voltage-controlled oscillators (VCOs), 525–535
- High-speed architectures, DRAM (dynamic random access memory) circuits, 353–359
 - extended data out (EDO), 354–355
 - fast page mode (FPM), 354
 - rambus DRAM (RDRAM), 355–358
 - synchronous DRAM (SDRAM), 355
 - synchronous link DRAM (SLDRAM), 358
- High-speed digital applications:
 - channel length and threshold voltage, 552–553
 - deep-submicrometer CMOS transistor drive current, 550–551
 - device design parameters, 551–552
 - gate/active-region sheet resistances, 564–568
 - gate dielectrics, boron suppression, 578–580
 - gate length, power supply voltage and maximum off-state leakage current, 552
 - gate length critical dimension (CD) control, 553–554
 - gate oxide thickness, 554–555
 - gate-to-drain overlap capacitance (C_{GD}), 577–578
 - performance figure of merit (FOM), 548–550
 - poly depletion, 575–577
 - self-aligned silicides, 568–575
 - CoSi₂ salicide, 571
 - raised source/drain and other advanced structures, 571–572
 - TiSi₂ salicide, 568–570
- short-channel effects (SCE), 555–564
 - channel dopant N_{sub} engineering, 557–560
 - junction depth X_j and source/drain resistance (RSD), 560–561
 - junction depth X_j processing techniques, 561–564
 - diffusions, 562–564
 - ion implantation, 562
- High-temperature circuits, silicon-on-insulator (SOI) devices, 254–256
- Hole traps:
 - charge-pumping detection, oxide traps N_{ot} , 304–307
 - hot-carrier effect (HCE):
 - high gate voltage, p-MOS devices, 292–293
 - low gate voltage stresses, 283–284

- lifetime estimation, n-MOS hot-hole-generated electron traps, 295–296
- Hooke's law, phonon scattering, 177–181
- Hot carrier effect (HCE):
 - Boltzmann transport equation (BTE), 200
 - damage identification, 278–281
 - interface states, 278–279
 - oxide traps, 279–280
 - relaxable states, 280–281
 - floating-gate memory physics, charge transfer, 379
 - future research, 322–324
 - lifetime estimation, AC and DC, 294–295
 - n-MOS AC stress lifetimes, 297–299
 - n-MOS static damage modes, 295–297
 - p-MOS static damage modes, 299–302
 - measurement techniques, 302–310
 - charge-pumping technique, 303–308
 - DCIV method, 310
 - floating-gate technique, 308–309
 - gated-diode technique, 309–310
 - gate overlap capacitance, 310
 - MOSFET parasitic effects, 111–119
 - draft engineering, 118–119
 - energy distribution functions, 116–118
 - experimental background, 111–112
 - phenomenological model, 112–116
 - n- and p-channel transistors:
 - heating systems, 276–278
 - research background, 275
 - process dependence, 314–322
 - back-end processing, 321–322
 - oxidation, 317–321
 - plasma damage, 314–317
 - stresses, gate voltage dependence:
 - high voltage stressing, electron trapping, 287–288
 - intermediate stresses, 281–283
 - low stress, 283–287
 - p-MOS systems, 288–294
 - structure dependence, 310–314
 - drain engineering, 311–312
 - length, 311
 - mechanical stress, 313–314
 - oxide thickness, 312–313
- Hot carrier negative-bias temperature instability (HC-NBTI), p-MOS devices, 293
- Hot-electron injection:
 - floating-gate memory, endurance failures, 423–424
 - floating-gate memory physics:
 - charge transfer, 379
 - charge transfer, avalanche injection, 381
- Hybrid matrix, bipolar transistors, 24
- Hydrodynamic modeling:
 - inner-box shaped transistor, static characteristics, 31–35
 - MOSFET simulations, 210–211
- Hydrogen, hot-carrier effects (HCE), back-end processing, 321–322
- Hydrophilic bonding, silicon-on-insulator (SOI) devices, UNIBOND material, 229
- Hysteresis, ferroelectric memory structure, 452–453
- ILD film, hot-carrier effects (HCE), back-end processing, 322
- Impedance transformation, CMOS low-noise amplifiers (LNA), 499–508
- Imprint phenomenon, ferroelectric memory, 455–457
- In-band intermodulation products, CMOS low-noise amplifiers (LNA), 496–508
- Induced-based transistor, ULSI applications, 8
- Inductor losses, voltage-controlled oscillators (VCOs), highly integrated transceivers, 529–535
- Inhibit function, floating-gate memory arrays, UV EPROMS, 394–395
- Inner-box shaped transistor, 25–51
 - high-frequency behavior, 35–39
 - static characteristics, 29–35
 - thermal effects, 39–51
 - heat flux and thermal diffusion, device performance, 43–45
 - one-dimensional heat transfer, analytical approach, 46–49
 - safe operation area (SOA), 49–51
 - second breakdown instabilities, 42–43
 - temperature current, 39–42
- Input-referred intercept point (IIP₃):
 - CMOS low-noise amplifiers (LNA), 497–508
 - CMOS radiofrequency mixers, highly integrated transceivers, 514–520
- Inside spacer, self-adjusted transistor structures, double-poly transistor, 53–55
- In situ annealing, hot-carrier devices, plasma damage, 316
- In situ doping of poly-Si (IDP), self-adjusted transistor structures, double-poly transistor, 54–55
- Instantaneous drain voltage, digital applications, large-signal power anad efficiency, 601–607
- Insulated-gate field-effect transistor (IGFET), evolution of, 221

- Integral nonlinearity (INL), analog-to-digital (A/D) conversion, 653–655
- Integrated circuits, developmental history, 4–6, 19
- Intercept point (IP3), CMOS low-noise amplifiers (LNA), 497–508
- Interconnect capacitance ($C_{\text{interconnect}}$), high-speed digital applications, gate oxide thickness (t_{ox}), 555
- Interface states N_{it} :
 - hot-carrier effect (HCE), 278–279
 - back-end processing, 321–322
 - charge-pumping detection, 303–304
 - floating-gate separation, 308–309
 - gated-diode technique, 309–310
 - lifetime estimation, n-MOS static hot-carrier damage modes, 295
- Interface trap generation, MOSFET parasitic effects, hot carriers, 112, 114
- Interference, block integration, system-on-chip concepts, 670–672
- Intermediate gate voltage stresses:
 - hot-carrier effect (HCE):
 - n-MOS, 281–283
 - p-MOS, 292
 - lifetime estimation, p-MOS static hot-carrier damage modes, 301
- International System of Units (SI units), 687
- Intracollisional field effect (ICFE), Boltzmann transport equation (BTE), 189–192
- Intrinsic conduction, inner-box shaped transistor, second breakdown thermal instabilities, 42–43
- Inversion-layer capacitance:
 - MOSFET inversion layers, 2DEG
 - quantization, Boltzmann transport equation (BTE) simulations, 200–210
 - MOSFET parasitic effects, gate capacitance degradation, 124–126
- Inversion-layer scaling, MOSFET models, device scaling, 86–88
- Inverted active region, bipolar transistors, 20
- Inverted transfer function, voltage-controlled oscillators (VCOs), 523–525
- Inverter delays, high-speed digital applications, gate oxide thickness (t_{ox}), 555
- Invervalley scattering, device simulations, 181–182
- Ion implantation:
 - high-speed digital applications, short-channel effects (SCE), junction depth processing, 562
 - silicon-on-insulator (SOI) devices, UNIBOND material, 228–229
 - system-on-chip technology limits, systematic and random mismatch, 646
- Ionized impurity scattering, device simulation, 173–174
- Ionizing radiation:
 - ferroelectric memory, radiation hardness, 458
 - silicon nitride memory, radiation hardness, 442
- Ishibashi model, ferroelectric memory structure, switching time and current, 453–454
- I-V characteristics:
 - CMOS technology, power amplifiers, 538–540
 - digital applications, large-signal power anad efficiency, 599–607
 - floating-gate memory physics, Fowler-Nordheim tunneling, 388–389
 - floating-gate separation, 308–309
 - hot-carrier effect (HCE):
 - interface states (N_{it}), 279
 - relaxable states (N_{nit}), 280–281
 - MOSFET parasitic effects, hot carriers, 112–113
 - single-gate silicon memory cells, 443–444
 - textured poly E²PROM cell, 405–406
- Jet-vapor-deposited (JVD) nitride, high-speed digital applications, gate dielectrics for penetration suppression, 580
- Junction capacitance, high-speed digital applications, performance figure of merit (FOM), 548–550
- Junction depth:
 - advanced MOSFET structures, 137–138
 - high-speed digital applications:
 - raised source/drain structures, 572–573
 - short-channel effects (SCE), 556
 - processing techniques, 561–563
 - source-drain resistance, 560–561
- Junction leakage, DRAM (dynamic random access memory) circuits, 342
- Kinetic-energy terms:
 - band structure:
 - dispersion relationship, 153–160
 - single-electron effect-mass equation, 151–152
 - Boltzmann transport equation (BTE), 192–193
- “Kink effect,” SOI MOSFETs, output characteristics and transconductance, 235–237
- Kirchhoff’s circuit laws:
 - bipolar transistors, 24–25
 - npn transistors, 20

- Kirchhoff transformation, inner-box shaped transistor, one-dimensional heat transfer, 46–49
- Kirk effect:
 inner-box shaped transistor, second breakdown thermal instabilities, 43
 self-adjusted structures, high current density operations, 55–57
- Knee voltage, CMOS technology, power amplifiers, 537–540
- Lai hole trap-interface, hot-carrier oxidation, 318–319
- Lanthanum strontium cobalt oxide (LSCO), ferroelectric memory, fatigue degradation, 457
- Laplace equation, MOSFETs, short-channel effects, characteristic length theory, 100–102
- Large-signal power anad efficiency, digital applications, 597–607
- Laser recrystallization, silicon-on-insulator (SOI) devices, 223–224
- Lateral diffusion MOS (LDMOS), power amplifiers, highly integrated transceivers, 540–541
- Layer thickness, system-on-chip technology limits, systematic and random mismatch, 646
- LC parallel tanks, voltage-controlled oscillators (VCOs), highly integrated transceivers, 527–535
- Lead zirconate titanate (PZT). *See* PZT films
- Leakage mechanisms:
 digital switching, CMOS technology, dynamic inverter, 482–483
 DRAM (dynamic random access memory) circuits, 341–344
 capacitor dielectric leakage, 343
 junction leakage, 342
 parasitic leakage paths, 344
 pass transistor subthreshold leakage, 342–343
 high-speed digital applications, maximum off-state leakage current, 552
- Least significant bits (LSB):
 analog-to-digital (A/D) conversion, 652–655
 system-on-chip concepts, analog-to-digital (A/D) conversion, 662–664
- Lifetime estimation, hot-carrier devices, 294–302
 gate length effects, 311
 n-MOS static damage modes, 295–297
 n-MOS stress lifetimes, 297–299
 p-MOS static hot-carrier damage modes, 299–302
- Light doped drain (LDD) structure:
 hot carriers, 311–312
 MOSFET parasitic effects, hot carrier reduction, 118–119
 MOSFETs, evolution of, historical development, 130–131
- Lithographic techniques, state-of-the-art bulk MOSFET, 131–133
- Load inductors, CMOS low-noise amplifiers (LNA), highly integrated transceivers, 506–508
- Local-field mobility models, MOSFET transport properties, 108–109
- Local oscillator (LO):
 CMOS technology, radiofrequency (RF) circuit design, mixer fundamentals, 508–513
 CMOS technology, radiofrequency (RF) circuits, highly integrated transceivers, 514–520
- Local oxidation (LOCOS):
 mixed-signal circuits, feature size, system-on-chip concepts, 641–642
 MOSFETs, evolution of, historical development, 130–131
 self-adjusted transistor structures, double-poly transistor, 53–55
- Local threshold variation, system-on-chip technology limits, component matching, 651
- Lombardi mobility model, MOSFET transport properties, 108–109
- Long-channel devices, MOSFETs, short-channel effects, 92–97
- Longitudinal acoustic deformation potential (LADP), phonon scattering, 180–181
- Loop gain, voltage-controlled oscillators (VCOs), 523–525
- Lorentz force:
 scattering theory, 172–173
 semiclassical electron dynamics, 161–163
- Lorenz number, inner-box shaped transistor, heat flux and thermal diffusion, 44–45
- Lower-dielectric-constant insulators, state-of-the-art bulk MOSFET, 132–133
- Low gate voltage stresses:
 hot-carrier effect (HCE):
 electron traps, 285–286
 hole traps, 283–284
 p-MOS devices, 289–292
 relaxable states, 286–287
 lifetime estimation:

- n-MOS hot-hole-generated electron traps, 295–296, 296–297
- p-MOS static hot-carrier damage modes, 299–301
- Low-noise amplifiers (LNA), CMOS technology:
 - fundamentals, 493–508
 - highly integrated transceivers, 504–508
- Low pressure chemical vapor deposition (LPCVD), silicon nitride memory, 435–438
- Low-voltage/low-power (LVLP) circuits:
 - digital applications, 581–585
 - device design, 582–585
 - dual- V_T CMOS, 582–583
 - dynamic threshold voltage CMOS (DTMOS), 584
 - fully or partially depleted SOI CMOS, 585
 - low- V_T CMOS, adjustable substrate bias, 583–584
 - SOI on active substrate (SOIAS), 585
 - silicon-on-insulator (SOI) devices, 256–259
 - system-on-chip technology limits, component matching, 649–650
- Low- V_T CMOS, low voltage/low power considerations, adjustable substrate bias, 583–584
- “Lucky electron” model:
 - floating-gate memory physics:
 - channel hot-electron (CHE) injection, 382–383
 - charge transfer, 379
 - hot carriers, MOSFET parasitic effects, 112–113, 116
 - energy distribution functions, 116–118
- Magnetic breakdown, semiclassical electron dynamics, 214–216
- Manley-Rowe relations, CMOS technology, radiofrequency (RF) circuit design, mixer fundamentals, 509–513
- Masks:
 - block integration, system-on-chip concepts, analog/digital partitioning, 673–675
 - extra masks:
 - major process adaptations, mixed-signal circuits, system-on-chip concepts, 643
 - minor process adaptations, mixed-signal circuits, system-on-chip concepts, 643
 - no extra masks:
 - minor process adaptations, mixed-signal circuits, system-on-chip concepts, 643
 - no process adaptations, mixed-signal circuits, system-on-chip concepts, 642–643
- Master-slave divider architecture, voltage-controlled oscillators (VCOs), highly integrated transceivers, 532–535
- Matching techniques, CMOS low-noise amplifiers (LNA), common gate/source configuration, 499–502
- Matthiessen’s rule:
 - lifetime estimation, n-MOS stress lifetimes, 298–299
 - MOSFET transport properties, 104
 - mobility models, 107–109
- Maximum available gain (MAG), microwave MOSFETs, 237–238
- Maximum collector current, inner-box shaped transistor, safe operation area (SOA), 49–51
- Maximum collector-emitter voltage, inner-box shaped transistor, safe operation area (SOA), 50–51
- Maximum off-state leakage current, high-speed digital applications, performance figure of merit (FOM), 552
- Maximum oscillation frequency:
 - digital applications, 585–597
 - inner-box shaped transistor, high-frequency behavior, 38–39
- Maxwell-Boltzmann functions:
 - equilibrium statistics, Fermi-Dirac distribution, 166–170
 - first-order MOSFET models, one-dimensional Poisson equation, charge density, 77–79
 - hot-carrier heating, 276–278
 - inner-box shaped transistor, temperature dependence, collector current, 41–42
 - MOSFET models, device scaling, 86–88
 - MOSFET parasitic effects, gate capacitance degradation, 124–127
 - MOSFET transport properties, universal mobility, 106–107
 - semiclassical electron dynamics, 149–150
- Maxwellian distribution:
 - Boltzmann transport equation (BTE), 192
 - Monte Carlo simulations, 182–185
- Maxwell’s equations:
 - Boltzmann transport equation (BTE), 197–200
 - ferroelectric memory structure, hysteresis and retention, 452–453
- Mean free path, ULSI simulations, 214–216
- Mean-square output current, high-speed digital applications, noise parameters, 610–616
- Mechanical stress, hot carrier structure, 313–314

- Memory cell structure:
 - DRAM (dynamic random access memory)
 - circuits, 333–336, 346, 348–351
 - scaling techniques, 335–339
 - ferroelectric memory, 458–463
 - destructive readout structure, 459–462
 - SRAM-type structure, 462
 - transistor structure, 462–463
 - floating-gate memory arrays, 393
 - byte-alterable E²PROMs, 402–406
 - flash EEPROM memories, 406–421
 - future trends in, 426–430
 - reliability, 421–426
 - UV EPROMs, 393–402
 - silicon nitride memory, 442–448
 - gated-access structure, 444–446
 - shadow RAM structure, 446–448
 - single-gate structure, 443–444
- Memory circuits, heterogeneous ICs, system-on-chip concepts, 635–636
- Memory window, silicon nitride memory, 431
 - radiation hardness, 442
- MESFET, ULSI applications, 7
- Mesoscopic systems, universal conductance fluctuations, 216
- Metal oxide silicon (MOS), developmental history, 19. *See also* MOSFETs
- Microwave MOSFETs, silicon-on-insulator (SOI) configuration, 237–238
- MICROX, microwave MOSFETs, 237–238
- Miller capacitance, high-speed digital applications:
 - gate-to-drain overlap capacitance (C_{GD}), 577–578
 - noise parameters, 607–616
- Miller opamp, low-voltage, low-power (LVLP) circuits, 256–259
- Minimum channel length (L_{min}), high-speed digital applications, SCE-channel dopant N_{sub} engineering, 558–559
- Minimum detectable signal (MDS) level, CMOS
 - low-noise amplifiers (LNA), 498–508
- Minimum noise factor, CMOS low-noise amplifiers (LNA), 495–508
 - optimum source impedance, 499–508
- Minority carrier concentration:
 - dual-gate MOSFET, 245–248
 - inner-box shaped transistor, 27–30
 - high-frequency behavior, 35–39
 - temperature dependence, 39–42
- Mixed-signal circuits, system-on-chip concepts, 639–644
 - feature size, 641–642
 - process options, 642–643
 - signal swing, 640–641
 - tolerances, 643–644
- Mixer design:
 - CMOS technology, radiofrequency (RF)
 - circuit design, fundamentals, 508–513
 - highly integrated transceivers, 513–520
 - mixer fundamentals, 508–513
 - monolithic mixers, comparison of, 520–521
- MNOS device, structural characteristics of, 431
- Mobile-carrier transport, inner-box shaped transistor, static characteristics, 31–35
- Mobility calculations:
 - inner-box shaped transistor, temperature dependence, collector current, 40–42
 - MOSFET transport properties, universal mobility, 104–107
- Modulated signals, block integration, system-on-chip concepts:
 - conversion of, 668
 - sampling of, 666–667
- Modulation-doped-base hot-electron transistor, ULSI applications, 8–9
- Modulation-doped field-effect transistor (MODFET):
 - clock frequency, 10
 - ULSI applications, 7–8
- Molecular beam epitaxy (MBE), double-heterojunction bipolar transistor (DHBT), narrow-bandgap base, 64–65
- Molybdenum doping, high-speed digital applications, TiSi₂ salicide narrow-line effect, 570–571
- Moment equations, Boltzmann transport equation (BTE), 195–200
- Monte Carlo simulations:
 - device modeling, 182–185
 - HBT simulations, 184–185
 - hot carriers, MOSFET parasitic effects, energy distribution functions, 118
 - inner-box shaped transistor, static characteristics, 31–35
 - MOSFET simulations, 210–211
 - semiclassical electron dynamics, 149–150
 - single-particle models, 182–185
- Moore's law:
 - embedded modules on IC, system-on-chip concepts, 633–639
 - technology trends in, 9–10
- MOSFETs:
 - Boltzmann transport equation (BTE)
 - simulations, 200–211
 - 2DEG quantization in inversion layers, 200–210
 - hydrodynamic studies, 210–211

- CMOS technology, radiofrequency (RF)
 - circuits, highly integrated transceivers, 513–520
- depletion and junction depths, 137–138
- developmental history, 4–9, 73–74
- device scaling, 84–88
 - projections, 87–88
 - scaling theory, 84–87
- double-gate FED family, 136–139
- first-order models, 74–83
 - one-dimensional drain current model, 75–76
 - one-dimensional Poisson equation, charge density, 76–79
 - strong inversion approximation drain current, 79–82
 - subthreshold drain current, 82–83
- gate length, 137
- gate oxide thickness, 137
- hot-carrier heating, 276–278
- hydrodynamic modeling, 210–211
- insulated-gate field-effect transistor (IGFET) precursor, 221
- Monte Carlo simulations, 210–211
- parasitic effects, 110–130
 - band-to-band tunneling, 119–124
 - gate capacitance degradation, 124–127
 - inversion-layer capacitance, 125–126
 - polysilicon gate depletion, 127
 - gate-induced drain leakage (GIDL), 119–124
 - gate tunneling current, 127–128
 - hot carriers, 111–119
 - draft engineering, 118–119
 - energy distribution functions, 116–118
 - experimental background, 111–112
 - phenomenological model, 112–116
 - output resistance, 128–130
- short-channel effects, 88–102
 - device scale length, 99–102
 - threshold voltage, 88–99
 - charge-sharing model, 89–97
 - quasi-two-dimensional analysis, 97–99
- SOI MOSFET, 230–239
 - body effect, 234–235
 - fully and partially depleted and accumulation-mode MOSFETs, 230–232
 - microwave MOSFETs, 2370238
 - output characteristics and transconductance, 235–237
 - source and drain capacitance, 230
 - subthreshold slope, 237
 - threshold voltage, 232–234
- structural evolution, 130–135
 - advanced MOSFET device structures, 133–135
 - historical development, 130–131
 - state-of-the-art bulk MOSFET, 131–133
- transport properties, 102–110
 - high field drift velocity, 109–110
 - mobility, 103–109
 - models, 107–108
 - numerical simulation models, 108–109
 - scattering mechanisms, 103–104
 - universal mobility, 104–107
- ULSI device compared with, 3
- Multichip packaging, partitioning, block integration, system-on-chip concepts, 675–676
- Multidie packaging, partitioning, block integration, system-on-chip concepts, 675–676
- Multilevel storage, floating-gate memory cells, 393
- Multiplicative noise, voltage-controlled oscillators (VCOs), 524–525
- Multithreshold CMOS (MTCMOS), bipolar-MOS “hybrid” devices, 241–242
- NAND cell:
 - CMOS digital switching, current mode logic (CML), 487
 - floating-gate memory arrays, 414–417
- Narrow-bandgap base, heterojunction bipolar transistor (HBT), 63–65
- Narrowband voltage gains, CMOS low-noise amplifiers (LNA), 503–508
- Narrow-line effect, high-speed digital applications, TiSi₂ salicide, 568–571
- n-channel first-order MOSFET model, layout of, 74–75
- Near-carrier noise, voltage-controlled oscillators (VCOs), 524–525
- Near-production device parameters, self-adjusted transistor structures, 58–60
- Negative-bias temperature instability (NBTI), p-MOS devices, 293
- Nitric oxide:
 - high-speed digital applications, gate dielectrics for penetration suppression, 579
 - hot-carrier oxidation, 319
- Nitridation, hot-carrier oxidation, 318–320
- Nitride/oxide (NO) composite film:
 - DRAM (dynamic random access memory) circuits:
 - capacitor dielectric leakage, 343–344
 - memory cell scaling, 337–339

- Nitride/oxide (NO) composite film (*continued*)
 high-speed digital applications, gate dielectrics
 for penetration suppression, 578–580
- Nitrogen, hot-carrier oxidation, pre- and post-
 oxidation introduction, 319–320
- Nitrous oxide:
 high-speed digital applications, gate dielectrics
 for penetration suppression, 579
 hot-carrier oxidation, 319
- NMOS transistors:
 charge-pumping detection, oxide traps N_{ot} ,
 304–307
- CMOS radiofrequency mixers, highly
 integrated transceivers, 515–520
- digital switching:
 current mode logic (CML), 485–487
 differential split-level (DSL) and cascode
 voltage switch logic (CVSL), 489–492
 folded source-coupled logic (FSCL),
 488–489
 static inverter characteristics, 479–481
- digital switching, CMOS technology, dynamic
 inverter, 481–484
- DRAM (dynamic random access memory)
 operation, sense amplifiers, 351–352
- intermediate gate voltage stresses, hot-carrier
 effect (HCE), 281–283
- lifetime estimation:
 static hot-carrier damage modes, 295–297
 stress lifetimes, 297–299
- oxidation, reoxidized nitrided oxides, 318–320
- oxide thickness, 312–313
- plasma damage, 315–317
- poly depletion, 575–577
- voltage-controlled oscillators (VCOs), highly
 integrated transceivers, 531–535
- No extra masks:
 minor process adaptations, mixed-signal
 circuits, system-on-chip concepts, 643
 no process adaptations, mixed-signal circuits,
 system-on-chip concepts, 642–643
- Noise figure (NF), digital applications, 614–616
- Noise floor, CMOS low-noise amplifiers (LNA),
 498–508
- Noise parameters. *See also* Signal-to-noise ratio
 (SNR)
 digital applications, 607–616
 self-adjusted transistor structures, 58–60
 voltage-controlled oscillators (VCOs),
 523–525
 highly integrated transceivers, 527–535
- Nonequilibrium Green functions (NEGF), device
 simulations, 215–216
- Nonlinear thermal diffusion equation, inner-box
 shaped transistor, one-dimensional heat
 transfer, 46–49
- Nonparabolic bandstructure (NPHD), MOSFET
 simulations, 211
- Nonplanar surfaces, floating-gate memory
 physics, Fowler-Nordheim tunneling,
 386–389
- Non-quasi-static (NQS) solution, digital
 applications of power transistors, cutoff
 and maximum oscillation frequencies,
 586–597
- Nonuniform erase, T-cell Flash EEPROM,
 407–410
- Nonvolatile semiconductor memory (NVSM):
 developmental history, 5–9
 ferroelectric memory:
 cell structure, 458–463
 destructive readout structure, 459–462
 SRAM-type structure, 462
 transistor structure, 462–463
 fatigue and imprint, 455–457
 hysteresis and retention, 452–453
 materials and structure, 449–452
 physics of, 449–458
 polarization temperature dependence,
 454–455
 radiation hardness, 458
 reliability, 458
 research background, 448–449
 scaling, 458
 switching time and current, 453–454
- floating-gate memory, 378–393
 physics of, 379–393
 cell sensing, 391–393
 charge transfer, 379–391
- floating-gate memory arrays:
 byte-alterable E²PROMs, 402–406
 flash EEPROM memories, 406–421
 future trends in, 426–430
 reliability, 421–426
 UV EPROMs, 393–402
- impact on semiconductors, 11–12
- power dissipation, 11
- research background, 377–378
- silicon nitride memory:
 cell structure, 442–448
 gated-access structure, 444–446
 shadow RAM structure, 446–448
 single-gate structure, 443–444
 endurance, 438
 nonvolatile memory reliability, 448
 physics of, 432–442
 radiation hardness, 442
 reliability, 440–441

- research background, 430–431
- retention, 435–438
- scaling, 441–442
- tunneling and emission mechanisms, 432–435
- write inhibit and repeated erase, 438–440
- Nonvolatile SRAM (nvSRAM), shadow RAM
 - silicon nitride memory cells, 446–448
- NPN transistor:
 - bipolar-MOS “hybrid” device, 240–242
 - bipolar transistors, 250
 - operating regions, 22
 - structure, symbols and nomenclature, 20–21
- Numerical simulation, MOSFET transport
 - properties, mobility models, 108–109
- Nyquist frequency:
 - band structure, dispersion relationship, 157–160
 - system-on-chip concepts, D/A converters, 658–659
- Offset storage, system-on-chip concepts, analog-to-digital (A/D) conversion, 661–662
- One-chip oscilloscope, system-on-chip concepts, 679–680
- One-chip television chip, system-on-chip concepts, 676–677
- One-dimensional drain current model, first-order MOSFET models, 75–76
- One-dimensional heat transfer, inner-box shaped transistor, analytical approach, 46–49
- One-dimensional Poisson equation, first-order MOSFET models, charge density, 76–79
- One-dimensional potential well, band structure, dispersion relationship, 155–160
- “One-time” trimming, system-on-chip concepts, analog-to-digital (A/D) conversion, 661
- One transistor/one ferroelectric capacitor (1T/1C) architecture, ferroelectric memory, destructive readout structure, 459–462
- Open bit line configuration, DRAM (dynamic random access memory) circuits, memory array architectures, 348
- Optical deformation potential (ODP), phonon scattering, 180–181
- Output characteristics:
 - inner-box shaped transistor, static characteristics, 33–35
 - SOI MOSFET, 235–237
- Output noise current, high-speed digital applications, noise parameters, 609–616
- Output-referred intercept point (OIP₃), CMOS low-noise amplifiers (LNA), 497–508
- Output resistance, MOSFET parasitic effects, 128–130
- Oversampling:
 - block integration, system-on-chip concepts, D/A converter, 668–670
 - system-on-chip concepts, D/A converters, 658–659
- Oxidation:
 - floating-gate memory physics, Fowler-Nordheim tunneling, 387–389
 - hot-carrier process and, 317–321
 - nitridation during, 318–319
 - pre- and postoxidation of nitrogen, 319–321
- Oxide/nitride/oxide (ONO) film:
 - DRAM (dynamic random access memory) circuits, memory cell scaling, 337–339
 - floating-gate memory arrays:
 - NAND cell, 414–417
 - scaling trends, 428–430
- Oxide thickness (t_{ox}):
 - advanced MOSFET structures, 137
 - high-speed digital applications:
 - gate thickness, 554–555
 - short-channel effects (SCE), 556
 - hot carrier structure, 312–313, 323–324
 - silicon nitride memory, tunneling-emission mechanisms, 433–435
- Oxide traps (N_{ot}):
 - charge-pumping detection, 304–307
 - damage characteristics, 293–294
 - hot-carrier effect (HCE), 279–280
 - low gate voltage stresses, 285–286
- Pair transition rate, carrier-carrier scattering, 176–177
- Parabolic-cylindrical function representation, digital applications of power transistors, cutoff and maximum oscillation frequencies, 586–597
- Parallel linear networks, CMOS low-noise amplifiers (LNA), 499–500
- Parasitic effects:
 - DRAM (dynamic random access memory) circuits, 344
 - MOSFETs, 110–130
 - band-to-band tunneling, 119–124
 - gate capacitance degradation, 124–127
 - inversion-layer capacitance, 125–126
 - polysilicon gate depletion, 127
 - gate-induced drain, 119–124
 - gate tunneling current, 127–128
 - hot carriers, 111–119
 - draft engineering, 118–119
 - energy distribution functions, 116–118

- Parasitic effects, MOSFETs (*continued*)
 experimental background, 111–112
 phenomenological model, 112–116
 output resistance, 128–130
- Partially-depleted (PD) SOI:
 advanced MOSFET structures, 133–135
 body effect, 235
 characteristics of, 230–232
 digital applications, low voltage/low power considerations, 585
 output characteristics and transconductance, 235–237
 threshold voltage, 233–234
- Particle conservation, inner-box shaped transistor, static characteristics, 30–35
- Partitioning, block integration, system-on-chip concepts, 672–676
 analog/digital partitioning, 673–675
 digital/memory choices, 672–673
 multichip packaging, 675–676
- Passive ring topology, CMOS radiofrequency mixers, highly integrated transceivers, 515–520
- Pass transistor subthreshold leakage, DRAM (dynamic random access memory) circuits, 342–343
- Pattern-constrained epitaxy (PACE), silicon-on-insulator (SOI) devices, epitaxial lateral overgrowth (ELO) variations, 225
- Patterning deviations, system-on-chip technology limits, systematic and random mismatch, 646
- Pauli exclusion principle:
 band structure, single-electron effect-mass equation, 150–152
 equilibrium statistics, Fermi-Dirac distribution, 166–170
- Peltier effect, inner-box shaped transistor, heat flux and thermal diffusion, 45
- Performance figure of merit (FOM):
 digital applications of power transistors, cutoff and maximum oscillation frequencies, 589–597
 high-speed digital applications, 548–550
 device design, 551–552
 gate length critical dimension (CD), 552–553
 gate oxide thickness (t_{ox}), 554–555
 gate sheet resistance, 564–568
 low voltage/low power considerations, device design, 582
- Periodic crystal potential, semiclassical electron dynamics, 161–163
- Periodic lattice potential, band structure:
 dispersion relationship, 157–160
 single-electron effect-mass equation, 151–152
- Perovskite ferroelectric materials, ferroelectric memory, 451–452
- Phase-locked loop (PLL) structures:
 CMOS technology, voltage-controlled oscillators (VCOs), 520, 522–525
 voltage-controlled oscillators (VCOs), highly integrated transceivers, 531–535
- Phase-modulated (PM) signals, voltage-controlled oscillators (VCOs), 524–525
- Phenomenological models, hot carriers, MOSFET parasitic effects, 112–113, 116
- Phonons, scattering theory, 172–173
 equations and calculations, 177–181
- Phosphorus compounds, high-speed digital applications:
 SCE-channel dopant N_{sub} engineering, 557–560
 SCE junction depth processing, 563
- Photon scattering, MOSFET transport properties, universal mobility, 106–107
- Physical constants, 693
- Pipelined architecture, system-on-chip concepts, analog-to-digital (A/D) conversion, 660–662
- Planar capacitor DRAM cell, structure of, 335
- Plasma-assisted chemical etching (PACE), silicon-on-insulator (SOI) devices, bonding and etchback, 228
- Plasma charging mechanism, hot-carrier devices, 314–317
- PMOS transistors:
 CMOS low-noise amplifiers (LNA), highly integrated transceivers, 507–508
 CMOS radiofrequency mixers, highly integrated transceivers, 515–520
 digital switching:
 CMOS technology:
 dynamic inverter, 481–484
 static inverter, 479–481
 current mode logic (CML), 485–487
 differential split-level (DSL) and cascode voltage switch logic (CVSL), 489–492
- DRAM (dynamic random access memory) operation, sense amplifiers, 351–352
 gate voltage dependence of stress, 288–294
 high gate range, 292–293
 intermediate range, 292
 low gate range, 289–292
- hot carriers, light doped drain (LDD) structure, 312

- lifetime estimation, static hot-carrier damage modes, 299–302
- oxide thickness, 312–313
- plasma damage, 315–317
- poly depletion, 575–577
- voltage-controlled oscillators (VCOs), highly integrated transceivers, 531–535
- PNP transistor:
 - bipolar-MOS “hybrid” device, 240–242
 - structure, symbols and nomenclature, 20–21
- Pocket implants, high-speed digital applications, SCE-channel dopant N_{sub} engineering, 558
- Poisson equation:
 - Boltzmann transport equation (BTE), 189–192
 - device scaling, MOSFET models, 84–88
 - first-order MOSFET models:
 - one-dimensional Poisson equation, 76–79
 - strong inversion approximation, 79–82
 - inner-box shaped transistor, static characteristics, 30–35
 - inversion layers, Boltzmann transport equation (BTE) simulations, 203–210
 - Monte Carlo simulations, 182–185
 - semiclassical electron dynamics, 149–150
 - short-channel effects, MOSFET models:
 - characteristic length theory, 100–102
 - quasi-two-dimensional analysis, 97–99
 - SOI MOSFETS, threshold voltage, 233–234
 - system-on-chip technology limits, component matching, 648–651
- Polarization, ferroelectric memory, temperature dependence, 454–455
- Poly depletion, high-speed digital applications, 575–577
- Polysilicon base:
 - floating-gate memory physics, Fowler-Nordheim tunneling, 387–389
 - self-adjusted transistor structures:
 - double-poly transistor with inside spacer, 53–55
 - emitter contacts and, 51–55
- Polysilicon emitter, inner-box shaped transistor, 25–29
- Polysilicon films, silicon-on-insulator (SOI) devices, recrystallization, 223–224
- Polysilicon gate depletion, MOSFET parasitic effects, gate capacitance degradation, 124, 127
- Poole-Frenkel (PF) emission, silicon nitride memory, 432–435
- Post-oxidation techniques, hot-carrier oxidation, 317
- Poststress interface trap creation, hot-carrier effects (HCE), back-end processing, 322
- Potential energy, band structure, dispersion relationship, 153–160
- Power-added efficiency (PAE):
 - CMOS technology, power amplifiers, 536–540
 - double heterojunction bipolar transistor (DHBT), narrow-bandgap base, 65
- Power amplifiers, CMOS technology, 535–540
 - fundamentals, 535–540
 - highly integrated transceivers, 540
- Power flux density, inner-box shaped transistor, heat flux and thermal diffusion, 44–45
- Power limits, system-on-chip technology limits, component matching, 650–651
- Power-series self-consistency technique, digital applications of power transistors, cutoff and maximum oscillation frequencies, 586–597
- Power supply noise rejection:
 - block integration, system-on-chip concepts, interference, 671–672
 - low voltage/low power considerations, dual- V_T CMOS, 582–583
 - voltage-controlled oscillators (VCOs), highly integrated transceivers, 527–535
- Power supply voltage (V_{DD}):
 - high-speed digital applications:
 - performance figure of merit (FOM), 552
 - short-channel effects (SCE), 556
 - mixed-signal circuits, system-on-chip concepts, 640–641
 - system-on-chip technology limits, component matching, 649–650
- Preamorphization implant (PAI), high-speed digital applications, TiSi_2 salicide narrow-line effect, 569–571
- Pre-oxidation techniques, hot-carrier oxidation, 317
- Process options, mixed-signal circuits, system-on-chip concepts, 642–643
- Programmable read-only memory (PROM), nonvolatile semiconductor memory (NVSM), 377
- Propagation delay, high-speed digital applications, performance figure of merit (FOM), 548–550
- Proximity effect, system-on-chip technology limits, systematic and random mismatch, 646
- Pseudomorphic layers, heterojunction bipolar transistor (HBT):
 - Si/Ge layers, 66

- Pseudomorphic layers, heterojunction bipolar transistor (HBT) (*continued*)
silicon/germanium/silicon (SiGe/Si) material system, 65–66
- Pullup/pulldown time, high-speed digital applications, performance figure of merit (FOM), 548–550
- Pumping current, charge-pumping detection, interface states (N_{it}), 304
- Punchthrough, inner-box shaped transistor, 27–29
- PZT films:
DRAM (dynamic random access memory) circuits, memory cell scaling, 337–339
ferroelectric memory:
destructive readout structure, 460–462
fatigue degradation, 457
scaling trends, 458
structure, 452
- Quadrature-cancellation technique:
CMOS technology, radiofrequency (RF) circuits, highly integrated transceivers, 518–520
voltage-controlled oscillators (VCOs), highly integrated transceivers, 526–535
- Quadrature outputs, voltage-controlled oscillators (VCOs), highly integrated transceivers, 529–535
- Quantization effects:
analog-to-digital (A/D) conversion, signal-to-noise ratio (SNR), 652–655
MOSFET simulations, 211–212
system-on-chip concepts, D/A converters, 658–659
- Quantum-effect devices, silicon-on-insulator (SOI) construction, 250–254
- Quantum mechanics:
Boltzmann transport equation (BTE), 214–216
MOSFET Monte Carlo simulations, 211, 213
MOSFET parasitic effects, gate capacitance degradation, 124–127
semiclassical electron dynamics, 149–150
- Quantum transport technology, device simulations, 215–216
- Quantum-wire MOSFET, structural components, 254
- Quarter-pitch array architecture, DRAM (dynamic random access memory) circuits, memory array architectures, 349–350
- Quasi-Fermi potentials:
equilibrium statistics, Fermi-Dirac distribution, 169–170
hot-carrier measurements, gated-diode technique, 309–310
inner-box shaped transistor, heat flux and thermal diffusion, 43–45
- Quasimomentum, semiclassical electron dynamics, 161–163
- Quasistatic (QS) solution, digital applications of power transistors, cutoff and maximum oscillation frequencies, 586–597
- Quasi-two-dimensional analysis, MOSFETs, short-channel effects, 97–99
- Quiescent power dissipation, CMOS technology, power amplifiers, 537–540
- Radiation hardness:
ferroelectric memory, 458
silicon nitride memory, 442
- Radiofrequency (RF) circuit design:
CMOS technology:
integration levels, 477–479
low-noise amplifiers, 493–508
highly integrated transceivers, 504–508
mixers and frequency translation, 508–520
highly integrated transceivers, 513–520
mixer fundamentals, 508–513
power amplifiers, 536–540
digital applications:
cutoff and maximum oscillation frequencies, 585–597
large-signal power and efficiency, 601–607
inner-box shaped transistor, static characteristics, 31–35
- Raised source/drain structures (R/SD), high-speed digital applications, salicides, 571–575
- Rambus DRAM (RDRAM), high-speed DRAM (dynamic random access memory) circuits, 355–358
- Random fluctuation, advanced MOSFET structures, 138–139
- Random mismatch, system-on-chip technology limits, 645–646
- Random offset, system-on-chip concepts, analog-to-digital (A/D) conversion, 661
- Random phase approximation (RPA), Boltzmann transport equation (BTE), 214–216
- Random scattering potential, semiclassical electron dynamics, 161–163
- RC (resistance \times capacitance) delays, minimization of, 11
- Reactance transformation techniques, CMOS low-noise amplifiers (LNA), 499–508
- Reactive ion etching (RIE), self-adjusted

- transistor structures, double-poly transistor, 53–55
- Read only memory (ROM), nonvolatile semiconductor memory (NVSM), 377
- Reciprocal lattice, band structure, dispersion relationship, 158–160
- Relaxable damage, hot-carrier effect (HCE), high gate voltage, p-MOS devices, 292
- Relaxable states (N_{niot}), hot-carrier effect (HCE), 280–281
 - high gate voltage stresses, 288
 - low gate voltage stresses, 286–287
 - p-MOS device, 292
- Relaxation oscillator design, voltage-controlled oscillators (VCOs), highly integrated transceivers, 526–535
- Relaxation-time approximation (RTA):
 - band structure, single-electron effect-mass equation, 151–152
 - Boltzmann transport equation (BTE), 191–192
 - device simulation, 174–176
- Reliability hazards:
 - ferroelectric memory, 458
 - floating-gate memory, 421–426
 - disturb failures, 424–426
 - endurance failures, 423–424
 - retention failures, 421–423
 - silicon nitride memory, 440–441
 - nonvolatile memory, 448
 - single-gate silicon memory cells, 443–444
- Remote-plasma nitrided oxide (RPNO)
 - technique, high-speed digital applications, gate dielectrics for penetration suppression, 580
- Reoxidized nitrided oxides (ROXNOXs, RNOs), hot-carrier oxidation, 317–321
- Resist damage:
 - CMOS low-noise amplifiers (LNA), signal-to-noise ratio (SNR), 494–508
 - plasma damage, 316
- Resistivity, system-on-chip technology limits, systematic and random mismatch, 646
- Resistor, system-on-chip concepts, D/A converters, 655–656
- Resonant tunneling diode, ULSI applications, 7
- Retention failure:
 - ferroelectric memory structure, 452–453
 - floating-gate memory, 421–423
 - silicon nitride memory, 435–438
 - nonvolatile, 448
 - reliability hazards, 440–441
- Retrograde (SSR) doping, high-speed digital applications, SCE-channel dopant N_{sub} engineering, 557–560
- Reuse technology, CMOS low-noise amplifiers (LNA), highly integrated transceivers, 506–508
- Reversed-biased junctions:
 - DRAM (dynamic random access memory) circuits, junction leakage, 342
 - floating-gate memory physics, charge transfer, avalanche injection, 379–381
- Richardson constant, DRAM (dynamic random access memory) circuits, capacitor dielectric leakage, 343–344
- Richardson-Schottky equation, floating-gate memory, retention failure, 422–423
- Ring oscillators:
 - CMOS digital switching, differential split-level (DSL) and cascode voltage switch logic (CVSL), 490–492
 - CMOS technology, radiofrequency (RF) circuits, double-balanced mixer design, 512–513
 - voltage-controlled oscillators (VCOs), highly integrated transceivers, 525–535
- Row address strobe (RAS) signal:
 - DRAM (dynamic random access memory) circuits, 346–348
 - high-speed DRAM (dynamic random access memory) circuits:
 - fast page mode (FPM), 354
 - synchronous DRAM (SDRAM), 355–356
- Rugged polysilicon, DRAM (dynamic random access memory) circuits, memory cell scaling, 338–339
- Safe operation area (SOA), inner-box shaped transistor, thermal effects, 49–51
- Salicides, high-speed digital applications, 568–575
 - CoSi₂ salicide, 571
 - raised source/drain and other advanced structures, 571–572
 - TiSi₂ salicide, 568–571
- Sampling rate, block integration, system-on-chip concepts:
 - accuracy and, 665–666
 - bandwidth and signal processing, 665
 - modulated signals, 666–667
- Saturation region, bipolar transistors, 20
- Scaling trends:
 - DRAM (dynamic random access memory) circuits, 335–339
 - ferroelectric memory, 458
 - floating-gate memory array, 426–430
 - MOSFETs, 84–88
 - projections, 87–88

- Scaling trends, MOSFETs (*continued*)
 - scaling theory, 84–87
 - silicon nitride memory, 441–442
- Scattering mechanisms, MOSFET transport properties, 103–104
- Scattering theory, device simulation, 170–182
 - carrier-carrier scattering, 176–177
 - Conwell-Weisskopf model, 176
 - intervalley scattering, 181–182
 - ionized impurity scattering, 173–174
 - phonon scattering, 177–181
 - relaxation time averages, 174–176
- Schottky approximation:
 - DRAM (dynamic random access memory)
 - circuits, capacitor dielectric leakage, 343–344
 - self-adjusted structures, high current density operations, 56–57
- Schrödinger equation:
 - band structure:
 - dispersion relationship, 152–160
 - single-electron effect-mass equation, 150–152
 - MOSFET inversion layers, 2DEG
 - quantization, Boltzmann transport equation (BTE) simulations, 201–210
 - MOSFET transport properties, universal mobility, 106–107
 - scattering theory, 171–173
- Screening techniques, band structure, single-electron effect-mass equation, 152
- Second breakdown instabilities, inner-box shaped transistor, thermal effects, 42–43
- Second-quantized operators, phonon scattering, 179
- Selective collector implantation (SCI), self-adjusted transistor structures, double-poly transistor, 54–55
- Selective epitaxial silicon growth (SEG), high-speed digital applications, raised source/drain structures, 572–575
- Selective scaling, MOSFET models, device scaling, 86–88
- Self-adjusted transistor structures, 51–60
 - device and circuit results, 57–60
 - high current density operations, 55–57
 - polysilicon base and emitter contacts, 51–55
 - double-poly transistor with inside spacer, 53–55
- Self-aligned silicides (salicides), high-speed digital applications, 568–575
 - CoSi₂ salicide, 571
 - raised source/drain and other advanced structures, 571–572
 - TiSi₂ salicide, 568–570
- Self-scattering, Monte Carlo simulations, 182–185
- Semiclassical electron dynamics:
 - band structure, single-electron effect-mass equation, 152
 - device simulation, 160–163
 - limits of, 214–216
 - principles of, 149–150
- Semiconductor industry:
 - composition of, 1–3
 - sales volume and GNP, 1–2
 - technology trends, 9–13
- Sense amplifiers:
 - DRAM (dynamic random access memory)
 - operation, 351–352
 - charge share sensing, 340–341
 - SRAM (static random access memory)
 - architecture, 368
- Sense operation, SRAM (static random access memory) circuits, 362
- Sensing mechanisms, floating-gate memory cells, 391–393
 - binary sensing, 391–393
 - multilevel storage, 393
- Sensitivity, CMOS low-noise amplifiers (LNA), 498–508
- Series linear networks, CMOS low-noise amplifiers (LNA), 499–500
- Series resistance correction, inner-box shaped transistor, static characteristics, 33–35
- Shadow RAM architecture:
 - silicon nitride memory cells, 446–448
 - SRAM-type ferroelectric memory cell, 462
- Shallow-trench isolation, state-of-the-art bulk MOSFET, 132–133
- Shockley-Read-Hall (SRH) combination theory:
 - Boltzmann transport equation (BTE), 196–200
 - charge-pumping detection, damage
 - localization, 308
- Short-channel effects (SCE):
 - high-speed digital applications, 555–564
 - channel dopant N_{sub} engineering, 557–560
 - junction depth X_j processing techniques, 561–564
 - diffusions, 562–564
 - ion implantation, 562
 - source/drain resistance (RSD), junction depth and, 560–561
- MOSFETs, 88–102
 - device scale length, 99–102
 - threshold voltage, 88–99
 - charge-sharing model, 89–97
 - quasi-two-dimensional analysis, 97–99

- Shot noise, high-speed digital applications, noise parameters, 609–616
- SIA *Roadmap*:
 embedded modules on IC, system-on-chip concepts, 633–639
 high-speed digital applications:
 gate oxide thickness (t_{ox}), 554–555
 short-channel effects (SCE), source-drain resistance (R_{SD}), 560–561
 low voltage/low power considerations, 581–585
- Signal processing strategy, block integration, system-on-chip concepts, 665–667
 modulated signal sampling, 666–667
 sample rate and accuracy, 665–666
 sampling rate and bandwidth, 665
- Signal swing, mixed-signal circuits, system-on-chip concepts, 640–641
- Signal-to-matching coefficient ratio, system-on-chip technology limits, component matching, 650
- Signal-to-noise-and-distortion (SINAD), analog-to-digital (A/D) conversion, 653–655
- Signal-to-noise ratio (SNR):
 analog-to-digital (A/D) conversion, 652–655
 CMOS low-noise amplifiers (LNA), fundamentals of, 493–508
 digital applications, noise figure (NF), 614–616
 mixer fundamentals, radiofrequency (RF) CMOS design, 509–513
 system-on-chip concepts, D/A converters, 658–659
- Silicon, properties at 300 K, 695
- Silicon-germanium (SiGe) alloys:
 heterojunction bipolar transistor (HBT), narrow-bandgap base, 63–65
 high-speed digital applications, poly depletion, 576–577
- Silicon/germanium/silicon (SiGe/Si):
 heterojunction bipolar transistor (HBT), 65–66
 heterojunctions, developmental history, 19
- Silicon nitride memory:
 cell structure, 442–448
 gated-access structure, 444–446
 shadow RAM structure, 446–448
 single-gate structure, 443–444
 endurance, 438
 nonvolatile memory reliability, 448
 physics of, 432–442
 radiation hardness, 442
 reliability, 440–441
 research background, 430–431
 retention, 435–438
 scaling, 441–442
 tunneling and emission mechanisms, 432–435
 write inhibit and repeated erase, 438–440
- Silicon-on-insulator (SOI) devices:
 advanced MOSFET structures, 133–135
 bipolar-MOS “hybrid” devices, 239–242
 bipolar transistors, 248–250
 circuit structures:
 high-temperature circuits, 254–256
 low-voltage low-power (LVLP) circuits, 256–259
 smart power circuits, 259–261
 SRAMs and DRAMs, 261–262
 dual-gate MOSFET, 242–248
 plasma damage, 316–317
 quantum-effect devices, 250–254
 SOI MOSFET, 230–239
 body effect, 234–235
 fully and partially depleted and accumulation-mode MOSFETs, 230–232
 microwave MOSFETs, 237–238
 output characteristics and transconductance, 235–237
 source and drain capacitance, 230
 subthreshold slope, 237
 threshold voltage, 232–234
- substrates:
 epitaxial lateral overgrowth, 225
 laser recrystallization, 223–224
 silicon-on-sapphire material, 222–223
 SIMOX (separation by implanted oxygen), 225–227
 UNIBOND material, 228–230
 wafer bonding and etchback, 227–228
 zone-melting recrystallization, 224–225
- Silicon-on-sapphire (SOS) material, silicon-on-insulator (SOI) devices, 222–223
- SIMOX (separation by implanted oxygen):
 microwave MOSFETs, 237–238
 quantum effect devices, 253–254
 silicon-on-insulator (SOI) devices, 225–227
 smart power circuits, 259–261
- Simple harmonic oscillator (SHO), phonon scattering, 178–181
- Single comparator architecture, system-on-chip concepts, analog-to-digital (A/D) conversion, 661–662
- Single crystalline materials, heterojunction bipolar transistor (HBT), 62–63
- Single-electron effect-mass equation, band structure, 150–152
- Single-electron memory cell (SEMC), developmental history, 6–7

- Single-electron transfer (SET), quantum effect devices, 253–254
- Single-ended mixer design, CMOS technology, radiofrequency (RF) circuit design, 509–513
- Single-ended power amplifiers, CMOS technology, 536–537
- Single-gate silicon nitride memory cells, structural characteristics, 443–444
- Single-stage differential low-noise amplifier, highly integrated transceivers, 505–508
- “ $\sin(x)x$ ” effect, analog-to-digital (A/D) conversion, 654
- Slow hole traps, hot-carrier effect (HCE), low gate voltage stresses, 287
- Small-signal conductance, drain current, strong inversion approximation, first-order MOSFET models, 81–82
- Smart-Cut process, silicon-on-insulator (SOI) devices, UNIBOND material, 228–230
- Smart power circuits, silicon-on-insulator (SOI) devices, 259–261
- SMI process, self-adjusted transistor structures: device results, 59–60
- double-poly transistor, 54–55
- SNOS device, structural characteristics of, 431
- Soft errors:
 - DRAM (dynamic random access memory) circuits, 344–346
 - alpha particles, 344–345
 - cosmic rays, 345–346
 - SRAM (static random access memory) circuits, operation principle, 365–367
 - alpha particles, 365–366
 - cosmic rays, 366–367
 - vs. DRAM, 367
- SOI on active substrate (SOIAS), digital applications, low voltage/low power considerations, 585
- Solid-phase epitaxy and regrowth (SPEAR), silicon-on-sapphire (SOS) material, 223
- SONOS device, structural characteristics of, 430–431
- Source capacitance:
 - floating-gate memory arrays, NAND cell, 414–417
 - SOI MOSFET, 230
- Source-coupled logic (SCL), CMOS digital switching, folded source-coupled logic (FSCL), 488–489
- Source current, MOSFET parasitic effects, hot carriers, 111–112
- Source-drain resistance (R_{SD}), high-speed digital applications:
 - gate/active region sheet resistances, 566–568
 - short-channel effects (SCE), 560–561
- Source lines, field-enhanced tunnel injector flash (FETIF) EPROM, 412
- Source resistance, CMOS low-noise amplifiers (LNA), signal-to-noise ratio (SNR), 495–508
- Source-side electron injection, floating-gate memory physics, charge transfer, 383–384
- Spectral density, high-speed digital applications, output noise current, 610–616
- Speed optimization, mixed-signal circuits, tolerances, system-on-chip concepts, 643–644
- SPICE (simulation program for IC emphasis): digital applications of power transistors, cutoff and maximum oscillation frequencies, 591–597
 - inner-box shaped transistor, static characteristics, 31–35
- Split-gate cell, floating-gate memory arrays, UV EPROMS, 399
- Spurious free dynamic range (SFDR), CMOS low-noise amplifiers (LNA), 496–508
- SRAM (static random access memory) circuits:
 - address transition detection circuit, 369–371
 - basic SRAM structure, 367
 - divided word line (DWL) structure, 368–369
 - ferroelectric memory architecture and, 458–463
 - shadow-ram architectures, 462
 - future trends in, 372
 - memory cell structure, 358, 360–363
 - depletion-load SRAM, 360–362
 - full-CMOS load, 360–362
 - resistor-load (R-load) cell, 360–362
 - TFT-load type, 360–362
 - operation principle:
 - cell stability analysis, 364–365
 - sense operation, 362
 - soft-error requirement, 365–367
 - alpha particles, 365–366
 - cosmic rays, 366–367
 - vs. DRAM, 367
 - sense amplifier, 368
 - shadow RAM silicon nitride memory cells, 446–448
 - silicon-on-insulator (SOI) devices, 261–262
- Stack capacitor DRAM cell:
 - memory cell scaling, 337–338
 - structure of, 335

- Stacked-gate cell, floating-gate memory arrays, UV EPROMS, 398–399
- Staggered virtual ground (SVG):
 - floating-gate memory arrays, UV EPROMS, 399–402
 - triple-poly, virtual ground (TPVG) flash cell, 413–414
- “Staircase” density of states, equilibrium statistics, 165
- Standard deviation, system-on-chip technology limits, component matching, 648–651
- State effective mass densities, inner-box shaped transistor, temperature dependence, collector current, 41–42
- State-of-the-art bulk MOSFET, structural evolution, 131–133
- Static characteristics:
 - CMOS inverter, digital switching, 479–481
 - inner-box shaped transistor, 29–35
 - safe operation area (SOA), 50–51
- Static-noise margin (SNM), SRAM (static random access memory) circuits, cell stability analysis, 364–365
- Static scattering, device simulation, 172–173
- Static write inhibit, silicon nitride memory, 438–440
- Stepped-gate silicon oxide memory cell, structural characteristics, 443–444
- Stresses, gate voltage dependence, hot-carrier effect (HCE):
 - high voltage stressing, electron trapping, 287–288
 - intermediate stresses, 281–283
 - low stress, 283–287
 - p-MOS systems, 288–294
- Stress-induced leakage current (SILC):
 - floating gate memory arrays:
 - disturb failures, 426
 - endurance failures, 423–424
 - floating-gate memory arrays, scaling trends, 428–430
- Strong inversion approximation, drain current, first-order MOSFET models, 79–82
- Strontium bismuth tantalate (SBT), ferroelectric memory:
 - destructive readout structure, 461–462
 - fatigue degradation, 457
 - scaling trends, 458
 - structure, 451–452
- Structure dependence, hot-carrier effect (HCE), 310–314
 - drain engineering, 311–312
 - length, 311
 - mechanical stress, 313–314
 - oxide thickness, 312–313
- Subband energy levels, quantum effect devices, 252–254
- Subgrain boundaries, silicon-on-insulator (SOI) devices, zone-melting recrystallization (ZMR), 224–225
- Substrate bias technique:
 - charge-pumping detection, damage localization, 308
 - low voltage/low power considerations, low- V_T CMOS, 583–584
- Substrate current, MOSFET parasitic effects, hot carriers, 111–114
- Substrate-current-induced body effect (SCBD), MOSFET parasitic effects, output resistance, 128–130
- Substrate hot-electron capability:
 - hot carrier structure, mechanical stress, 313–314
 - MOSFET parasitic effects, gate-induced drain leakage (GIDL), 123–124
- Substrate injection, floating-gate memory physics, charge transfer, 384–385
- Substrate resistance network, digital applications of power transistors, cutoff and maximum oscillation frequencies, 591–597
- Substrates, silicon-on-insulator (SOI) devices:
 - epitaxial lateral overgrowth, 225
 - laser recrystallization, 223–224
 - silicon-on-sapphire (SOS) material, 222–223
 - SIMOX (separation by implanted oxygen), 225–227
 - UNIBOND material, 228–230
 - wafer bonding and etchback, 227–228
 - zone-melting recrystallization, 224–225
- Subthreshold current, DRAM (dynamic random access memory) circuits, pass transistor subthreshold leakage, 342–343
- Subthreshold region:
 - drain current, first-order MOSFET models, 82–83
 - MOSFET models:
 - device scaling, 86–88
 - short-channel effects, 88–99
 - charge-sharing model, 89–97
 - quasi-two-dimensional analysis, 97–99
- Subthreshold slope:
 - dual-gate MOSFET, 244–248
 - SOI MOSFET, 237
- Surface potential, drain current, subthreshold region, first-order MOSFET models, 83
- Surface roughness scattering, MOSFET transport properties, 104
- mobility models, 107–108

- Surface roughness scattering, MOSFET transport properties (*continued*)
 - universal mobility, 106–107
- Switching noise, CMOS digital switching, current mode logic (CML), 486–487
- Switching time and current, ferroelectric memory structure, 453–454
- Symbols table, 685–686
- Synchronous DRAM (SDRAM), high-speed DRAM (dynamic random access memory) circuits, 355–356
- Synchronous-link DRAM (SLDRAM), high-speed DRAM (dynamic random access memory) circuits, 358–359
- Systematic mismatch, system-on-chip technology limits, 645–646
- System function implementation, block integration, system-on-chip concepts, 667–670
 - modulated signal conversion, 668
 - oversampled D/A converter, 668–669
- System-on-chip concepts:
 - analog interfaces:
 - analog-to-digital conversion terminology, 651–655
 - conversion architectures, 655–664
 - A/D converters, 659–662
 - comparator matching, 662–664
 - D/A converters, 655–659
 - block integration:
 - interference, 670–672
 - partitioning, 672–676
 - analog/digital partitioning, 673–675
 - digital/memory choices, 672–673
 - multichip packaging, 675–676
 - signal processing strategy, 665–667
 - modulated signal sampling, 666–667
 - sample rate and accuracy, 665–666
 - sampling rate and bandwidth, 665
 - system function implementation, 667–670
 - modulated signal conversion, 668
 - oversampled D/A converter, 668–669
 - definitions, 632–633
 - digital video front-end chip, 676, 678
 - embedded modules on IC, 633–639
 - digital blocks, 636–639
 - functionality, 633–635
 - heterogeneous ICs, 635–636
 - future applications, 679–681
 - limits of technology:
 - component matching, 646–651
 - low voltage, 649–650
 - local threshold variation, 651
 - power and accuracy limits, 650–651
 - systematic and random mismatch, 645–646
 - mixed-signal circuits, 639–644
 - feature size, 641–642
 - process options, 642–643
 - signal swing, 640–641
 - tolerances, 643–644
 - one-chip oscilloscope, 679–680
 - one-chip television chip, 676–677
 - research background, 631–633
- Tantalum pentoxide, DRAM (dynamic random access memory) circuits, memory cell scaling, 337–339
- Taylor series expansion:
 - drain current, subthreshold region, first-order MOSFET models, 82–83
 - phonon scattering, 178–181
 - semiclassical electron dynamics, 162
- T-cell Flash EEPROM:
 - cell structure and array architecture, 407–410
 - disturb failures, 425–426
- Technology trends, semiconductors, 9–13
- Temperature dependence:
 - ferroelectric memory:
 - materials and structure, 450–452
 - polarization, 454–455
 - inner-box shaped transistor, collector current, 39–42
- Textured poly E²PROM cell, layout and cross section, 404–406
- Theoretical maximum efficiency, CMOS technology, power amplifiers, 538–540
- Thermal diffusion, inner-box shaped transistor, heat flux, 43–45
- Thermal effects, inner-box shaped transistor, 39–51
 - heat flux and thermal diffusion, device performance, 43–45
 - one-dimensional heat transfer, analytical approach, 46–49
 - safe operation area (SOA), 49–51
 - second breakdown instabilities, 42–43
 - temperature-dependent current, 39–42
- Thermal lenses, inner-box shaped transistor, second breakdown thermal instabilities, 43
- Thermal noise values, high-speed digital applications, noise parameters, 607–616
- Thermionic field emission, DRAM (dynamic random access memory) circuits, junction leakage, 342
- Thick-film SOI device:
 - floating-gate memory arrays, endurance failures, 423–424

- SOI MOSFETS, 232
- Thin-film SOI device:
 - floating-gate memory arrays, endurance failures, 423–424
- SOI MOSFETS, 232
- Third-order intermodulation (IMD3), CMOS low-noise amplifiers (LNA), 496–508
- Three-dimensional parabolic density of states, equilibrium statistics, 165
- Three-stage ring oscillators, voltage-controlled oscillators (VCOs), highly integrated transceivers, 526–535
- Threshold matching, system-on-chip technology limits, component matching, 647–651
- Threshold mismatch coefficients, system-on-chip technology limits, component matching, 648–649
- Threshold variation, system-on-chip technology limits, component matching, 651
- Threshold voltage (V_T):
 - dual-gate MOSFET, 246–248
 - high-speed digital applications, 552–553
 - low voltage/low power considerations:
 - dual- V_T CMOS, 582–583
 - low- V_T CMOS, substrate bias, 583–584
 - mixed-signal circuits, system-on-chip concepts, 640–641
 - MOSFETs, short-channel effects, 88–99
 - charge-sharing model, 89–97
 - quasi-two-dimensional analysis, 97–99
 - silicon nitride memory, decay rate, 435–438
 - SOI MOSFET, 232–234
 - body effect, 234–235
- Time constants:
 - hot-carrier effect (HCE), low gate voltage stresses, 287
 - system-on-chip concepts, D/A converters, 656
- Time-domain multiplication, CMOS technology, radiofrequency (RF) circuit design, mixer fundamentals, 509–513
- Time power laws, hot-carrier effect (HCE), intermediate gate voltage stresses, 281–283
- Timing signals, block integration, system-on-chip concepts, interference, 671–672
- TiN capping, high-speed digital applications,
 - TiSi₂ salicide narrow-line effect, 569–571
- TiSi₂ salicide, high-speed digital applications, 568–570
- TMA MEDICI, MOSFET simulations, 200–211
- Tolerances, mixed-signal circuits, system-on-chip concepts, 643–644
- Total harmonic distortion (THD), analog-to-digital (A/D) conversion, 653–655
- Total power consumption, low voltage/low power considerations, 581–585
- Transceivers:
 - CMOS technology, low-noise amplifiers (LNA), 493–508
 - highly integrated structure, CMOS low-noise amplifiers (LNA), 504–508
- Transconductance:
 - change threshold, dual-gate MOSFET, 245–248
 - CMOS low-noise amplifiers (LNA), highly integrated transceivers, 506–508
 - SOI MOSFET, 235–237
- Transient dissipated power, inner-box shaped transistor, safe operation area (SOA), 51
- Transistor, system-on-chip concepts, D/A converters, 656
- Transistor ferroelectric memory cell, architecture of, 462–463
- Transport properties:
 - high field drift velocity, 109–110
 - inner-box shaped transistor, static characteristics, 31–35
 - mobility, 103–109
 - models, 107–108
 - numerical simulation models, 108–109
 - scattering mechanisms, 103–104
 - universal mobility, 104–107
 - MOSFETs, 102–110
- Transversal filtering, embedded modules on IC, system-on-chip concepts, 634–635
- Transverse acoustic deformation potential (TADP) scattering, phonon scattering, 179–181
- Trapped charge, silicon nitride memory, tunneling-emission mechanisms, 434–435
- Trench capacitor DRAM cell, structure of, 335
- Triple-poly, virtual ground (TPVG) flash cell:
 - disturb failures, 425–426
 - floating-gate memory arrays, 412–414
- Tuning techniques, digital applications, large-signal power and efficiency, 606–607
- Tunnel epitaxy, silicon-on-insulator (SOI) devices, epitaxial lateral overgrowth (ELO) variations, 225
- Tunneling mechanisms:
 - floating-gate memory physics, Fowler-Nordheim tunneling, 386–389
 - silicon nitride memory, 432–435
- Tunnel oxide thickness, floating-gate memory arrays, scaling trends, 429–430
- Two-dimensional constant density of states, equilibrium statistics, 165

- Two-dimensional electron gas (2DEG)
 - quantization:
 - dual-gate MOSFET, quantum-effect devices, 250–254
 - MOSFET inversion layers, Boltzmann transport equation (BTE) simulations, 200–210
- Two-dimensional simulations:
 - charge-pumping detection, damage localization, 308
 - MOSFETs, short-channel effects, quasi-two-dimensional analysis, 97–99
 - self-adjusted structures, high current density operations, 56–57
- Two-phase heating, silicon-on-insulator (SOI) devices, UNIBOND material, 229
- Two-port circuit theory, bipolar transistors, 23
- ULSI-related devices:
 - developmental milestones, 3–9
 - information sources concerning, 14–15
- Ultraviolet light erase, floating-gate memory physics, charge transfer, 389–391
- Unary representation, system-on-chip concepts, D/A converters, 656–658
- UNIBOND material, silicon-on-insulator (SOI) devices, 228–230
- Uniform (100) orientation, silicon-on-insulator (SOI) devices, laser recrystallization, 224
- Unilateral gain (ULG):
 - digital applications of power transistors, cutoff and maximum oscillation frequencies, 590–597
 - microwave MOSFETs, 237–238
- Unit prefixes, 689
- Universal conductance fluctuations, mesoscopic systems, 216
- Universal mobility, MOSFET transport properties, 104–107
- UV EPROMs, floating-gate memory arrays, 393–402
- Variational techniques, MOSFET inversion layers, Boltzmann transport equation (BTE) simulations, 209–210
- Velocity-field relationships, MOSFET transport properties, 109–110
- Velocity saturation:
 - high-speed digital applications, deep-submicrometer CMOS transistor, 550–551
 - inner-box shaped transistor, temperature dependence, collector current, 40–42
- Velocity-saturation-limited drain saturation
 - current, MOSFET transport properties, 109–110
- Voltage-controlled oscillators (VCOs):
 - CMOS technology:
 - fundamentals, 520–525
 - highly integrated transceivers, 525–535
 - power amplifiers, 535–540
 - monolithic designs, comparison of, 531–533
- Voltage drop, inner-box shaped transistor, static characteristics, 34–35
- Voltage scaling, floating-gate memory arrays, 429–430
- Voltage standing-wave ratio (VSWR), CMOS low-noise amplifiers (LNA), highly integrated transceivers, 504–508
- Voltage-to-frequency converter, CMOS technology, 520–525
- Voltage transfer characteristics (VTC), digital switching, CMOS technology, static inverter, 479–481
- Volume inversion, dual-gate MOSFET, 244–248
- V_T rolloff, high-speed digital applications:
 - SCE-channel dopant N_{sub} engineering, 557–560
 - short-channel effects (SCE), 556
- Wafer bonding, silicon-on-insulator (SOI) devices, 227–228
- Wide-bandgap emitter, heterojunction bipolar transistor (HBT), 62–63
- Wiedemann-Franz law, Boltzmann transport equation (BTE), 198–199
- Wigner distribution, Boltzmann transport equation (BTE), 214–216
- Wigner-Seitz cell, band structure, dispersion relationship, 159–160
- Wiring systems, state-of-the-art bulk MOSFET, 132–133
- WKB approximation, floating-gate memory physics, Fowler-Nordheim tunneling, 386–389
- Word line (WL):
 - byte-alterable E²PROMs, floating-gate tunnel oxide (FLOTOX), 403–406
 - DRAM (dynamic random access memory) operation:
 - boosters, 352–353
 - sense amplifiers, 351–352
 - floating-gate memory arrays:
 - disturb failures, 425–426
 - UV EPROMs, 400–402
- Write-enable (WE) pin, DRAM (dynamic random access memory) circuits, 346–348

- Write/erase characteristics:
 - silicon nitride memory, endurance failures, 438
 - silicon nitride memory, tunneling-emission mechanisms, 434–435
- Write operations:
 - silicon nitride memory, write inhibit and repeated erase, 438–440
 - textured poly E²PROM cell, 406
- W/TiN metal gates, high-speed digital applications, poly depletion, 577
- X cells, floating-gate memory arrays, UV EPROMS, 395–396
- XMOS device, dual-gate MOSFET, 244–248
- Zener tunneling, semiclassical electron dynamics, 214–216
- Zero-temperature coefficient (ZTC), silicon-on-insulator (SOI) devices, high-temperature circuits, 255–256
- Zeroth moment, Boltzmann transport equation (BTE), 196–200
- Zone-melting recrystallization (ZMR), silicon-on-insulator (SOI) devices, 224–225
- Z_{th} curvature, inner-box shaped transistor, one-dimensional heat transfer, 47–49

A complete guide to current knowledge and future trends in ULSI devices

Ultra-Large-Scale Integration (ULSI), the next generation of semiconductor devices, has become a hot topic of investigation. *ULSI Devices* provides electrical and electronic engineers, applied physicists, and anyone involved in IC design and process development with a much-needed overview of key technology trends in this area. Edited by two of the foremost authorities on semiconductor device physics, with contributions by some of the best-known researchers in the field, this comprehensive reference examines such major ULSI devices as MOSFET, nonvolatile semiconductor memory (NVM), and the bipolar transistor, and the improvements these devices offer in power consumption, low-voltage and high-speed operation, and system-on-chip for ULSI applications. Supplemented with introductory material and references for each chapter as well as more than 400 illustrations, coverage includes:

- The physics and operational characteristics of the different components
- The evolution of device structures
- The ultimate limitations on device and circuit performance
- Device miniaturization and simulation
- Issues of reliability and the hot carrier effect
- Digital and analog circuit building blocks

C. Y. CHANG, PhD, is a National Chair Professor at the National Chiao Tung University in Hsinchu, Taiwan and a Foreign Associate of the National Academy of Engineering. **S. M. SZE, PhD**, is UMC Chair Professor at the National Chiao Tung University.

Cover Design: David Levy

Cover Photograph: First monolithic integrated circuit invented by R. N. Noyce in 1959. It is a flip-flop circuit containing six devices using 20 μ m design rules.

WILEY-INTERSCIENCE

John Wiley & Sons, Inc.

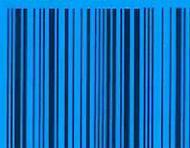
Scientific, Technical, and Medical Division

605 Third Avenue, New York, N.Y. 10158-0012

New York • Chichester • Weinheim

Brisbane • Singapore • Toronto

ISBN 0-471-24067-2



90000



9 780471 240679