Fundamentals of Modern VLSI Devices

SECOND EDITION

YUAN TAUR
University of California, San Diego

TAK H. NING
IBM T. J. Watson Research Center, New York

CAMBRIDGE
UNIVERSITY PRESS

SAMSUNG EXHIBIT 1015

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521832946

© Cambridge University Press 1998, 2009

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 1998
Second edition 2009

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Taur, Yuan, 1946-
Fundamentals of modern VLSI devices / Yuan Taur, Tak H. Ning. - 2nd ed.
p. cm.
ISBN 978-0-521-83294-6
1. Metal oxide semiconductors, Complementary. 2. Bipolar transistors.
3. Integrated circuits – Very large scale integration.
I. Ning, Tak H., 1943– II. Title.
TK7871.99.M44T38 2009
621.39'5–dc22

2009007334

ISBN 978-0-521-83294-6 hardback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party Internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

Contents

Preface to the first edition *page* xi
Preface to the second edition xiii
Physical constants and unit conversions xv
List of symbols xvi

1 Introduction 1

1.1 Evolution of VLSI Device Technology 1
1.1.1 Historical Perspective 1
1.1.2 Recent Developments 4
1.2 Modern VLSI Devices 4
1.2.1 Modern CMOS Transistors 4
1.2.2 Modern Bipolar Transistors 5
1.3 Scope and Brief Description of the Book 6

2 Basic Device Physics 11

2.1 Electrons and Holes in Silicon 11
2.1.1 Energy Bands in Silicon 11
2.1.2 n-Type and p-Type Silicon 17
2.1.3 Carrier Transport in Silicon 23
2.1.4 Basic Equations for Device Operation 27
2.2 p-n Junctions 35
2.2.1 Energy-Band Diagrams for a p–n Diode 35
2.2.2 Abrupt Junctions 38
2.2.3 The Diode Equation 46
2.2.4 Current–Voltage Characteristics 51
2.2.5 Time-Dependent and Switching Characteristics 64
2.2.6 Diffusion Capacitance 70
2.3 MOS Capacitors 72
2.3.1 Surface Potential: Accumulation, Depletion, and Inversion 72
2.3.2 Electrostatic Potential and Charge Distribution in Silicon 78
2.3.3 Capacitances in an MOS Structure 85
2.3.4 Polysilicon-Gate Work Function and Depletion Effects 91
2.3.5 MOS under Nonequilibrium and Gated Diodes 94

2.3.6 Charge in Silicon Dioxide and at the Silicon–Oxide Interface 98
2.3.7 Effect of Interface Traps and Oxide Charge on Device Characteristics 103
2.4 Metal–Silicon Contacts 108
2.4.1 Static Characteristics of a Schottky Barrier Diode 108
2.4.2 Current Transport in a Schottky Barrier Diode 115
2.4.3 Current–Voltage Characteristics of a Schottky Barrier Diode 115
2.4.4 Ohmic Contacts 120
2.5 High-Field Effects 122
2.5.1 Impact Ionization and Avalanche Breakdown 122
2.5.2 Band-to-Band Tunneling 125
2.5.3 Tunneling into and through Silicon Dioxide 127
2.5.4 Injection of Hot Carriers from Silicon into Silicon Dioxide 133
2.5.5 High-Field Effects in Gated Diodes 135
2.5.6 Dielectric Breakdown 137
Exercises 141

3 MOSFET Devices 148

3.1 Long-Channel MOSFETs 148
3.1.1 Drain-Current Model 149
3.1.2 MOSFET *I–V* Characteristics 155
3.1.3 Subthreshold Characteristics 163
3.1.4 Substrate Bias and Temperature Dependence of Threshold Voltage 166
3.1.5 MOSFET Channel Mobility 169
3.1.6 MOSFET Capacitances and Inversion-Layer Capacitance Effect 172
3.2 Short-Channel MOSFETs 175
3.2.1 Short-Channel Effect 176
3.2.2 Velocity Saturation and High-Field Transport 186
3.2.3 Channel Length Modulation 195
3.2.4 Source–Drain Series Resistance 196
3.2.5 MOSFET Degradation and Breakdown at High Fields 196
Exercises 201

4 CMOS Device Design 204

4.1 MOSFET Scaling 204
4.1.1 Constant-Field Scaling 204
4.1.2 Generalized Scaling 207
4.1.3 Nonscaling Effects 210
4.2 Threshold Voltage 212
4.2.1 Threshold-Voltage Requirement 213
4.2.2 Channel Profile Design 217
4.2.3 Nonuniform Doping 224
4.2.4 Quantum Effect on Threshold Voltage 234
4.2.5 Discrete Dopant Effects on Threshold Voltage 239

4.3 MOSFET Channel Length 242
4.3.1 Various Definitions of Channel Length 242
4.3.2 Extraction of the Effective Channel Length 244
4.3.3 Physical Meaning of Effective Channel Length 248
4.3.4 Extraction of Channel Length by *C–V* Measurements 252
Exercises 254

5 CMOS Performance Factors 256

5.1 Basic CMOS Circuit Elements 256
5.1.1 CMOS Inverters 256
5.1.2 CMOS NAND and NOR Gates 266
5.1.3 Inverter and NAND Layouts 270
5.2 Parasitic Elements 273
5.2.1 Source–Drain Resistance 274
5.2.2 Parasitic Capacitances 277
5.2.3 Gate Resistance 280
5.2.4 Interconnect *R* and *C* 283
5.3 Sensitivity of CMOS Delay to Device Parameters 289
5.3.1 Propagation Delay and Delay Equation 289
5.3.2 Delay Sensitivity to Channel Width, Length, and Gate Oxide Thickness 296
5.3.3 Sensitivity of Delay to Power-Supply Voltage and Threshold Voltage 299
5.3.4 Sensitivity of Delay to Parasitic Resistance and Capacitance 301
5.3.5 Delay of Two-Way NAND and Body Effect 304
5.4 Performance Factors of Advanced CMOS Devices 307
5.4.1 MOSFETs in RF Circuits 308
5.4.2 Effect of Transport Parameters on CMOS Performance 311
5.4.3 Low-Temperature CMOS 312
Exercises 315

6 Bipolar Devices 318

6.1 n–p–n Transistors 318
6.1.1 Basic Operation of a Bipolar Transistor 322
6.1.2 Modifying the Simple Diode Theory for Describing Bipolar Transistors 322
6.2 Ideal Current–Voltage Characteristics 327
6.2.1 Collector Current 329
6.2.2 Base Current 330
6.2.3 Current Gains 334
6.2.4 Ideal I_C–V_{CE} Characteristics 336
6.3 Characteristics of a Typical n–p–n Transistor 337
6.3.1 Effect of Emitter and Base Series Resistances 338
6.3.2 Effect of Base–Collector Voltage on Collector Current 340
6.3.3 Collector Current Falloff at High Currents 343
6.3.4 Nonideal Base Current at Low Currents 347

6.4 Bipolar Device Models for Circuit and Time-Dependent Analyses 352
6.4.1 Basic dc Model 352
6.4.2 Basic ac Model 355
6.4.3 Small-Signal Equivalent-Circuit Model 356
6.4.4 Emitter Diffusion Capacitance 359
6.4.5 Charge-Control Analysis 361
6.5 Breakdown Voltages 366
6.5.1 Common-Base Current Gain in the Presence of Base–Collector Junction Avalanche 367
6.5.2 Saturation Currents in a Transistor 369
6.5.3 Relation Between BV_{CEO} and BV_{CBO} 370
Exercises 371

7 Bipolar Device Design 374

7.1 Design of the Emitter Region 374
7.1.1 Diffused or Implanted-and-Diffused Emitter 375
7.1.2 Polysilicon Emitter 376
7.2 Design of the Base Region 377
7.2.1 Relationship between Base Sheet Resistivity and Collector Current Density 378
7.2.2 Intrinsic-Base Dopant Distribution 380
7.2.3 Electric Field in the Quasineutral Intrinsic Base 381
7.2.4 Base Transit Time 384
7.3 Design of the Collector Region 385
7.3.1 Collector Design When There Is Negligible Base Widening 387
7.3.2 Collector Design When There Is Appreciable Base Widening 388
7.4 SiGe-Base Bipolar Transistors 389
7.4.1 Transistors Having a Simple Linearly Graded Base Bandgap 390
7.4.2 Base Current When Ge Is Present in the Emitter 396
7.4.3 Transistors Having a Trapezoidal Ge Distribution in the Base 401
7.4.4 Transistors Having a Constant Ge Distribution in the Base 406
7.4.5 Effect of Emitter Depth Variation on Device Characteristics 410
7.4.6 Some Optimal Ge Profiles 414
7.4.7 Base-Width Modulation by V_{BE} 419
7.4.8 Reverse-Mode I–V Characteristics 423
7.4.9 Heterojunction Nature of a SiGe-Base Bipolar Transistor 426
7.5 Modern Bipolar Transistor Structures 429
7.5.1 Deep-Trench Isolation 429
7.5.2 Polysilicon Emitter 430
7.5.3 Self-Aligned Polysilicon Base Contact 430
7.5.4 Pedestal Collector 431
7.5.5 SiGe-Base 431
Exercises 432

8 Bipolar Performance Factors 437

8.1 Figures of Merit of a Bipolar Transistor 437
8.1.1 Cutoff Frequency 437
8.1.2 Maximum Oscillation Frequency 440
8.1.3 Ring Oscillator and Gate Delay 440
8.2 Digital Bipolar Circuits 441
8.2.1 Delay Components of a Logic Gate 442
8.2.2 Device Structure and Layout for Digital Circuits 445
8.3 Bipolar Device Optimization for Digital Circuits 447
8.3.1 Design Points for a Digital Circuit 447
8.3.2 Device Optimization When There Is Significant Base Widening 448
8.3.3 Device Optimization When There Is Negligible Base Widening 449
8.3.4 Device Optimization for Small Power–Delay Product 453
8.3.5 Bipolar Device Optimization from Some Data Analyses 455
8.4 Bipolar Device Scaling for ECL Circuits 457
8.4.1 Device Scaling Rules 458
8.4.2 Limits in Bipolar Device Scaling for ECL Circuits 460
8.5 Bipolar Device Optimization and Scaling for RF and Analog Circuits 463
8.5.1 The Single-Transistor Amplifier 463
8.5.2 Optimizing the Individual Parameters 464
8.5.3 Technology for RF and Analog Bipolar Devices 467
8.5.4 Limits in Scaling Bipolar Transistors for RF and Analog Applications 468
8.6 Comparing a SiGe-Base Bipolar Transistor with a GaAs HBT 469
Exercises 472

9 Memory Devices 476

9.1 Static Random-Access Memory 477
9.1.1 CMOS SRAM Cell 478
9.1.2 Other Bistable MOSFET SRAM Cells 486
9.1.3 Bipolar SRAM Cell 487
9.2 Dynamic Random-Access Memory 495
9.2.1 Basic DRAM Cell and Its Operation 496
9.2.2 Device Design and Scaling Considerations for a DRAM Cell 499
9.3 Nonvolatile Memory 500
9.3.1 MOSFET Nonvolatile Memory Devices 501
9.3.2 Flash Memory Arrays 507
9.3.3 Floating-Gate Nonvolatile Memory Cells 511
9.3.4 Nonvolatile Memory Cells with Charge Stored in Insulator 514
Exercise 516

10 Silicon-on-Insulator Devices 517

10.1 SOI CMOS 517
10.1.1 Partially Depleted SOI MOSFETs 518
10.1.2 Fully Depleted SOI MOSFETs 520
10.2 Thin-Silicon SOI Bipolar 523
10.2.1 Fully Depleted Collector Mode 524
10.2.2 Partially Depleted Collector Mode 526
10.2.3 Accumulation Collector Mode 527
10.2.4 Discussion 527
10.3 Double-Gate MOSFETs 529
10.3.1 An Analytic Drain Current Model for Symmetric DG MOSFETs 529
10.3.2 The Scale Length of Double-Gate MOSFETs 533
10.3.3 Fabrication Requirements and Challenges of DG MOSFETs 534
10.3.4 Multiple-Gate MOSFETs 536
Exercise 537

Appendix 1 CMOS Process Flow 538
Appendix 2 Outline of a Process for Fabricating Modern n–p–n Bipolar Transistors 542
Appendix 3 Einstein Relations 543
Appendix 4 Spatial Variation of Quasi-Fermi Potentials 546
Appendix 5 Generation and Recombination Processes and Space-Charge-Region Current 553
Appendix 6 Diffusion Capacitance of a p–n Diode 562
Appendix 7 Image-Force-Induced Barrier Lowering 569
Appendix 8 Electron-Initiated and Hole-Initiated Avalanche Breakdown 573
Appendix 9 An Analytical Solution for the Short-Channel Effect in Subthreshold 575
Appendix 10 Generalized MOSFET Scale Length Model 582
Appendix 11 Drain Current Model of a Ballistic MOSFET 588
Appendix 12 Quantum-Mechanical Solution in Weak Inversion 594
Appendix 13 Power Gain of a Two-Port Network 598
Appendix 14 Unity-Gain Frequencies of a MOSFET Transistor 601
Appendix 15 Determination of Emitter and Base Series Resistances 605
Appendix 16 Intrinsic-Base Resistance 610
Appendix 17 Energy-Band Diagram of a Si–SiGe n–p Diode 614
Appendix 18 f_T and f_{max} of a Bipolar Transistor 617
References 623
Index 644

Preface to the first edition

It has been fifty years since the invention of the bipolar transistor, more than forty years since the invention of the integrated-circuit (IC) technology, and more than thirty-five years since the invention of the MOSFET. During this time, there has been a tremendous and steady progress in the development of the IC technology with a rapid expansion of the IC industry. One distinct characteristic in the evolution of the IC technology is that the physical feature sizes of the transistors are reduced continually over time as the lithography technologies used to define these features become available. For almost thirty years now, the minimum lithography feature size used in IC manufacturing has been reduced at a rate of 0.7× every three years. In 1997, the leading-edge IC products have a minimum feature size of 0.25 μm.

The basic operating principles of large and small transistors are the same. However, the relative importance of the various device parameters and performance factors for transistors of the 1-μm and smaller generations is quite different from those for transistors of larger-dimension generations. For example, in the case of CMOS, the power-supply voltage was lowered from the standard 5 V, starting with the 0.6- to 0.8-μm generation. Since then CMOS power supply voltage has been lowered in steps once every few years as the device physical dimensions are reduced. At the same time, many physical phenomena, such as short-channel effect and velocity saturation, which are negligible in large-dimension MOSFETs, are becoming more and more important in determining the behavior of MOSFETs of deep-submicron dimensions. In the case of bipolar devices, breakdown voltage and base-widening effects are limiting their performance, and power dissipation is limiting their level of integration on a chip. Also, the advent of SiGe-base bipolar technology has extended the frequency capability of small-dimension bipolar transistors into the range previously reserved for GaAs and other compound-semiconductor devices.

The purpose of this book is to bring together the device fundamentals that govern the behavior of CMOS and bipolar transistors into a single text, with emphasis on those parameters and performance factors that are particularly important for VLSI (very-large-scale-integration) devices of deep-submicron dimensions. The book starts with a comprehensive review of the properties of the silicon material, and the basic physics of p–n junctions and MOS capacitors, as they relate to the fundamental principles of MOSFET and bipolar transistors. From there, the basic operation of MOSFET and bipolar devices, and their design and optimization for VLSI applications are developed. A great deal of the volume is devoted to in-depth discussions of the intricate interdependence and subtle tradeoffs of the various device parameters pertaining to circuit performance and manufacturability. The effects which are particularly important in small-dimension devices,

e.g., quantization of the two-dimensional surface inversion layer in a MOSFET device and the heavy-doping effect in the intrinsic base of a bipolar transistor, are covered in detail. Also included in this book are extensive discussions on scaling and limitations to scaling of MOSFET and bipolar devices.

This book is suitable for use as a textbook by senior undergraduate or graduate students in electrical engineering and microelectronics. The necessary background assumed is an introductory understanding of solid-state physics and semiconductor physics. For practicing engineers and scientists actively involved in research and development in the IC industry, this book serves as a reference in providing a body of knowledge in modern VLSI devices for them to stay up to date in this field.

VLSI devices are too huge a subject area to cover thoroughly in one book. We have chosen to cover only the fundamentals necessary for discussing the design and optimization of the state-of-the-art CMOS and bipolar devices in the sub-0.5-μm regime. Even then, the specific topics covered in this book are based on our own experience of what the most important device parameters and performance factors are in modern VLSI devices.

Many people have contributed directly and indirectly to the topics covered in this book. We have benefited enormously from the years of collaboration and interaction we had with our colleagues at IBM, particularly in the areas of advanced silicon-device research and development. These include Douglas Buchanan, Hu Chao, T. C. Chen, Wei Chen, Kent Chuang, Peter Cook, Emmanuel Crabbé, John Cressler, Bijan Davari, Robert Dennard, Max Fischetti, David Frank, Charles Hsu, Genda Hu, Randall Isaac, Khalid Ismail, G. P. Li, Shih-Hsien Lo, Yuh-Jier Mii, Edward Nowak, George Sai-Halasz, Stanley Schuster, Paul Solomon, Hans Stork, Jack Sun, Denny Tang, Lewis Terman, Clement Wann, James Warnock, Siegfried Wiedmann, Philip Wong, Matthew Wordeman, Ben Wu, and Hwa Yu.

We would like to acknowledge the secretarial support of Barbara Grady and the support of our management at IBM Thomas J. Watson Research Center where this book was written. Finally, we would like to give special thanks to our families – Teresa, Adrienne, and Brenda Ning and Betty, Ying, and Hsuan Taur – for their support and understanding during this seemingly endless task.

Yuan Taur
Tak H. Ning
Yorktown Heights, New York, October, 1997

Preface to the second edition

Since the publication of the first edition of *Fundamentals of Modern VLSI Devices* by Cambridge University Press in 1998, we received much praise and many encouraging reviews on the book. It has been adopted as a textbook for first-year graduate courses on microelectronics in many major universities in the United States and worldwide. The first edition was translated into Japanese by a team led by Professor Shibahara of Hiroshima University in 2002.

During the past 10 years, the evolution and scaling of VLSI (very-large-scale-integration) technology has continued. Now, sixty years after the first invention of the transistor, the number of transistors per chip for both microprocessors and DRAM (dynamic random access memory) has increased to over one billion, and the highest clock frequency of microprocessors has reached 5 GHz. In 2007, the worldwide IC (integrated circuits) sales grew to $250 billion. In 2008, the IC industry reached the 45-nm generation, meaning that the leading-edge IC products employ a minimum lithography feature size of 45 nm. As bulk CMOS (complementary metal–oxide–semiconductor field-effect transistor) technologies are scaled to dimensions below 100 nm, the very factor that makes CMOS technology the technology of choice for digital VLSI circuits, namely, its low standby power, can no longer be taken for granted. Not only has the off-state current gone up with the power supply voltage down scaled to the 1 V level, the gate leakage has also increased exponentially from quantum mechanical tunneling through gate oxides only a few atomic layers thick. Power management, both active and standby, has become a key challenge to continued increase of clock frequency and transistor count in microprocessors. New materials and device structures are being explored to replace conventional bulk CMOS in order to extend scaling to 10 nm.

The purpose of writing the second edition is to update the book with additional material developed after the completion of the first edition. Key new material added includes MOSFET scale length theory and high-field transport model, and the section on SiGe-base bipolar devices has been greatly expanded. We have also expanded the discussions on basic device physics and circuits to include metal–silicon contacts, noise margin of CMOS circuits, and figures of merit for RF applications. Furthermore, two new chapters are added to the second edition. Chapter 9 is on memory devices and covers the fundamentals of read and write operations of commonly used SRAM, DRAM, and nonvolatile memory arrays. Chapter 10 is on silicon-on-insulator (SOI) devices, including advanced devices of future potential.

We would like to take this opportunity to thank all the friends and colleagues who gave us encouragement and valuable suggestions for improvement of the book. In particular, Professor Mark Lundstrom of Purdue University who adopted the first edition early on,

and Dr. Constantin Bulucea of National Semiconductor Corporation who suggested the treatment on diffusion capacitance. Thanks also go to Professor James Meindl of Georgia Institute of Technology, Professor Peter Asbeck of University of California, San Diego, and Professor Jerry Fossum of University of Florida for their support of the book.

We would like to thank many of our colleagues at IBM, particularly in the areas of advanced silicon-device research and development, for their direct or indirect contributions. Yuan Taur would like to thank many of his students at University of California, San Diego, in particular Jooyoung Song and Bo Yu, for their help with the completion of the second edition. He would also like to thank Katie Kahng for her love, support, and patience during the course of the work.

We would like to give special thanks to our families for their support and understanding during this seemingly endless task.

Yuan Taur
Tak H. Ning
June, 2008

Physical constants and unit conversions

Description	Symbol	Value and unit
Electronic charge	q	1.6×10^{-19} C
Boltzmann's constant	k	1.38×10^{-23} J/K
Vacuum permittivity	ε_0	8.85×10^{-14} F/cm
Silicon permittivity	ε_{si}	1.04×10^{-12} F/cm
Oxide permittivity	ε_{ox}	3.45×10^{-13} F/cm
Velocity of light in vacuum	c	3×10^{10} cm/s
Planck's constant	h	6.63×10^{-34} J-s
Free-electron mass	m_0	9.1×10^{-31} kg
Thermal voltage ($T = 300$ K)	kT/q	0.0259 V
Angstrom	Å	$1\text{Å} = 10^{-8}$ cm
Nanometer	nm	$1\text{ nm} = 10^{-7}$ cm
Micrometer (micron)	μm	$1\ \mu\text{m} = 10^{-4}$ cm
Millimeter	mm	1 mm = 0.1 cm
Meter	m	$1\text{ m} = 10^{2}$ cm
Electron-volt	eV	$1\text{eV} = 1.6 \times 10^{-19}$ J
Energy = charge × voltage	$E = qV$	Joule = Coulomb × Volt
Charge = capacitance × voltage	$Q = CV$	Coulomb = Farad × Volt
Power = current × voltage	$P = IV$	Watt = Ampere × Volt
Time = resistance × capacitance	$t = RC$	second = Ω (ohm) × Farad
Current = charge/time	$I = Q/t$	Ampere = Coulomb/second
Resistance = voltage/current	$R = V/I$	Ω (ohm) = Volt/Ampere

A word of caution about the length units: strictly speaking, MKS units should be used for all the equations in the book. As a matter of convention, electronics engineers often work with centimeter as the unit of length. While some equations work with lengths in either meter or centimeter, not all of them do. It is prudent always to check for unit consistency when doing calculations. It may be necessary to convert the length unit to meter before plugging into the equations.

List of symbols

Symbol	Description	Unit
A	Area	cm^2
A_E	Emitter area	cm^2
α	Common-base current gain	None
α_0	Static common-base current gain	None
α_F	Forward common-base current gain in the Ebers–Moll model	None
α_R	Reverse common-base current gain in the Ebers–Moll model	None
α_T	Base transport factor	None
α_n	Electron-initiated rate of electron–hole pair generation per unit distance	cm^{-1}
α_p	Hole-initiated rate of electron–hole pair generation per unit distance	cm^{-1}
BV	Breakdown voltage	V
BV_{CBO}	Collector–base junction breakdown voltage with emitter open circuit	V
BV_{CEO}	Collector–emitter breakdown voltage with base open circuit	V
BV_{EBO}	Emitter–base junction breakdown voltage with collector open circuit	V
β	Current gain	None
β_0	Static common-emitter current gain	None
β_F	Forward common-emitter current gain in the Ebers–Moll model	None
β_R	Reverse common-emitter current gain in the Ebers–Moll model	None
c	Velocity of light in vacuum ($=3\times10^{10}$ cm/s)	cm/s
C	Capacitance	F
C_d	Depletion-layer capacitance per unit area	F/cm^2
$C_{d,tot}$	Total depletion-layer capacitance	F
C_{dBC}	Base–collector diode depletion-layer capacitance per unit area	F/cm^2
$C_{dBC,tot}$	Total base–collector diode depletion-layer capacitance	F
C_{dBE}	Base–emitter diode depletion-layer capacitance per unit area	F/cm^2
$C_{dBE,tot}$	Total base–emitter diode depletion-layer capacitance	F
C_{dm}	Maximum depletion-layer capacitance (per unit area)	F (F/cm^2)
C_D	Diffusion capacitance	F

C_{Dn}	Diffusion capacitance due to excess electrons	F
C_{Dp}	Diffusion capacitance due to excess holes	F
C_{DE}	Emitter diffusion capacitance	F
C_{DoS}	Equivalent density-of-states capacitance	F/m^2
C_{fb}	MOS capacitance at flat band per unit area	F/cm^2
C_{FC}	Capacitance between the floating gate and the control gate of a MOSFET nonvolatile memory device	F
C_g	Intrinsic gate capacitance per unit area	F/cm^2
C_G	Total gate capacitance of MOSFET	F
C_i	Inversion-layer capacitance per unit area	F/cm^2
C_{it}	Interface trap capacitance per unit area	F/cm^2
C_j	Junction capacitance per unit area	F/cm^2
C_J	Junction capacitance	F
C_L	Load capacitance	F
C_{in}	Equivalent input capacitance of a logic gate	F
C_{inv}	MOSFET capacitance in inversion per unit area	F/cm^2
$C_{\min}$	Minimum MOS capacitance per unit area	F/cm^2
C_{out}	Equivalent output capacitance of a logic gate	F
C_{ov}	Gate-to-source (-drain) overlap capacitance (per edge)	F
C_{ox}	Oxide capacitance per unit area	F/cm^2
C_p	Polysilicon-gate depletion-layer capacitance per unit area	F/cm^2
C_{si}	Silicon capacitance per unit area	F/cm^2
C_w	Wire capacitance per unit length	F/cm
C_π	Base–emitter capacitance in the small-signal hybrid-π equivalent-circuit model	F
C_μ	Base–collector capacitance in the small-signal hybrid-π equivalent-circuit model	F
d	Width of diffusion region in a MOSFET	cm
D_n	Electron diffusion coefficient	cm^2/s
D_{nB}	Electron diffusion coefficient in the base of an n–p–n transistor	cm^2/s
D_p	Hole diffusion coefficient	cm^2/s
D_{pE}	Hole diffusion coefficient in the emitter of an n–p–n transistor	cm^2/s
ΔV_t	Threshold voltage rolloff due to short-channel effect	V
ΔE_g	Apparent bandgap narrowing	J
ΔE_{gB}	Bandgap-narrowing parameter in the base region	J
$\Delta E_{g,\max}$	Maximum bandgap narrowing due to the presence of Ge	J
$\Delta E_{g,SiGe}$	Local bandgap narrowing due to the presence of Ge	J
ΔL	Channel length modulation in MOSFET	cm
ΔQ_{total}	Total charge stored in a nonvolatile memory device	C
E	Energy	J
E_c	Conduction-band edge	J
E_v	Valence-band edge	J
E_a	Ionized-acceptor energy level	J

E_d	Ionized-donor energy level	J
E_f	Fermi energy level	J
E_g	Energy gap of silicon	J
E_i	Intrinsic Fermi level	J
E_{fn}	Fermi energy level on the n-side of a p–n diode	J
E_{fp}	Fermi energy level on the p-side of a p–n diode	J
$\mathscr{E}$	Electric field	V/cm
$\mathscr{E}_c$	Critical field for velocity saturation	V/cm
$\mathscr{E}_{eff}$	Effective vertical field in MOSFET	V/cm
$\mathscr{E}_{ox}$	Oxide electric field	V/cm
$\mathscr{E}_s$	Electric field at silicon surface	V/cm
$\mathscr{E}_x$	Vertical field in silicon	V/cm
$\mathscr{E}_y$	Lateral field in silicon	V/cm
ε_0	Vacuum permittivity ($= 8.85 \times 10^{-14}$ F/cm)	F/cm
ε_i	Permittivity of gate insulator	F/cm
ε_{si}	Silicon permittivity ($= 1.04 \times 10^{-12}$ F/cm)	F/cm
ε_{ox}	Oxide permittivity ($= 3.45 \times 10^{-13}$ F/cm)	F/cm
f_D	Probability that an electronic state is filled	None
f	Frequency, clock frequency	Hz
f_{max}	Unity power gain frequency	Hz
f_T	Unity current gain frequency	Hz
FI	Fan-in	None
FO	Fan-out	None
ϕ	Barrier height	V
ϕ_{ox}	Silicon–silicon dioxide interface potential barrier for electrons	V
ϕ_{ms}	Work-function difference between metal and silicon	V
ϕ_n	Electron quasi-Fermi potential	V
ϕ_p	Hole quasi-Fermi potential	V
ϕ_{Bn}	Schottky barrier height for electrons	V
ϕ_{Bp}	Schottky barrier height for holes	V
g	Number of degeneracy	None
g_{ds}	Small-signal output conductance	A/V
g_m	Small-signal transconductance	A/V
G_E	Emitter Gummel number	s/cm^4
G_B	Base Gummel number	s/cm^4
G_n	Electron emission rate (also called electron generation rate)	1/cm^3-s
G_p	Hole emission rate (also called hole generation rate)	1/cm^3-s
γ	Emitter injection efficiency	None
h	Planck's constant ($= 6.63 \times 10^{-34}$ J-s)	J-s
i	Time-dependent current	A
i_B	Time-dependent base current in a bipolar transistor	A
i_b	Time-dependent small-signal base current	A
i_C	Time-dependent collector current in a bipolar transistor	A

i_c	Time-dependent small-signal collector current	A
i_E	Time-dependent emitter current in a bipolar transistor	A
I	Current	A
I_B	Static base current in a bipolar transistor	A
I_C	Static collector current in a bipolar transistor	A
I_E	Static emitter current in a bipolar transistor	A
I_S	Switch current in an ECL circuit	A
I_g	Gate current in a MOSFET	A
I_0	MOSFET current per unit width to length ratio for threshold definition	A
I_{dsat}	MOSFET saturation current	A
I_{on}	MOSFET on current	A
I_{off}	MOSFET off current	A
I_n	nMOSFET current per unit width	A/cm
I_p	pMOSFET current per unit width	A/cm
I_N	nMOSFET current	A
I_P	pMOSFET current	A
I_{ds}	Drain-to-source current in a MOSFET	A
I_{sx}	Substrate current in a MOSFET	A
$I_{ds,Vt}$	MOSFET current at threshold	A
$I_{on,n}$	nMOSFET on current per device width	A/cm
I_{onN}	nMOSFET on current	A
$I_{on,p}$	pMOSFET on current per device width	A/cm
I_{onP}	pMOSFET on current	A
λ	MOSFET scale length	cm
J	Current density	A/cm^2
J_B	Base current density	A/cm^2
J_C	Collector current density	A/cm^2
J_n	Electron current density	A/cm^2
J_p	Hole current density	A/cm^2
k	Boltzmann's constant ($= 1.38 \times 10^{-23}$ J/K)	J/K
κ	Scaling factor (> 1)	None
l	Mean free path	cm
L	Length, MOSFET channel length	cm
L_D	Debye length	cm
L_n	Electron diffusion length	cm
L_p	Hole diffusion length	cm
L_{met}	Metallurgical channel length of MOSFET	cm
L_{eff}	Effective channel length of MOSFET	cm
L_w	Wire length	cm
m	MOSFET body-effect coefficient	None
m_0	Free-electron mass ($= 9.1 \times 10^{-31}$ kg)	kg
m^*	Electron effective mass	kg

M	Avalanche multiplication factor	None
m_l	Electron effective mass in the longitudinal direction	kg
m_t	Electron effective mass in the transverse direction	kg
μ	Carrier mobility	cm^2/V-s
μ_{eff}	Effective mobility	cm^2/V-s
μ_n	Electron mobility	cm^2/V-s
μ_p	Hole mobility	cm^2/V-s
n	Density of free electrons	cm^{-3}
n_0	Density of free electrons at thermal equilibrium	cm^{-3}
n_i	Intrinsic carrier density	cm^{-3}
n_{ie}	Effective intrinsic carrier density	cm^{-3}
n_{ieB}	Effective intrinsic carrier density in base of bipolar transistor	cm^{-3}
n_{ieE}	Effective intrinsic carrier density in emitter of bipolar transistor	cm^{-3}
n_n	Density of electrons in n-region	cm^{-3}
n_p	Density of electrons in p-region	cm^{-3}
N_a	Acceptor impurity density	cm^{-3}
N_d	Donor impurity density	cm^{-3}
N_b	Impurity concentration in bulk silicon	cm^{-3}
N_c	Effective density of states of conduction band	cm^{-3}
N_v	Effective density of states of valence band	cm^{-3}
N_B	Base doping concentration	cm^{-3}
N_C	Collector doping concentration	cm^{-3}
N_E	Emitter doping concentration	cm^{-3}
$N(E)$	Density of electronic states per unit energy per volume	$1/J\text{-}m^3$
p	Density of free holes	cm^{-3}
p_0	Density of free holes at thermal equilibrium	cm^{-3}
p_n	Density of holes in n-region	cm^{-3}
p_p	Density of holes in p-region	cm^{-3}
P	Power dissipation	W
P_{ac}	Active power dissipation	W
P_{off}	Standby power dissipation	W
q	Electronic charge ($= 1.6 \times 10^{-19}$ C)	C
Q	Charge	C
Q_B	Excess minority charge per unit area in the base	C/cm^2
$Q_{B,tot}$	Total excess minority charge in the base	C
Q_{BE}	Excess minority charge per unit area in the base–emitter space-charge region	C/cm^2
$Q_{BE,tot}$	Total excess minority charge in the base–emitter space-charge region	C
Q_{BC}	Excess minority charge per unit area in the base–collector space-charge region	C/cm^2
$Q_{BC,tot}$	Total excess minority charge in the base–collector space-charge region	C

Q_{DE}	Total stored minority-carrier charge in a bipolar transistor biased in the forward-active mode	C
Q_E	Excess minority charge per unit area in the emitter	C/cm^2
$Q_{E,tot}$	Total excess minority charge in the emitter	C
Q_{pB}	Hole charge per unit area in base of n–p–n transistor	C/cm^2
Q_s	Total charge per unit area in silicon	C/cm^2
Q_d	Depletion charge per unit area	C/cm^2
Q_i	Inversion charge per unit area	C/cm^2
Q_f	Fixed oxide charge per unit area	C/cm^2
Q_g	Charge on MOS gate per unit area	C/cm^2
Q_m	Mobile charge per unit area	C/cm^2
Q_{it}	Interface trapped charge per unit area	C/cm^2
Q_n	Excess electron charge per unit area	C/cm^2
Q_{ot}	Oxide trapped charge per unit area	C/cm^2
Q_{ox}	Equivalent oxide charge density per unit area	C/cm^2
Q_p	Excess hole charge per unit area	C/cm^2
r, R	Resistance	Ω
r_b	Base resistance	Ω
r_{bi}	Intrinsic base resistance	Ω
r_{bx}	Extrinsic base resistance	Ω
r_c	Collector series resistance	Ω
r_e	Emitter series resistance	Ω
r_0	Output resistance in small-signal hybrid-π equivalent-circuit model	Ω
r_π	Input resistance in small-signal hybrid-π equivalent-circuit model	Ω
R_L	Load resistance in a circuit	Ω
R_s	Source series resistance	Ω
R_d	Drain series resistance	Ω
R_n	Electron capture rate (also called electron recombination rate)	$1/cm^3\text{-}s$
R_p	Hole capture rate (also called hole recombination rate)	$1/cm^3\text{-}s$
R_{sd}	Source–drain series resistance	Ω
R_{ch}	MOSFET channel resistance	Ω
R_w	Wire resistance per unit length	Ω/cm
R_{Sbi}	Sheet resistance of intrinsic-base layer	Ω/□
R_{sw}	Equivalent switching resistance of a CMOS gate	Ω
R_{swn}	Equivalent switching resistance of nMOSFET pulldown	Ω
R_{swp}	Equivalent switching resistance of pMOSFET pullup	Ω
ρ	Resistivity	Ω-cm
ρ_{sh}	Sheet resistivity	Ω/□
ρ_{ch}	Sheet resistivity of MOSFET channel	Ω/□
ρ_{sd}	Sheet resistivity of source or drain region	Ω/□
ρ_c	Specific contact resistivity	$\Omega\text{-}cm^2$

ρ_{net}	Volume density of net charge	C/cm^3
S	MOSFET inverse subthreshold current slope	V/decade
S_p	Surface recombination velocity for holes	cm/s
σ_L	Lateral straggle of Gaussian doping profile	cm
t	Time	s
t_B	Base transit time	s
t_E	Emitter transit time	s
t_{BE}	Base–emitter depletion-layer transit time	s
t_{BC}	Base–collector depletion-layer transit time	s
t_i	Thickness of gate insulator	cm
t_{inv}	Equivalent oxide thickness for inversion charge calculations	cm
t_{ox}	Oxide thickness	cm
t_r	Transit time	s
t_w	Thickness of wire	cm
t_{si}	Thickness of silicon film	cm
T	Absolute temperature	K
τ	Lifetime	s
τ	Circuit delay	s
τ_b	Buffered delay	s
τ_{int}	Intrinsic, unloaded delay	s
τ_F	Forward transit time of bipolar transistor	s
τ_n	Electron lifetime	s
τ_n	nMOSFET pulldown delay	s
τ_{nB}	Electron lifetime in base of n–p–n transistor	s
τ_p	Hole lifetime	s
τ_p	pMOSFET pullup delay	s
τ_{pE}	Hole lifetime in emitter of n–p–n transistor	s
τ_R	Reverse transit time of bipolar transistor	s
τ_w	Wire RC delay	s
τ_E	Emitter delay time	s
τ_B	Base delay time	s
τ_{BE}	Base–emitter depletion-region delay time	s
τ_{BC}	Base–collector depletion-region delay time	s
U	Net recombination rate	1/cm^3-s
v	Velocity	cm/s
v	Small-signal voltage	V
v_{th}	Thermal velocity	cm/s
v_d	Carrier drift velocity	cm/s
v_{sat}	Saturation velocity of carriers	cm/s
v_T	Thermal injection velocity at MOSFET source	cm/s
V	Voltage	V
V	Quasi-Fermi potential along MOSFET channel	V
V_A	Early voltage	V

V_{app}	Applied voltage across p–n diode	V
V'_{app}	Applied voltage appearing immediately across p–n junction (smaller than V_{app} by IR drops in series resistances)	V
V_{BE}	Base–emitter bias voltage	V
V_{BC}	Base–collector bias voltage	V
V_{CE}	Collector-to-emitter voltage	V
V_{CG}	Control gate voltage in a nonvolatile memory device	V
V_{FG}	Floating gate voltage in a nonvolatile memory device	V
V_{dd}	Power-supply voltage	V
V_{ds}	Source-to-drain voltage	V
V_{dsat}	MOSFET drain saturation voltage	V
V_{fb}	Flat-band voltage	V
V_{ox}	Potential drop across oxide	V
V_g	Gate voltage in MOS	V
V_{gs}	Gate-to-source voltage in a MOSFET	V
V_{bs}	MOSFET body bias voltage	V
V_t	Threshold voltage ($2\psi_B$ definition)	V
V_{on}	Linearly extrapolated threshold voltage	V
V_{in}	Input node voltage of a logic gate	V
V_{out}	Output node voltage of a logic gate	V
V_x	Node voltage between stacked nMOSFETs of a NAND gate	V
$V_{t,high}$	The higher threshold voltage of a nonvolatile memory device	V
$V_{t,low}$	The lower threshold voltage of a nonvolatile memory device	V
W	Width, MOSFET width	cm
W_n	nMOSFET width	cm
W_p	pMOSFET width	cm
W_B	Intrinsic-base width	cm
W_d	Depletion-layer width	cm
W_{dBE}	Base–emitter junction depletion-layer width	cm
W_{dBC}	Base–collector junction depletion-layer width	cm
W_{dm}	Maximum depletion-layer width in MOS	cm
W_E	Emitter-layer width (thickness)	cm
W_S	Source junction depletion-layer width	cm
W_D	Drain junction depletion-layer width	cm
ω	Angular frequency	rad/s
x_j	Junction depth	cm
x_c, x_i	Depth of inversion channel	cm
ψ	Potential	V
ψ_B	Difference between Fermi potential and intrinsic potential	V
ψ_{bi}	Built-in potential	V
ψ_f	Fermi potential	V
ψ_i	Intrinsic potential	V
ψ_s	Surface potential	V

1 Introduction

Since the invention of the bipolar transistor in 1947, there has been an unprecedented growth of the semiconductor industry, with an enormous impact on the way people work and live. In the last thirty years or so, by far the strongest growth area of the semiconductor industry has been in silicon very-large-scale-integration (VLSI) technology. The sustained growth in VLSI technology is fueled by the continued shrinking of transistors to ever smaller dimensions. The benefits of miniaturization – higher packing densities, higher circuit speeds, and lower power dissipation – have been key in the evolutionary progress leading to today's computers, wireless units, and communication systems that offer superior performance, dramatically reduced cost per function, and much reduced physical size, in comparison with their predecessors. On the economic side, the integrated-circuit (IC) business has grown worldwide in sales from $1 billion in 1970 to $20 billion in 1984 and has reached $250 billion in 2007. The electronics industry is now among the largest industries in terms of output as well as employment in many nations. The importance of microelectronics in economic, social, and even political development throughout the world will no doubt continue to ascend. The large worldwide investment in VLSI technology constitutes a formidable driving force that will all but guarantee the continued progress in IC integration density and speed, for as long as physical principles will allow.

1.1 Evolution of VLSI Device Technology

1.1.1 Historical Perspective

An excellent account of the evolution of the metal–oxide–semiconductor field-effect transistor (MOSFET), from its initial concept to VLSI applications in the mid 1980s, can be found in the paper by Sah (Sah, 1988). Figure 1.1 gives a chronology of the major milestone events in the development of VLSI technology. The bipolar transistor technology was developed early on and was applied to the first integrated-circuit memory in mainframe computers in the 1960s. Bipolar transistors have been used all along where raw circuit speed is most important, for bipolar circuits remain the fastest at the individual-circuit level. However, the large power dissipation of bipolar circuits has severely limited their integration level, to about 10^4 circuits per chip. This integration level is quite low by today's VLSI standard.

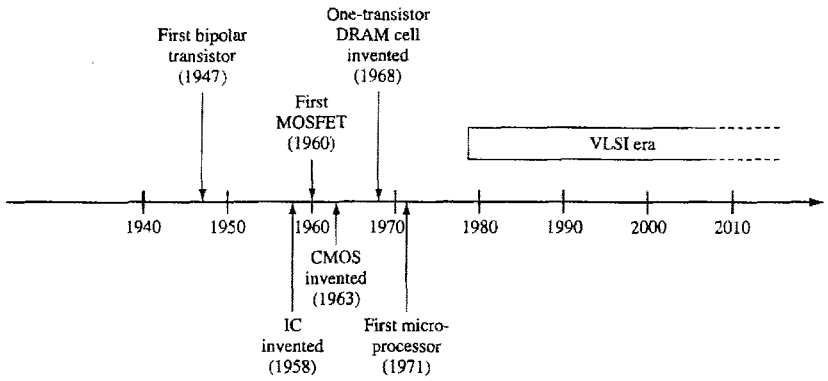


Figure 1.1. A brief chronology of the major milestones in the development of VLSI.

The idea of modulating the surface conductance of a semiconductor by the application of an electric field was first reported in 1930. However, early attempts to fabricate a surface-field-controlled device were not successful because of the presence of large densities of surface states which effectively shielded the surface potential from the influence of an external field. The first MOSFET on a silicon substrate using SiO_2 as the gate insulator was fabricated in 1960 (Kahng and Atalla, 1960). During the 1960s and 1970s, n-channel and p-channel MOSFETs were widely used, along with bipolar transistors, for implementing circuit functions on a silicon chip. Although the MOSFET devices were slow compared to the bipolar devices, they had a higher layout density and were relatively simple to fabricate; the simplest MOSFET chip could be made using only four masks and a single doping step. However, just like bipolar circuits, single-polarity MOSFET circuits suffered from large standby power dissipation, and hence were limited in the level of integration on a chip.

The major breakthrough in the level of integration came in 1963 with the invention of CMOS (complementary MOS) (Wanlass and Sah, 1963), in which n-channel and p-channel MOSFETs are constructed side by side on the same substrate. A CMOS circuit typically consists of an n-channel MOSFET and a p-channel MOSFET connected in series between the power-supply terminals, so that there is negligible standby power dissipation. Significant power is dissipated only during switching of the circuit (i.e., only when the circuits are active.) By cleverly designing the "switch activities" of the circuits on a chip to minimize active power dissipation, engineers have been able to integrate hundreds of millions of CMOS transistors on a single chip and still have the chip readily air-coolable. Until the minimum feature size of lithography reached 180 nm, the integration level of CMOS was not limited by chip-level power dissipation, but by chip fabrication technology. Another advantage of CMOS circuits comes from the ratioless, full rail-to-rail logic swing, which improves the noise margin and makes a CMOS chip easier to design.

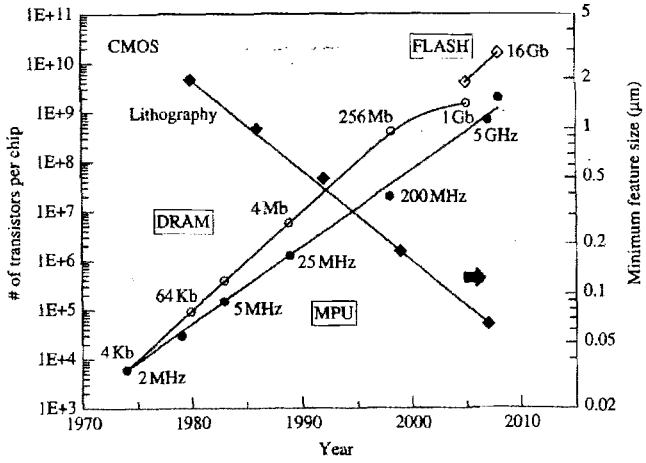


Figure 1.2. Trends in lithographic feature size, number of transistors per chip for DRAM and microprocessors (MPU), and number of memory bits per chip for Flash. The transistor count for DRAM is computed as 1.5 times the number of bits on the chip to account for the peripheral circuits. Recent data points represent announced leading edge products.

As linear dimensions reached the 0.5-μm level in the early 1990s, the performance advantage of bipolar transistors was outweighed by the significantly greater circuit density of CMOS devices. The system performance benefit of integrated functionality superseded that of raw transistor performance. Even the designers of high-end computer systems were able to meet their performance targets using CMOS instead of bipolar (Rao *et al.*, 1997). Since then, CMOS has become the technology for digital circuits, and bipolar is used primarily in radio-frequency (RF) and analog circuits only.

Advances in lithography and etching technologies have enabled the industry to scale down transistors in physical dimensions, and to pack more transistors in the same chip area. Such progress, combined with a steady growth in chip size, resulted in an exponential growth in the number of transistors and memory bits per chip. The history and recent trends in these areas are illustrated in Fig. 1.2. Traditionally, dynamic random-access memories (DRAMs) have contained the highest component count of any IC chips. This has been so because of the small size of the one-transistor memory cell (Dennard, 1968) and because of the large and often insatiable demand for more memory in computing systems. It is interesting to note that the entire content of this book can be stored in one 64-Mb DRAM chip, which was in volume production in 1997 and has an area equivalent to a square of about 1.2×1.2 cm^2.

One remarkable feature of silicon devices that fuels the rapid growth of the information technology industry is that their speed increases and their cost decreases as their size is reduced. The transistors manufactured today are 10 times faster and occupy less than 1% of the area of those built 20 years ago. This is illustrated in the trend of microprocessor units (MPUs) in Fig. 1.2. The increase in the clock frequency of microprocessors is the

result of a combination of improvements in microprocessor architecture and improvements in transistor speed.

1.1.2 Recent Developments

Since the publication of the first edition of this book in 1998, there have been major developments in the VLSI industry that are worth mentioning. These include the following.

- Up until the mid 1990s, DRAM has been the technology driver (ITRS, 1999). However, since the mid 1990s, microprocessor has replaced DRAM as the driver of VLSI technology. This shift occurred because microprocessors push the CMOS devices to shorter gate lengths and lower supply voltages and require many more wiring levels than DRAM (ITRS, 2007). The demand in microprocessor performance has spun recent research activities in high-κ gate dielectrics as a replacement for SiO_2 and in materials and device structures with enhanced transport properties. Some of the advanced features have already shown up in selected leading edge products.
- Driven by the need for low-power and light-weight data storage in battery-operated personal systems, NAND Flash (the highest density version of the electrically programmable and erasable nonvolatile memory) development has been on an exceptionally steep trajectory since the mid 1990s. Today, NAND Flash has overtaken DRAM as the IC chip with the highest component count, as shown in Fig. 1.2 (Kim, 2008).
- Two silicon derivative technologies, SOI (silicon on insulator) CMOS and SiGe bipolar, have gone into volume manufacturing. SOI CMOS is used primarily in high-end computers and interactive game systems for additional device performance. SiGe-base bipolar, with its greatly improved frequency response and analog-circuit attributes, is used in many RF and analog circuits today.
- With the bulk CMOS devices scaled to nearing their limits, researchers in the VLSI area have been exploring double-gate MOSFETs, and in general, multiple-gate MOSFETs which in principle can extend CMOS scaling to 10 nm gate lengths and below.

1.2 Modern VLSI Devices

It is clear from Fig. 1.2 that modern transistors of practical interest have feature sizes of 0.5 μm and smaller. Although the basic operation principles of large and small transistors are the same, the relative importance of the various device parameters and performance factors for the small-dimension modern transistors is quite different from that for the transistors of the early 1980s or earlier. It is our intention to focus our discussion in this book on the fundamentals of silicon devices of sub-0.5-μm generations.

1.2.1 Modern CMOS Transistors

A schematic cross section of modern CMOS transistors, consisting of an n-channel MOSFET and a p-channel MOSFET integrated on the same chip, is shown in Fig. 1.3.

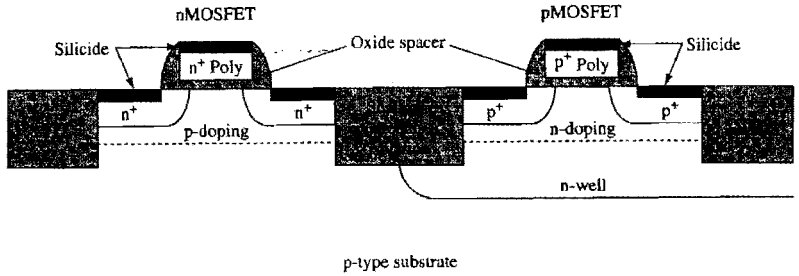


Figure 1.3. Schematic device cross section for an advanced CMOS technology.

A generic process flow for fabricating the CMOS transistors is outlined in Appendix 1. The key physical features of the modern CMOS technology, as illustrated in Fig. 1.3, include: p-type polysilicon gate for the p-channel MOSFET and n-type polysilicon gate for the n-channel MOSFET, refractory metal silicide on the polysilicon gate as well as on the source and drain diffusion regions, and shallow-trench oxide isolation.

In the electrical design of the modern CMOS transistor, the power-supply voltage is reduced with the physical dimensions in some coordinated manner. A great deal of design detail goes into decreasing the channel length, or separation between the source and drain, maximizing the on current of the transistor while maintaining an adequately low off current, minimizing variation of the transistor characteristics with process tolerances, and minimizing the parasitic resistances and parasitic capacitances.

1.2.2 Modern Bipolar Transistors

Figure 1.4 shows the schematic cross sections of two modern bipolar transistors: (a) with a Si-base and (b) with a SiGe-base. The process outline for fabricating transistor (a) is shown in Appendix 2. The salient features of the modern bipolar transistors include: shallow-trench field oxide and deep-trench isolation, polysilicon emitter, polysilicon base contact which is self-aligned to the emitter contact, and a pedestal collector which is doped to the desired level only directly underneath the emitter. A SiGe-base transistor is superior to a Si-base transistor for RF and analog circuit applications.

Unlike CMOS, the power-supply voltage for a bipolar transistor is usually kept constant as the transistor physical dimensions are reduced. Without the ability to reduce the operating voltage, electrical breakdown is a severe concern in the design of modern bipolar transistors. In designing a modern bipolar transistor, a lot of effort is spent tailoring the doping profile of the various device regions in order to maintain adequate breakdown-voltage margins while maximizing the device performance. At the same time, unlike the bipolar transistors before the early 1980s when the device performance was mostly limited by the device physical dimensions practical at the time, a modern bipolar transistor often has its performance limited by its current-density capability and

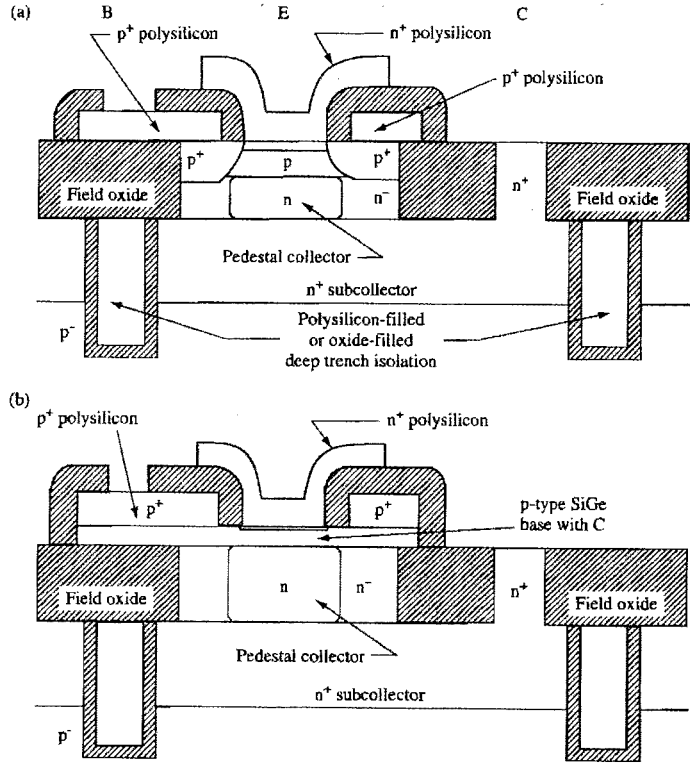


Figure 1.4. Schematic cross sections of modern silicon n–p–n bipolar transistors. (a) A transistor having a Si-base doped by ion implantation. (b) A transistor having a SiGe-base doped *in situ* with boron. Carbon is often added to suppress boron diffusion in the base layer.

not by its physical dimensions. Attempts to improve the current-density capability of a transistor usually lead to reduced breakdown voltages.

1.3 Scope and Brief Description of the Book

In writing this book, it is our goal to address the factors governing the performance of modern VLSI devices in depth. This is carried out by first discussing the device physics that goes into the design of individual device parameters, and then discussing the effects of these parameters on the performance of small-dimension modern transistors at the basic circuit level. A substantial part of the book is devoted to in-depth discussions on the interdependency among the device parameters and the subtle tradeoffs in the design of modern CMOS and bipolar transistors.

This book contains sufficient background tutorials to be used as a textbook for students taking a graduate or advanced undergraduate course in microelectronics. The prerequisite will be one semester of either solid-state physics or semiconductor physics. For the practicing engineer, this book provides an extensive source of reference material that covers the fundamentals of CMOS and bipolar technologies, devices, and circuits. It should be useful to VLSI process engineers and circuit designers interested in learning basic device principles, and to device design or characterization engineers who desire more in-depth knowledge in their specialized areas. Below is a brief description of each chapter. Two new chapters are added in the second edition: one on memory devices and the other on SOI devices.

Chapter 2: Basic Device Physics

Chapter 2 covers the appropriate level of basic device physics to make the book self-contained, and to prepare the reader with the necessary background on device operation and material physics to follow the discussion in the rest of the book.

Starting with the energy bands in silicon, Chapter 2 first introduces the basic concepts of Fermi level, carrier concentration, drift and diffusion current transport, and Poisson's equation. The next two sections focus on the most elementary building blocks of silicon devices: the p–n junction and the MOS capacitor. Basic knowledge of their characteristics is a prerequisite to further understand the operation of the VLSI devices they lead into: bipolar and MOSFET transistors. The rest of Chapter 2 covers high-field effects, $Si–SiO_2$ systems, metal–silicon contacts, hot carriers, and the physics of tunneling and breakdown relevant to VLSI device reliability.

Chapter 3: MOSFET Devices

Chapter 3 describes the basic characteristics of MOSFET devices, using the n-channel MOSFET as an example for most of the discussions. It is divided into two parts. The first part deals with the more elementary long-channel MOSFETs, including subsections on drain current models, $I–V$ characteristics, subthreshold currents, channel mobility, and intrinsic capacitances. These serve as a foundation for understanding the more important but more complex short-channel MOSFETs, which have lower capacitances and carry higher currents per gate voltage swing. The second part of Chapter 3 covers the specific features of short-channel MOSFETs important for device design purposes. The subsections include short-channel effects, velocity saturation and high-field transport, channel-length modulation, and source–drain series resistance.

Chapter 4: CMOS Device Design

Chapter 4 considers the major device design issues in a CMOS technology. It begins with the concept of MOSFET scaling – the most important guiding principle for achieving density, speed, and power improvements in VLSI evolution. Several non-scaling factors are addressed, notably, the thermal voltage and the silicon bandgap,

which have significant implications on the deviation of the CMOS evolution path from ideal scaling. Two key CMOS device design parameters – threshold voltage and channel length – are then discussed in detail. Subsections on threshold voltage include off-current requirement, choice of gate work function, channel profile design, nonuniform doping, and quantum-mechanical and discrete dopant effects on threshold voltage. Subsections on channel length include the definition of effective channel length, its extraction by the conventional method and the shift-and-ratio method, and the physical interpretation of effective channel length.

Chapter 5: CMOS Performance Factors

Chapter 5 examines the key factors that govern the switching performance and power dissipation of basic digital CMOS circuits which form the building blocks of a VLSI chip. Starting with a brief description of static CMOS logic gates, their layout and noise margin, we examine the parasitic resistances and capacitances that may adversely affect the delay of a CMOS circuit. These include source and drain series resistance, junction capacitance, overlap capacitance, gate resistance, and interconnect capacitance and resistance. Next, we formulate a delay equation and use it to study the sensitivity of CMOS delay performance to a variety of device and circuit parameters such as wire loading, device width and length, gate oxide thickness, power-supply voltage, threshold voltage, parasitic components, and substrate sensitivity in stacked circuits. The last section of Chapter 5 further extends the discussion of performance factors to several advanced CMOS materials and device structures. These include RF CMOS, effect of mobility on CMOS delay, and low-temperature CMOS.

Chapter 6: Bipolar Devices

The basic components of a bipolar transistor are described in Chapter 6. The discussion is based entirely on the vertical n–p–n transistor, since practically all high-speed bipolar transistors used in digital circuits are of the vertical n–p–n type. However, the basic device operation concept and device physics can be readily extended to other types of bipolar transistors, such as p–n–p bipolar transistors and lateral bipolar transistors.

The basic operation of a bipolar transistor is described in terms of two p–n diodes connected back to back. The basic theory of a p–n diode is modified and applied to derive the current equations for a bipolar transistor. From these current equations, other important device parameters and phenomena, such as current gain, Early voltage, base–collector junction avalanche, emitter–collector punch-through, base widening, and diffusion capacitance, are examined. Finally, the basic equivalent-circuit models relating the device parameters to circuit parameters are developed. These equivalent-circuit models form the starting point for discussing the performance of a bipolar transistor in circuit applications.

Chapter 7: Bipolar Device Design

Chapter 7 covers the basic design of a bipolar transistor. The design of the individual device regions, namely the emitter, the base, and the collector, are discussed separately. Since the detailed characteristics of a bipolar transistor depend on its operating point, the focus of this chapter is on optimizing the device design according to its intended operating condition and environment, and on the tradeoffs that must be made in the optimization process. The sections include an examination of the effect of grading the base doping profile to enhance the drift field in the intrinsic base, and a derivation of the collector-current equations when there is significant heavy doping effect in the base. In addition, the physics and characteristics of SiGe-base bipolar transistors are discussed in much greater depth than in the first edition. The chapter concludes with a discussion of the salient features of the most commonly used modern bipolar device structure.

Chapter 8: Bipolar Performance Factors

The major factors governing the performance of bipolar transistors in circuit applications are discussed in Chapter 8. Several of the commonly used figures of merit, namely, cutoff frequency, maximum oscillation frequency, and logic gate delay, are examined, and how a bipolar transistor can be optimized for a given figure of merit is discussed. Sections are devoted to examining the important delay components of a logic gate, and how these components can be minimized. The power–delay tradeoffs in the design of a bipolar transistor under various circuit-loading conditions are also examined. The scaling properties of bipolar transistors, and how the large standby power dissipation of bipolar circuits limits the integration level of bipolar chips, are discussed. A discussion of the optimization of bipolar transistors for RF and analog circuit applications is given. The chapter concludes with a discussion comparing SiGe-base bipolar transistors with GaAs heterojunction bipolar transistors.

Chapter 9: Memory Devices

In Chapter 9, the basic operational and device design principles of commonly used memory devices are discussed. The memory devices covered include CMOS SRAM, DRAM, bipolar SRAM, and several commonly used nonvolatile memories including Flash. Typical read, write, and erase operations of the various memory arrays are explained. The issue of noise margin in scaled CMOS SRAM cells is discussed.

Chapter 10: Silicon-on-Insulator Devices

The last chapter of this book deals with silicon-on-insulator (SOI) devices, which include SOI CMOS, SOI bipolar, and double-gate MOSFETs. Both partially depleted and fully depleted SOI MOSFETs and their scaling characteristics are covered. A recently developed analytic-potential model for the drain current of a symmetric double-gate MOSFET is discussed at the end.

Appendices

There are altogether 18 appendices in the back of this book, covering in more detail various topics ranging from generation and recombination, analytic short-channel threshold model, quantum mechanical solution in weak inversion, emitter and base series resistance, to unity-gain frequencies of MOSFET and bipolar transistors. They usually involve mathematical treatments too tedious and lengthy to be included in the main text. Ten of the 18 appendices are new additions to the second edition.

2 Basic Device Physics

This chapter reviews the basic concepts of semiconductor device physics. Starting with electrons and holes and their transport in silicon, we focus on the most elementary types of devices in VLSI technology: p–n junction, metal–oxide–semiconductor (MOS) capacitor, and metal-semiconductor contacts. The rest of the chapter deals with subjects of importance to VLSI device reliability: high-field effects, the Si–SiO_2 system, and dielectric breakdown.

2.1 Electrons and Holes in Silicon

The first section covers energy bands in silicon, Fermi level, n-type and p-type silicon, electrostatic potential, drift and diffusion current transport, and basic equations governing VLSI device operation. These will serve as the basis for understanding the more advanced device concepts discussed in the rest of the book.

2.1.1 Energy Bands in Silicon

The starting material used in the fabrication of VLSI devices is silicon in the crystalline form. The silicon wafers are cut parallel to either the ⟨111⟩ or ⟨100⟩ planes (Sze, 1981), with ⟨100⟩ material being the most commonly used. This is largely due to the fact that ⟨100⟩ wafers, during processing, produce the lowest charges at the oxide–silicon interface as well as higher mobility (Balk *et al.*, 1965). In a silicon crystal each atom has four valence electrons to share with its four nearest neighboring atoms. The valence electrons are shared in a paired configuration called a *covalent bond*. ***The most important result of the application of quantum mechanics to the description of electrons in a solid is that the allowed energy levels of electrons are grouped into bands*** (Kittel, 1976). ***The bands are separated by regions of energy that the electrons in the solid cannot possess: forbidden gaps***. The highest energy band that is completely filled by electrons at 0 K is called the *valence band*. The next higher energy band, separated by a forbidden gap from the valence band, is called the *conduction band*, as shown in Fig. 2.1.

2.1.1.1 Bandgap of Silicon

What sets a semiconductor such as silicon apart from a metal or an insulator is that at absolute zero temperature, the valence band is completely filled with electrons, while

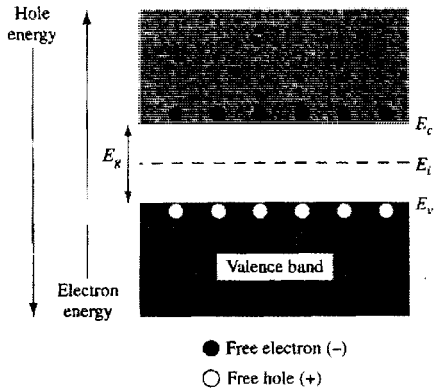


Figure 2.1. Energy-band diagram of silicon.

the conduction band is completely empty, and that the separation between the conduction band and valence band, or the *bandgap*, is on the order of 1 eV. On one hand, no electrical conduction is possible at 0 K, since there are no current carriers in the conduction band, whereas the electrons in the completely filled valence band cannot be accelerated by an electric field and gain energy. On the other hand, the bandgap is small enough that at room temperature a small fraction of the electrons are excited into the conduction band, leaving behind vacancies, or *holes*, in the valence band. This allows limited conduction to take place from the motion of both the electrons in the conduction band and the holes in the valence band. In contrast, an insulator has a much larger forbidden gap of at least several electron volts, making room-temperature conduction virtually impossible. Metals, on the contrary, have partially filled conduction bands even at absolute zero temperature, so that the electrons can gain an infinitesimal amount of energy from the applied electric field. This makes them good conductors at any temperature.

As shown in Fig. 2.1, the energy of the electrons in the conduction band increases upward, while the energy of the holes in the valence band increases downward. The bottom of the conduction band is designated E_c, and the top of the valence band E_v. Their separation, or the bandgap, is $E_g = E_c - E_v$. For silicon, E_g is 1.12 eV at room temperature or 300 K. The bandgap decreases slightly as the temperature increases, with a temperature coefficient of $dE_g/dT \approx -2.73 \times 10^{-4}$ eV/K for silicon near 300 K. Other important physical parameters of silicon and silicon dioxide are listed in Table 2.1 (Green, 1990).

2.1.1.2 Density of States

The density of available electronic states within a certain energy range in the conduction band is determined by the number of different momentum values that can be acquired by electrons in this energy range. Based on quantum mechanics, there is one allowed state in a phase space of volume $(\Delta x\,\Delta p_x)(\Delta y\,\Delta p_y)(\Delta z\,\Delta p_z) = h^3$, where p_x, p_y, p_z are the x-, y-, z-components of the electron momentum and h is Planck's constant. If we

Table 2.1 Physical Properties of Si and SiO_2 at Room Temperature (300 K)

Property	Si	SiO_2
Atomic/molecular weight	28.09	60.08
Atoms or molecules/cm^3	5.0×10^{22}	2.3×10^{22}
Density (g/cm^3)	2.33	2.27
Crystal structure	Diamond	Amorphous
Lattice constant (Å)	5.43	–
Energy gap (eV)	1.12	8–9
Dielectric constant	11.7	3.9
Intrinsic carrier concentration (cm^{-3})	1.0×10^{10}	–
Carrier mobility (cm^2/V-s)	Electron: 1430	–
	Hole: 480	–
Effective density of states (cm^{-3})	Conduction band, N_c: 2.9×10^{19}	–
	Valence band, N_v: 3.1×10^{19}	–
Breakdown field (V/cm)	3×10^5	$>10^7$
Melting point (°C)	1415	1600–1700
Thermal conductivity (W/cm-°C)	1.5	0.014
Specific heat (J/g-°C)	0.7	1.0
Thermal diffusivity (cm^2/s)	0.9	0.006
Thermal expansion coefficient (°C^{-1})	2.5×10^{-6}	0.5×10^{-6}

let $N(E)\,dE$ be the number of electronic states per unit volume with an energy between E and $E + dE$ in the conduction band, then

$$N(E)\,dE = 2g\,\frac{dp_x\,dp_y\,dp_z}{h^3}, \tag{2.1}$$

where $dp_x\,dp_y\,dp_z$ is the volume in the momentum space within which the electron energy lies between E and $E + dE$, g is the number of equivalent minima in the conduction band, and the factor of two arises from the two possible directions of electron spin. The conduction band of silicon has a sixfold degeneracy, so $g = 6$. Note that MKS units are used here (e.g., length must be in meters, not centimeters).

If the electron kinetic energy is not too high, one can consider the energy–momentum relationship near the conduction-band minima as being parabolic and write

$$E - E_c = \frac{p_x^2}{2m_x} + \frac{p_y^2}{2m_y} + \frac{p_z^2}{2m_y}, \tag{2.2}$$

where $E - E_c$ is the electron kinetic energy, and m_x, m_y, m_z are the effective masses. The constant energy surface in momentum space is an ellipsoid with the lengths of the symmetry axes proportional to the square roots of m_x, m_y, and m_z. For the silicon conduction band in the <100> direction, two of the effective masses are the transverse mass $m_t = 0.19m_0$, and the third is the longitudinal mass $m_l = 0.92m_0$, where m_0 is the free electron mass. The volume of the ellipsoid given by Eq. (2.2) in momentum space is $(4\pi/3)(8m_xm_ym_z)^{1/2}(E - E_c)^{3/2}$. Therefore, the volume $dp_x dp_y dp_z$ within which the electron energy lies between E and $E + dE$ is $4\pi(2m_xm_ym_z)^{1/2}(E - E_c)^{1/2}dE$ and Eq. (2.1) becomes

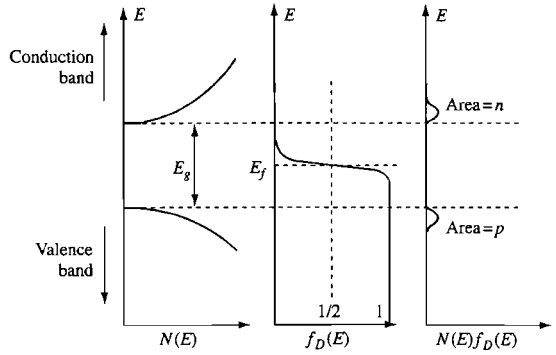


Figure 2.2. Schematic plots of density of states, Fermi–Dirac distribution function, and their products versus electron energy in a band diagram. (After Sze, 1981.)

$$N(E)\,dE = \frac{8\pi g\sqrt{2m_x m_y m_z}}{h^3}\sqrt{E - E_c}\,dE = \frac{8\pi g\sqrt{2m_t^2 m_l}}{h^3}\sqrt{E - E_c}\,dE. \quad (2.3)$$

The 3-D electron density of states in an energy diagram is then a parabolic function with its downward apex at the conduction-band edge, and vice versa for the hole density of states in the valence band. These are shown schematically in Fig. 2.2 (Sze, 1981).

2.1.1.3 Statistical Distribution Function

The energy distribution of electrons in a solid is governed by the laws of Fermi–Dirac statistics. For a system in thermal equilibrium, the principal result of these statistics is the *Fermi–Dirac distribution function*, which gives the probability that an electronic state at energy E is occupied by an electron,

$$f_D(E) = \frac{1}{1 + e^{(E - E_f)/kT}}. \quad (2.4)$$

Here $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant, and T is the absolute temperature. This function contains a parameter, E_f, called the *Fermi level*. The Fermi level is the energy at which the probability of occupation of an energy state by an electron is exactly one-half. At absolute zero temperature, $T = 0$ K, all the states below the Fermi level are filled ($f_D = 1$ for $E < E_f$), and all the states above the Fermi level are empty ($f_D = 0$ for $E > E_f$). At finite temperatures, some states above the Fermi level are filled as some states below become empty. In other words, the probability distribution $f_D(E)$ makes a smooth transition from unity to zero as the energy increases across the Fermi level. The width of the transition is governed by the thermal energy, kT. This is plotted schematically in Fig. 2.2, with a Fermi level in the middle of the forbidden gap (for reasons that will soon be clear). It is important to keep in mind that the thermal energy at room temperature is 0.026 eV, or roughly $\frac{1}{40}$ of the silicon bandgap. In most cases when the energy is at least several kT above or below the Fermi level, Eq. (2.4) can be approximated by the simple formulas

$$f_D(E) \approx e^{-(E - E_f)/kT} \quad \text{for} \quad E > E_f \quad (2.5)$$

and

$$f_D(E) \approx 1 - e^{-(E_f - E)/kT} \quad \text{for} \quad E < E_f. \quad (2.6)$$

Equation (2.6) should be interpreted as stating that the probability of finding a hole (i.e., an empty state *not* occupied by an electron) at an energy $E < E_f$ is $e^{-(E_f - E)/kT}$. The last two equations follow directly from the Maxwell–Boltzmann statistics for classical particles, which is a good approximation to the Fermi–Dirac statistics when the energy is at least several kT away from E_f.

Fermi level plays an essential role in characterizing the equilibrium state of a system. Consider two electronic systems brought into contact with Fermi levels E_{f1} and E_{f2}, and corresponding distribution functions $f_{D1}(E)$ and $f_{D2}(E)$. If $E_{f1} > E_{f2}$, then $f_{D1}(E) > f_{D2}(E)$, which means that at every energy E where electronic states are available in both systems, a larger fraction of the states in system 1 are occupied by electrons than in system 2. Equivalently, a larger fraction of the states in system 2 are empty than in system 1 at energies where electronic states exist. Since the two systems in contact are free to exchange electrons, there is a higher probability for the electrons in system 1 to re-distribute to system 2 than vice versa. This leads to a net electron transport from system 1 to system 2, i.e., current flows (defined in terms of positive charges) from system 2 to system 1. If there are no power sources connected to the systems to sustain the Fermi level imbalance, eventually the two systems will come to an equilibrium and $E_{f1} = E_{f2}$. No further net electron flow takes place once the same fractions of the electronic states in the two systems are occupied at every energy E. Note that this conclusion is reached regardless of the specific density of states in each of the two systems. For example, the two systems can be two metals, a metal and a semiconductor, two semiconductors of different doping or different composition. ***When two systems are in thermal equilibrium with no current flow between them, their Fermi levels must be equal.*** A direct extension is that, ***for a continuous region of metals and/or semiconductors in contact, the Fermi level at thermal equilibrium is flat, i.e., spatially constant, throughout the region.*** The role of Fermi level at the contacts when there is an applied voltage driving a steady-state current is further discussed in Section 2.1.4.5.

2.1.1.4 Carrier Concentration

Since $f_D(E)$ is the probability that an electronic state at energy E is occupied by an electron, the total number of electrons per unit volume in the conduction band is given by

$$n = \int_{E_c}^{\infty} N(E) f_D(E)\,dE. \quad (2.7)$$

Here the upper limit of integration (the top of the conduction band) is taken as infinity. Both the product $N(E)f_D(E)$ and n, p are shown schematically in Fig. 2.2. In general, Eq. (2.7) is a Fermi integral of the order 1/2 and must be evaluated numerically (Ghandhi, 1968). For nondegenerate silicon with a Fermi level at least $3kT/q$ below the edge of the

conduction band, the Fermi–Dirac distribution function can be approximated by the Maxwell–Boltzmann distribution, Eq. (2.5). Equation (2.7) then becomes

$$n = \frac{8\pi g\sqrt{2m_t^2 m_l}}{h^3}\int_{E_c}^{\infty}\sqrt{E-E_c}\,e^{-(E-E_f)/kT}dE. \tag{2.8}$$

With a change of variable, the integral can be expressed in the form of a gamma function, $\Gamma(3/2)$, which equals $\pi^{1/2}/2$. The electron concentration in the conduction band is then

$$n = N_c e^{-(E_c-E_f)/kT}, \tag{2.9}$$

where the pre-exponential factor is defined as the *effective density of states*,

$$N_c = 2g\sqrt{m_t^2 m_l}\left(\frac{2\pi kT}{h^2}\right)^{3/2}. \tag{2.10}$$

A similar expression can be derived for the hole density in the valence band,

$$p = N_v e^{-(E_f-E_v)/kT}, \tag{2.11}$$

where N_v is the effective density of states of the valence band, which depends on the hole effective mass and the valence band degeneracy. Both N_c and N_v are proportional to $T^{3/2}$. Their values at room temperature are listed in Table 2.1 (Green, 1990).

For an intrinsic silicon, $n = p$, since for every electron excited into the conduction band, a vacancy or hole is left behind in the valence band. The Fermi level for intrinsic silicon, or the *intrinsic Fermi level*, E_i, is then obtained by equating Eq. (2.9) and Eq. (2.11) and solving for E_f:

$$E_i = E_f = \frac{E_c+E_v}{2} - \frac{kT}{2}\ln\left(\frac{N_c}{N_v}\right). \tag{2.12}$$

By substituting Eq. (2.12) for E_f in Eq. (2.9) or Eq. (2.11), one obtains the intrinsic carrier concentration, $n_i = n = p$:

$$n_i = \sqrt{N_c N_v}\,e^{-(E_c-E_v)/2kT} = \sqrt{N_c N_v}\,e^{-E_g/2kT}. \tag{2.13}$$

Since the thermal energy, kT, is much smaller than the silicon bandgap E_g, ***the intrinsic Fermi level is very close to the midpoint between the conduction band and the valence band***. In fact, E_i is sometimes referred to as the midgap energy level, since the error in assuming E_i to be $(E_c + E_v)/2$ is only about 0.3 kT. The intrinsic carrier concentration n_i at room temperature is 1.0×10^{10} cm^{-3}, as given in Table 2.1, which is very small compared with the atomic density of silicon.

Equations (2.9) and (2.11) can be rewritten in terms of n_i and E_i:

$$n = n_i e^{(E_f-E_i)/kT}, \tag{2.14}$$

$$p = n_i e^{(E_i-E_f)/kT}. \tag{2.15}$$

These equations give the equilibrium electron and hole densities for any Fermi level position (not too close to the band edges) relative to the intrinsic Fermi level at the midgap. In the next section, we will show how the Fermi level varies with the type and concentration of impurity atoms in silicon. Since any change in E_f causes reciprocal changes in n and p, ***a useful, general relationship is that the product***

$$pn = n_i^2 \tag{2.16}$$

in equilibrium is a constant, independent of the Fermi level position.

2.1.2 n-Type and p-Type Silicon

Intrinsic silicon at room temperature has an extremely low free-carrier concentration; therefore, its resistivity is very high. In practice, intrinsic silicon hardly exists at room temperature, since it would require materials with an unobtainably high purity. Most impurities in silicon introduce additional energy levels in the forbidden gap and can be easily ionized to add either electrons to the conduction band or holes to the valence band, depending on where the impurity level is (Kittel, 1976). ***The electrical conductivity of silicon is then dominated by the type and concentration of the impurity atoms, or dopants***, and the silicon is called *extrinsic*.

2.1.2.1 Donors and Acceptors

Silicon is a column-IV element with four valence electrons per atom. There are two types of impurities in silicon that are electrically active: those from column V such as arsenic or phosphorus, and those from column III such as boron. As is shown in Fig. 2.3, a column-V atom in a silicon lattice tends to have one extra electron loosely bonded after forming covalent bonds with other silicon atoms. In most cases, the thermal energy at room temperature is sufficient to ionize the impurity atom and free the extra electron to the conduction band. Such types of impurities are called *donors*; they become positively charged when ionized. Silicon material doped with column-V impurities or donors is

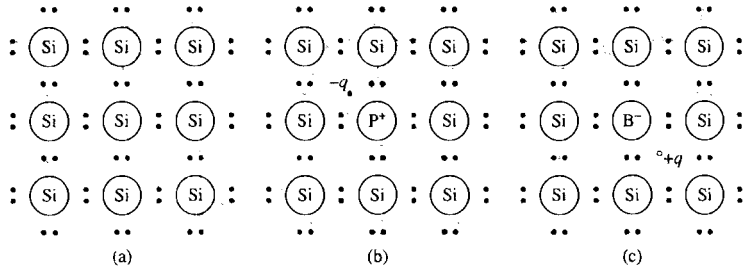


Figure 2.3. Three basic bond pictures of silicon: (a) intrinsic Si with no impurities, (b) n-type silicon with donor (phosphorus), (c) p-type silicon with acceptor (boron). (After Sze, 1981.)

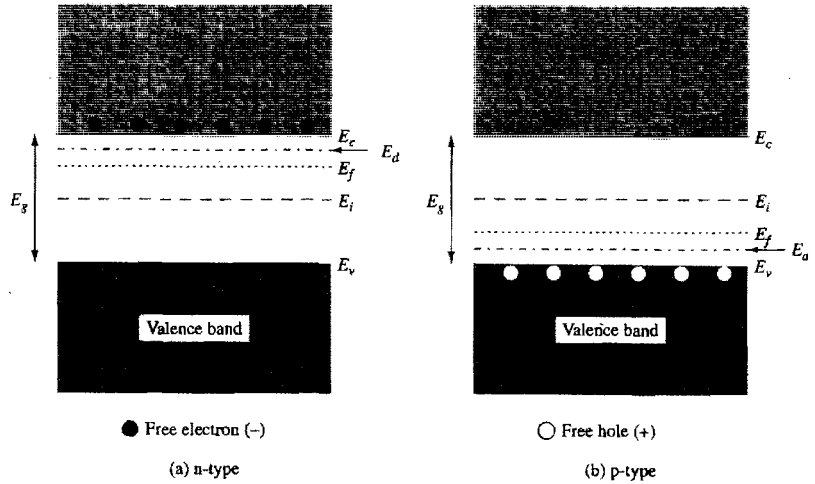


Figure 2.4. Energy-band diagram representation of (a) donor level E_d and Fermi level E_f in n-type silicon, (b) acceptor level E_a and Fermi level E_f in p-type silicon.

called *n-type* silicon, and its electrical conductivity is dominated by electrons in the conduction band. On the other hand, a column-III impurity atom in a silicon lattice tends to be deficient by one electron when forming covalent bonds with other silicon atoms (Fig. 2.3). Such an impurity atom can also be ionized by accepting an electron from the valence band, which leaves a free-moving hole that contributes to electrical conduction. These impurities are called *acceptors*; they become negatively charged when ionized. Silicon material doped with column-III impurities or acceptors is called *p-type* silicon, and its electrical conductivity is dominated by holes in the valence band. It should be noted that impurity atoms must be in a *substitutional* site (as opposed to *interstitial*) in silicon in order to be electrically active.

In terms of the energy-band diagrams in Fig. 2.4, donors add allowed electron states in the bandgap close to the conduction-band edge, while acceptors add allowed states just above the valence-band edge. Donor levels contain positive charge when ionized (emptied). Acceptor levels contain negative charge when ionized (filled). The ionization energies are denoted by $E_c - E_d$ for donors and $E_a - E_v$ for acceptors, respectively. Figure 2.5 shows the donor and acceptor levels of common impurities in silicon and their ionization energies (Sze, 1981). Phosphorus and arsenic are commonly used donors, or n-type dopants, with low ionization energies on the order of $2kT$, while boron is a commonly used acceptor or p-type dopant with a comparable ionization energy. Figure 2.6 shows the *solid solubility* of important impurities in silicon as a function of annealing temperature (Trumbore, 1960). Arsenic, boron, and phosphorus have the highest solid solubility among all the impurities, which makes them the most important doping species in VLSI technology.

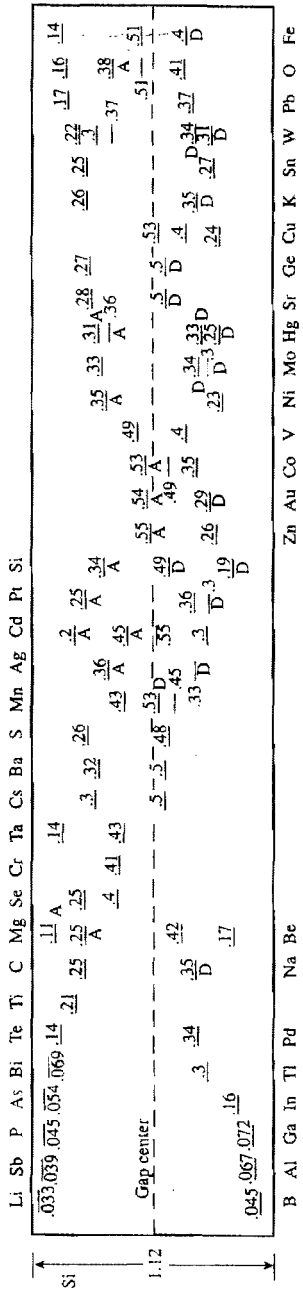


Figure 2.5. Donor and acceptor levels of various impurities in silicon. Numbers next to the level indicate ionization energies $E_c - E_d$ (donors) or $E_a - E_v$ (acceptors) in electron volts. (After Sze, 1981.)

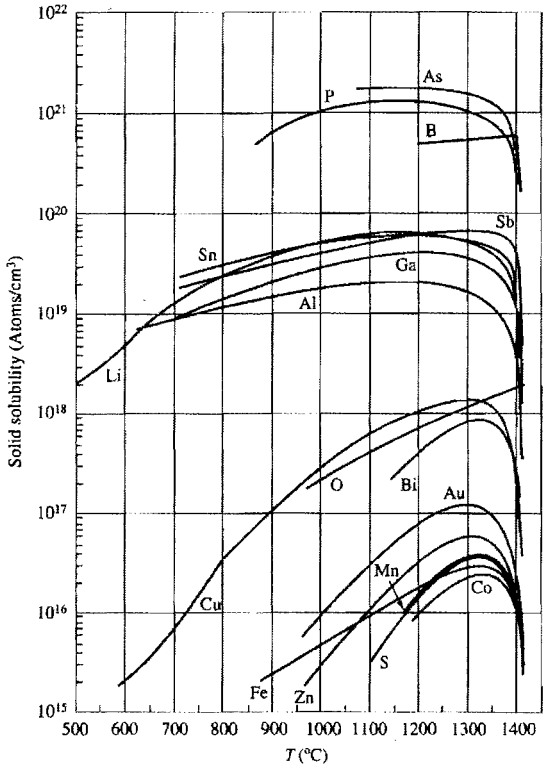


Figure 2.6. Solid solubility of various elements in silicon as a function of temperature. (After Trumbore, 1960.)

2.1.2.2 Fermi Level in Extrinsic Silicon

In contrast to intrinsic silicon, the Fermi level in an extrinsic silicon is not located at the midgap. The Fermi level in n-type silicon moves up towards the conduction band, consistent with the increase in electron density as described by Eq. (2.9). On the other hand, the Fermi level in p-type silicon moves down towards the valence band, consistent with the increase in hole density as described by Eq. (2.11). These cases are depicted in Fig. 2.4. The exact position of the Fermi level depends on both the ionization energy and the concentration of dopants. For example, for an n-type material with a donor impurity concentration N_d, the charge neutrality condition in silicon requires that

$$n = N_d^+ + p, \tag{2.17}$$

where N_d^+ is the density of ionized donors given by

$$N_d^+ = N_d[1 - f_D(E_d)] = N_d\left(1 - \frac{1}{1 + \frac{1}{2}e^{(E_d - E_f)/kT}}\right), \tag{2.18}$$

since the probability that a donor state is occupied by an electron (i.e., in the neutral state) is $f_D(E_d)$. The factor $\frac{1}{2}$ in the denominator of $f_D(E_d)$ arises from the spin degeneracy (up or down) of the available electronic states associated with an ionized donor level[1] (Ghandhi, 1968). Substituting Eq. (2.9) and Eq. (2.11) for n and p in Eq. (2.17), one obtains

$$N_c e^{-(E_c - E_f)/kT} = \frac{N_d}{1 + 2e^{-(E_d - E_f)/kT}} + N_v e^{-(E_f - E_v)/kT}, \tag{2.19}$$

which is an algebraic equation that can be solved for E_f. In n-type silicon, electrons are the majority current carriers, while holes are the minority current carriers, which means that the second term on the right-hand side (RHS) of Eq. (2.19) can be neglected. For shallow donor impurities with low to moderate concentration at room temperature, $(N_d/N_c)\exp[(E_c - E_d)/kT] \ll 1$, a good approximate solution for E_f is

$$E_c - E_f = kT\ln\left(\frac{N_c}{N_d}\right). \tag{2.20}$$

In this case, the Fermi level is at least a few kT below E_d and essentially all the donor levels are empty (ionized), i.e., $n = N_d^+ = N_d$.

It was shown earlier (Eq. (2.16)) that, in equilibrium, the product of majority and minority carrier densities equals n_i^2, independent of the dopant type and Fermi level position. The hole density in n-type silicon is then given by

$$p = n_i^2/N_d. \tag{2.21}$$

Likewise, for p-type silicon with a shallow acceptor concentration N_a, the Fermi level is given by

$$E_f - E_v = kT\ln\left(\frac{N_v}{N_a}\right), \tag{2.22}$$

the hole density is $p = N_a^- = N_a$, and the electron density is

$$n = n_i^2/N_a. \tag{2.23}$$

Figure 2.7 plots the Fermi-level position in the energy gap versus temperature for a wide range of impurity concentration (Grove, 1967). The slight variation of the silicon bandgap with temperature is also incorporated in the figure. It is seen that as the temperature increases, the Fermi level approaches the intrinsic value near midgap. When the intrinsic carrier concentration becomes larger than the doping concentration, the silicon is intrinsic. In an intermediate range of temperature including room temperature, all the donors or acceptors are ionized. The majority carrier concentration is then given by the doping concentration, independent of temperature. For temperatures below this range, *freeze-out* occurs, i.e., the thermal energy is no longer sufficient to ionize all the impurity atoms even with their shallow levels (Sze, 1981). In this case, the

[1] Detailed study showed that there are no other degeneracy with the electronic ground state in a donor except for spin (Ning and Sah, 1971).

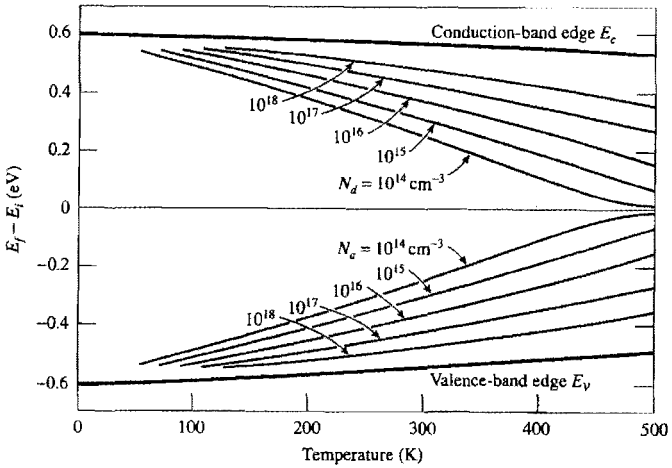


Figure 2.7. The Fermi level in silicon as a function of temperature for various impurity concentrations. (After Grove, 1967.)

majority-carrier concentration is less than the doping concentration, and one would have to solve Eq. (2.19) numerically to find E_f, n, and p (Shockley, 1950).

Instead of using N_c, N_v and referring to E_c and E_v, Eq. (2.20) and Eq. (2.22) can be written in a more useful form in terms of n_i and E_i defined by Eq. (2.12) and Eq. (2.13):

$$E_f - E_i = kT\ln\left(\frac{N_d}{n_i}\right) \tag{2.24}$$

for n-type silicon, and

$$E_i - E_f = kT\ln\left(\frac{N_a}{n_i}\right) \tag{2.25}$$

for p-type silicon. In other words, ***the distance between the Fermi level and the intrinsic Fermi level near the midgap is a logarithmic function of doping concentration***. These expressions will be used extensively throughout the book.

2.1.2.3 Fermi Level in Degenerately Doped Silicon

For heavily doped silicon, the impurity concentration N_d or N_a can exceed the effective density of states N_c or N_v, so that $E_f > E_c$ or $E_f < E_v$ according to Eq. (2.20) and Eq. (2.22). In other words, the Fermi level moves into the conduction band for n^+ silicon, and into the valence band for p^+ silicon. In addition, when the impurity concentration is higher than 10^{18}–10^{19} cm^{-3}, the donor (or acceptor) levels broaden into bands. This results in an effective decrease in the ionization energy until finally the impurity band merges with the conduction (or valence) band and the ionization energy becomes zero. Under these circumstances, the silicon is said to be *degenerate*. Strictly speaking, Fermi statistics

should be used for the electron concentration in calculation of the Fermi level when $E_c - E_f \leq kT$ (Ghandhi, 1968). For practical purposes, it is a good approximation [within $(1\text{–}2)kT$] to assume that the Fermi level of the degenerate n^+ silicon is at the conduction-band edge, and that of the degenerate p^+ silicon is at the valence-band edge.

2.1.3 Carrier Transport in Silicon

Carrier transport or current flow in silicon is driven by two different mechanisms: (a) the *drift* of carriers, which is caused by the presence of an electric field, and (b) the *diffusion* of carriers, which is caused by an electron or hole concentration gradient in silicon. The drift current will be discussed first.

2.1.3.1 Drift Current and Mobility

When an electric field is applied to a conducting medium containing free carriers, the carriers are accelerated and acquire a drift velocity superimposed upon their random thermal motion. This is described in more detail in Appendix 3. The drift velocity of holes is in the direction of the applied field, and the drift velocity of electrons is opposite to the field. The velocity of the carriers does not increase indefinitely under field acceleration, since they are scattered frequently and lose their acquired momentum after each collision. At low electric fields, the drift velocity v_d is proportional to the electric field strength $\mathscr{E}$ with a proportionality constant μ, defined as the *mobility*, in units of $cm^2/V\text{-}s$, i.e.,

$$v_d = \mu\mathscr{E}. \tag{2.26}$$

The mobility is proportional to the time interval between collisions and is inversely proportional to the effective mass of the carriers (Appendix 3). Electron and hole mobilities in silicon at low impurity concentrations are listed in Table 2.1. The electron mobility is approximately three times the hole mobility, since the effective mass of electrons in the conduction band is much lighter than that of holes in the valence band.

Figure 2.8 plots the electron and hole mobilities at room temperature versus n-type or p-type doping concentration. At low impurity levels, the mobilities are mainly limited by carrier collisions with the silicon lattice or acoustic phonons (Kittel, 1976). As the doping concentration increases beyond 10^{15}–$10^{16}/cm^3$, collisions with the charged (ionized) impurity atoms through Coulomb interaction become more and more important and the mobilities decrease. In general, one can use *Matthiessen's rule* to include different contributions to the mobility:

$$\frac{1}{\mu} = \frac{1}{\mu_L} + \frac{1}{\mu_I} + \cdots, \tag{2.27}$$

where μ_L and μ_I correspond to the lattice- and impurity-scattering-limited components of mobility, respectively. At high temperatures, the mobility tends to be limited by lattice scattering and is proportional to $T^{-3/2}$, relatively insensitive to the doping concentration (Sze, 1981). At low temperatures, the mobility is higher, but is a strong function of

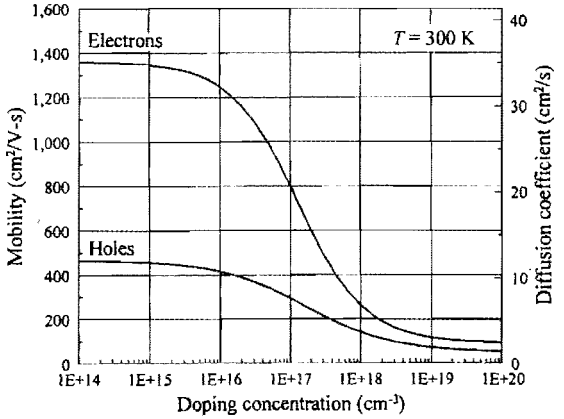


Figure 2.8. Electron and hole mobilities in bulk silicon at 300 K as a function of doping concentration.

doping concentration as it becomes more limited by impurity scattering. What is shown in Fig. 2.8 is the *bulk mobility* applicable to conduction in silicon substrates far from the surface. ***In the inversion layer of a MOSFET device, the current flow is governed by the surface mobility, which is much lower than the bulk mobility.*** This is mainly due to additional scattering mechanisms between the carriers and the Si–SiO_2 interface in the presence of high electric fields normal to the surface. Surface scattering adds another term to Eq. (2.27). Carrier mobility in the surface inversion channel of a MOSFET will be discussed in more detail in Section 3.1.5.

2.1.3.2 Resistivity

For a homogeneous n-type silicon with a free-electron density n, the drift current density under an electric field $\mathscr{E}$ is

$$J_{n,drift} = qnv_d = qn\mu_n\mathscr{E}, \tag{2.28}$$

where $q = 1.6 \times 10^{-19}$ C is the electronic charge and μ_n is the electron mobility. The resistivity, ρ_n, of n-type silicon defined by $J_{n,drift} = \mathscr{E}/\rho_n$ (a form of Ohm's law) is then given by

$$\rho_n = \frac{1}{qn\mu_n}. \tag{2.29}$$

Similarly, for p-type silicon,

$$J_{p,drift} = qp\mu_p\mathscr{E} \tag{2.30}$$

and

$$\rho_p = \frac{1}{qp\mu_p}, \tag{2.31}$$

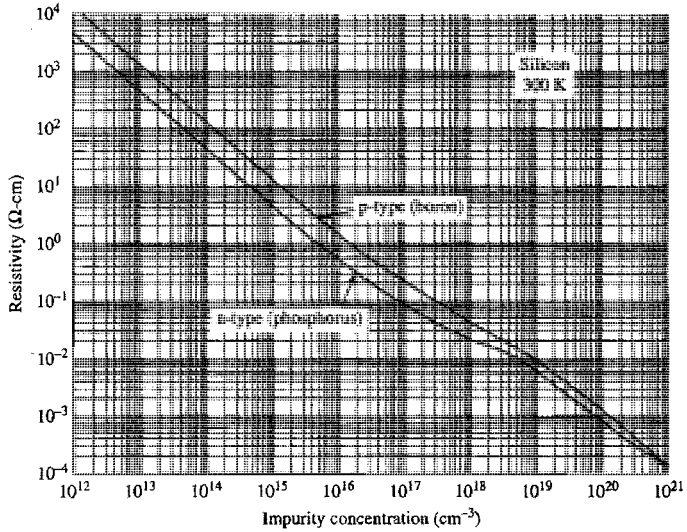


Figure 2.9. Resistivity versus impurity concentration for n-type and p-type silicon at 300 K. (After Sze, 1981.)

where μ_p is the hole mobility. In general, the total resistivity should include both the majority and the minority carrier components:

$$\rho = \frac{1}{qn\mu_n + qp\mu_p}, \tag{2.32}$$

since both electrons and holes contribute to electrical conduction. Figure 2.9 shows the measured resistivity of n-type (phosphorus-doped) and p-type (boron-doped) silicon versus impurity concentration at room temperature.

2.1.3.3 Sheet Resistivity

The resistance of a uniform conductor of length L, width W, and thickness t is given by

$$R = \rho\frac{L}{Wt}, \tag{2.33}$$

where ρ is the resistivity in ohm-centimeters. In a planar IC technology, the thickness of conducting regions is uniform and normally much less than both the length and the width of the regions. It is then useful to define a quantity, called the *sheet resistivity*, as

$$\rho_{sh} = \frac{\rho}{t} \tag{2.34}$$

in units of Ω/□ (ohms per square). Then

$$R = \frac{L}{W}\rho_{sh}, \tag{2.35}$$

i.e., the total resistance is equal to the number of squares ($L/W = 1$ is one square) of the line times the sheet resistivity. Note that sheet resistivity does not depend on the size of the square. The most common technique of measuring the sheet resistivity of a thin film is the four-point method, in which a small current is passed through the two outer probes and the voltage is measured between the two inner probes (Sze, 1981). If the spacing between the probes is much greater than the film thickness but much smaller than the linear dimension of the conducting film, the resistance measured can be approximated by $V/I = \rho_{sh}(\ln 2)/\pi \approx 0.22\rho_{sh}$, from which ρ_{sh} can be easily determined.

2.1.3.4 Velocity Saturation

The linear velocity–field relationship discussed above is valid only when the electric field is not too high and the carriers are in thermal equilibrium with the lattice. At high fields, the average carrier energy increases and carriers lose their energy by optical-phonon emission nearly as fast as they gain it from the field. This results in a decrease of the mobility as the field increases until finally the drift velocity reaches a limiting value, $v_{sat} \approx 10^7$ cm/s. This phenomenon is called *velocity saturation*.

Figure 2.10 shows the measured velocity–field relationship of electrons and holes in high-purity bulk silicon at room temperature. At low fields, the drift velocity is proportional to the field (45° slope on a log–log scale) with a proportionality constant given by the electron or the hole mobility. When the field becomes higher than 3×10^3 V/cm for electrons, velocity saturation starts to occur. The saturation velocity of holes is similar to or slightly lower than that of electrons, but saturation for holes takes place at a much higher field because of their lower mobility. For more highly doped material, low-field mobilities are lower because of impurity scattering (Fig. 2.8). However, the saturation velocity remains essentially the same, independent of impurity concentration. There is a weak dependence of v_{sat} on temperature. It decreases slightly as the temperature increases (Arora, 1993).

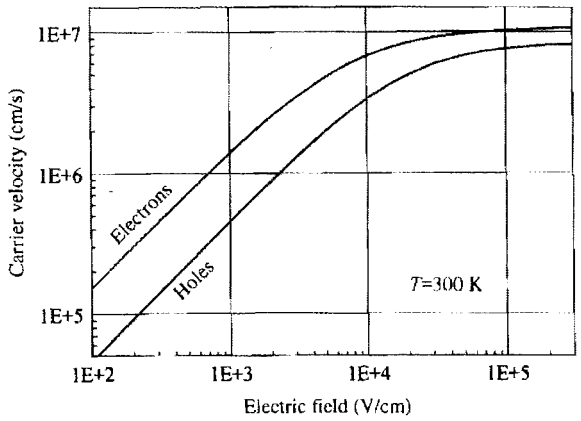


Figure 2.10. Velocity–field relationship of electrons and holes in silicon at 300 K.

2.1.3.5 Diffusion Current

The discussion in this section so far has dealt only with the case when the carrier concentration within the silicon is uniform and the carriers move under the influence of an electric field. ***If the carrier concentration is not uniform, carriers will also diffuse as a result of the concentration gradient.*** This leads to an additional current contribution in proportion to the concentration gradient:

$$J_{n,diff} = q D_n \frac{dn}{dx} \tag{2.36}$$

for electrons, and

$$J_{p,diff} = -q D_p \frac{dp}{dx} \tag{2.37}$$

for holes. The proportionality constants D_n and D_p are called the electron and hole *diffusion coefficients* and have units of cm^2/s. There is a negative sign in Eq. (2.37), since diffusion current flows in the direction of decreasing hole (positive charge) concentration.

Physically, both drift and diffusion are closely associated with the random thermal motion of carriers and their collisions with the silicon lattice in thermal equilibrium. A simple relationship between the diffusion coefficient and the mobility is derived from the basic principles in Appendix 3 (Muller and Kamins, 1977):

$$D_n = \frac{kT}{q}\mu_n \tag{2.38}$$

for electrons, and

$$D_p = \frac{kT}{q}\mu_p \tag{2.39}$$

for holes. These are known as the *Einstein relations*. The values of diffusion coefficient at room temperature can be read from Fig. 2.8 using the vertical scale on the right-hand side.

2.1.4 Basic Equations for Device Operation

2.1.4.1 Poisson's Equation

One of the key equations governing the operation of VLSI devices is *Poisson's equation*. It comes from Maxwell's first equation, which in turn is based on Coulomb's law for electrostatic force of a charge distribution. Poisson's equation is expressed in terms of the electrostatic potential, which is defined as the potential energy of carriers divided by the electronic charge q. The potential energies of carriers are either at the conduction-band edge or at the valence-band edge, as discussed before in connection with the energy-band diagram, Fig. 2.1. Since one is only interested in the spatial variation of the electrostatic potential, it can be defined with an arbitrary additive constant. It makes no difference whether E_c, E_v, or any other quantity displaced from the band edges by a fixed amount is

used to represent the potential. Conventionally, the electrostatic potential is defined in terms of the intrinsic Fermi level,

$$\psi_i = -\frac{E_i}{q}. \tag{2.40}$$

There is a negative sign because E_i is defined as electron energy while ψ_i is defined for a positive charge. The band diagram can thus be considered also as a potential diagram with the potential increasing downward, opposite to the electron energy.

The electric field $\mathscr{E}$, which is defined as the electrostatic force per unit charge, is equal to the negative gradient of ψ_i,

$$\mathscr{E} = -\frac{d\psi_i}{dx}. \tag{2.41}$$

Now we can write Poisson's equation as

$$\frac{d^2\psi_i}{dx^2} = -\frac{d\mathscr{E}}{dx} = -\frac{\rho_{net}(x)}{\varepsilon_{si}}, \tag{2.42}$$

where $\rho_{net}(x)$ is the net charge density per unit volume at x, and ε_{si} is the permittivity of silicon equal to $11.7\varepsilon_0$. Here $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm is the vacuum permittivity.

Another form of Poisson's equation is *Gauss's law*, which is obtained by integrating Eq. (2.42):

$$\mathscr{E} = \frac{1}{\varepsilon_{si}} \int \rho_{net}(x)\, dx = \frac{Q_s}{\varepsilon_{si}}, \tag{2.43}$$

where Q_s is the integrated charge density per unit area.

There are two sources of charge in silicon: *mobile charge* and *fixed charge*. Mobile charges are electrons and holes, whose densities are represented by n and p. Fixed charges are ionized donor (positively charged) and acceptor (negatively charged) atoms whose densities are represented by N_d^+ and N_a^-, respectively. Equation (2.42) can then be written as

$$\frac{d^2\psi_i}{dx^2} = -\frac{d\mathscr{E}}{dx} = -\frac{q}{\varepsilon_{si}}[p(x) - n(x) + N_d^+(x) - N_a^-(x)]. \tag{2.44}$$

For a homogeneous n-type or p-type silicon with no applied field, the RHS of Eq. (2.44) is zero and the potential is constant throughout the sample.

2.1.4.2 Dielectric Boundary Conditions

Equation (2.42) is one dimensional and is for a homogeneous material, silicon. It is adequate for describing most of the basic device operations. In some cases, e.g., in short-channel MOSFETs, the two-dimensional Poisson's equation is needed. In addition, the device could consist of two different materials with different dielectric constants, as depicted in Fig. 2.11. There are two basic boundary conditions for the components of the electric field across a dielectric interface.

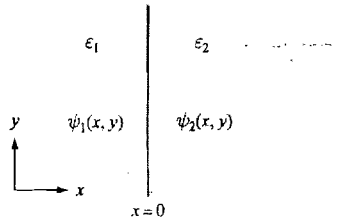


Figure 2.11. Diagram for discussing the boundary conditions of electric field at the interface between two dielectric media.

The two-dimensional Poisson's equation takes the following general form:

$$\frac{\partial(\varepsilon\mathscr{E}_x)}{\partial x} + \frac{\partial(\varepsilon\mathscr{E}_y)}{\partial y} = \rho_{net}, \tag{2.45}$$

where $\mathscr{E}_x = -\partial\psi/\partial x$ and $\mathscr{E}_y = -\partial\psi/\partial y$ are the components of the electric field perpendicular and parallel to the dielectric boundary, respectively. Note that in the geometry defined in Fig. 2.11, ε is a step function of x only, independent of y. If the potential functions in the two dielectric regions are represented by $\psi_1(x, y)$ and $\psi_2(x, y)$, they must be continuous at the interface, i.e., $\psi_1(0, y) = \psi_2(0, y)$. It follows that their y-derivatives or ***the tangential fields are also continuous***,

$$\mathscr{E}_{1y}(0, y) = \mathscr{E}_{2y}(0, y). \tag{2.46}$$

For a finite volume charge density ρ_{net}, the x-derivative in Eq. (2.45) must be finite at $x = 0$. This means that $\varepsilon\mathscr{E}_x$ must be continuous in the x-direction across the dielectric boundary. Therefore,

$$\varepsilon_1\mathscr{E}_{1x}(0, y) = \varepsilon_2\mathscr{E}_{2x}(0, y), \tag{2.47}$$

i.e., ***the perpendicular component of the displacement, $D = \varepsilon\mathscr{E}$, is continuous***. In the presence of a sheet charge at the dielectric interface, ρ_{net} can be considered as a delta function with an area equal to the interface charge density per unit area, Q_{it}. Then the last condition changes to $\varepsilon_2\mathscr{E}_{2x} - \varepsilon_1\mathscr{E}_{1x} = Q_{it}$. Interface trapped charge in an MOS capacitor is discussed in Section 2.3.6.

2.1.4.3 Carrier Concentration as a Function of Electrostatic Potential

Many of the parameters discussed before can be expressed in terms of the electrostatic potential ψ_i. For example, Eq. (2.24) and Eq. (2.25) can be combined into one equation for both n-type and p-type silicon:

$$\psi_B \equiv |\psi_f - \psi_i| = \frac{kT}{q}\ln\left(\frac{N_b}{n_i}\right), \tag{2.48}$$

where $\psi_f = -E_f/q$ is the Fermi potential and N_b is either the donor or the acceptor concentration. Equation (2.48) is a very useful expression relating the separation of the

Fermi potential from the midgap, ψ_B, to the doping concentration (or the ionized dopant concentration in the case of incomplete ionization). It is based on charge neutrality and is valid only if the local mobile carrier concentration equals the ionized dopant concentration, i.e., only if the RHS of Eq. (2.44) is zero and the field is either zero or constant.

In general, for any Fermi level position in the bandgap, carrier densities are given by Eqs. (2.14) and (2.15), which can be expressed in terms of the electrostatic potential:

$$n = n_i e^{q(\psi_i - \psi_f)/kT} \tag{2.49}$$

and

$$p = n_i e^{q(\psi_f - \psi_i)/kT}. \tag{2.50}$$

The last two equations are often referred to as *Boltzmann's relations* and are valid for either n-type or p-type silicon in thermal equilibrium. Note that Eqs. (2.49) and (2.50) are derived from Fermi level and density of states considerations, regardless of charge neutrality. They are generally applicable in the presence of net charge (due to an imbalance between the mobile and the fixed charge densities) and band bending (spatial variation of ψ_i) where $|\psi_f - \psi_i|$ is no longer given by Eq. (2.48).

2.1.4.4 Debye Length

In an inhomogeneous silicon, the doping concentration varies spatially. The bands (E_c, E_v, E_i) are not flat as in the uniformly doped case. Both the intrinsic Fermi level and the bands generally follow the doping variation according to Eq. (2.48). However, if the doping concentration changes abruptly on a very short length scale, the bands do not respond immediately, since both ψ_i and its first-order spatial derivative are continuous owing to the thermal diffusion effects of mobile carriers. The length scale in an n-type silicon, for example, can be estimated by substituting Eq. (2.49) into Eq. (2.44):

$$\frac{d^2\psi_i}{dx^2} = -\frac{q}{\varepsilon_{si}}[N_d(x) - n_i e^{q(\psi_i - \psi_f)/kT}]. \tag{2.51}$$

Without any applied bias or current, the Fermi level is spatially flat, i.e., ψ_f is independent of x, as discussed in Section 2.1.1.3. For an incremental change of doping concentration $\Delta N_d(x)$ with respect to a uniformly doped background, the corresponding change in the intrinsic potential $\Delta\psi_i(x)$ can be found by expanding the exponential term in Eq. (2.51) and keeping only the first-order term (zeroth-order terms have no spatial dependence):

$$\frac{d^2(\Delta\psi_i)}{dx^2} - \frac{q^2 N_d}{\varepsilon_{si} kT}\Delta\psi_i = -\frac{q}{\varepsilon_{si}}\Delta N_d(x). \tag{2.52}$$

This is a second-order differential equation whose solution $\Delta\psi_i$ takes the form $\exp(-x/L_D)$, where

$$L_D \equiv \sqrt{\frac{\varepsilon_{si} kT}{q^2 N_d}} \tag{2.53}$$

is called the *Debye length*. Physically, this means that ***it takes a distance on the order of L_D for the silicon bands to respond to an abrupt change in N_d.*** A small electric field is set up in this region due to the charge imbalance. The Debye length is usually much smaller than the lateral device dimension. For example, L_D = 0.04 μm for $N_d = 10^{16}$ cm^{-3}.

2.1.4.5 Current-Density Equations

The next set of equations are current-density equations. The total current density is the sum of the drift current density given by Eq. (2.28) and Eq. (2.30) and the diffusion current density given by Eq. (2.36) and Eq. (2.37). In other words,

$$J_n = qn\mu_n\mathscr{E} + qD_n\frac{dn}{dx} \tag{2.54}$$

for the electron current density, and

$$J_p = qp\mu_p\mathscr{E} - qD_p\frac{dp}{dx} \tag{2.55}$$

for the hole current density. The total conduction current density is $J = J_n + J_p$.

Using Eq. (2.41) and the Einstein relations, Eq. (2.38) and Eq. (2.39), one can write the current densities as

$$J_n = -qn\mu_n\left(\frac{d\psi_i}{dx} - \frac{kT}{qn}\frac{dn}{dx}\right) \tag{2.56}$$

and

$$J_p = -qp\mu_p\left(\frac{d\psi_i}{dx} + \frac{kT}{qp}\frac{dp}{dx}\right). \tag{2.57}$$

If both n and p take on their equilibrium values, Eqs. (2.49) and (2.50) can be substituted into the above to yield:

$$J_n = -qn\mu_n\frac{d\psi_f}{dx} \tag{2.58}$$

and

$$J_p = -qp\mu_p\frac{d\psi_f}{dx}. \tag{2.59}$$

The total current is then

$$J = J_n + J_p = -\frac{1}{\rho}\frac{d\psi_f}{dx}, \tag{2.60}$$

where ρ is the resistivity of silicon given by Eq. (2.32). Equation (2.60) resembles Ohm's law, $J_{drift} = \mathscr{E}/\rho$, discussed before. With the addition of the diffusion current, ***the total current is proportional to the gradient of the Fermi potential*** instead of proportional to the electric field, $\mathscr{E} = -d\psi_i/dx$. This reinforces and goes farther than the

concept on Fermi level and equilibrium discussed in Section 2.1.1.3: for a connected system of metals and/or semiconductors in thermal equilibrium with no current flow, the Fermi level is flat, i.e., spatially constant, throughout the system. It is important to keep in mind that Fermi level difference is the driving force for current flow, much like voltage difference drives currents in a circuit.

Strictly speaking, when current flows, the system is not in equilibrium and the Fermi level is not well defined. The electron distribution function is no longer a function of energy only. It becomes asymmetric in the current flow direction to favor population of the electronic states with a forward momentum. However, if the current is not too large and the net velocity of electron transport is small compared with the thermal velocity, there is only a slight departure from equilibrium. It is then useful to consider a *local* Fermi level, $E_f(x)$ or $\psi_f(x)$, based on the *local* equilibrium state at any given point. In this way, we generalize the Fermi level concept so that the current density equations (2.58) and (2.59) are valid as long as the *local* electron and hole densities are equal to their equilibrium values, Eqs. (2.49) and (2.50).

When an external battery or voltage source is connected to a device, it pumps all the electrons and states at one contact to a higher energy with respect to another contact. By the definition of contacts, both the electron and hole densities equal their equilibrium values at the contacts; therefore, one can define Fermi levels, e.g., E_{f1} and E_{f2}, respectively for the two contacts being considered. Without any externally applied voltage, $E_{f1} = E_{f2}$. ***When a voltage source is connected, $E_{f1} - E_{f2} = qV_{app}$, where V_{app} is the applied voltage*** with the lower voltage (higher electron energy) side connected to contact 1. It is this Fermi level imbalance at the contacts, sustained by the external power source, that drives a steady state current in the device. Fermi level will be used to reference terminal voltage and to establish the alignment relationship of electronic energy bands throughout the book.

2.1.4.6 Quasi-Fermi Potentials

The above discussion applies only when both the electron and hole densities take on their local equilibrium values and a local Fermi level can be defined. It is often in VLSI device operation to encounter nonequilibrium situations where the densities of one or both types of carriers depart from their equilibrium values given by Eqs. (2.49) and (2.50). In particular, the minority carrier concentration can be easily overwhelmed by injection from neighboring regions. This happens on a distance scale much larger than VLSI device dimensions. It results from the slow generation–recombination processes (discussed in the next section) that are inefficient to establish equilibrium between electrons and holes. Under these circumstances, while the electrons are in local equilibrium with themselves and so are the holes, electrons and holes are not in equilibrium with each other. In order to extend the kind of relationship between Fermi level and current densities discussed above, one can introduce *separate* Fermi levels for electrons and holes, respectively. They are called *quasi-Fermi levels*, E_{fn} and E_{fp}, defined so as to replace E_f in Eqs. (2.14) and (2.15):

$$n = n_i e^{(E_{fn} - E_i)/kT}, \tag{2.61}$$

$$p = n_i e^{(E_i - E_{fp})/kT}. \tag{2.62}$$

In this regard, quasi-Fermi levels have a similar physical interpretation in terms of the state occupancy as the Fermi level. ***That is, the electron density in the conduction band can be calculated as if the Fermi level is at E_{fn}, and the hole density in the valence band can be calculated as if the Fermi level is at E_{fp}.***[2] With these definitions, the current densities, Eqs. (2.56) and (2.57), become

$$J_n = -qn\mu_n \frac{d\phi_n}{dx} \tag{2.63}$$

and

$$J_p = -qp\mu_p \frac{d\phi_p}{dx}, \tag{2.64}$$

where the *quasi-Fermi potentials* ϕ_n and ϕ_p are defined by

$$\phi_n \equiv -\frac{E_{fn}}{q} = \psi_i - \frac{kT}{q}\ln\left(\frac{n}{n_i}\right) \tag{2.65}$$

and

$$\phi_p \equiv -\frac{E_{fp}}{q} = \psi_i + \frac{kT}{q}\ln\left(\frac{p}{n_i}\right). \tag{2.66}$$

Equations (2.63) and (2.64) are more generally applicable than Eqs. (2.58) and (2.59). They show that electron and hole currents are driven by separate entities: ***the gradient of electron quasi-Fermi potential drives the electron current, and the gradient of hole quasi-Fermi potential drives the hole current.*** When current flows, $E_{fn}(x)$ and $E_{fp}(x)$ (or $\phi_n(x)$ and $\phi_p(x)$) should be interpreted in the same sense of *local* equilibrium at point x as with the case of local $E_f(x)$ discussed earlier.

When electrons and holes are not in equilibrium with each other, the pn product is given by

$$pn = n_i^2 e^{q(\phi_p - \phi_n)/kT}. \tag{2.67}$$

It equals n_i^2 when $\phi_p = \phi_n = \psi_f$. Quasi-Fermi potentials are used extensively in the rest of the book for current calculations.

2.1.4.7 Continuity Equations

The next set of equations are *continuity equations* based on the conservation of mobile charge:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - R_n + G_n \tag{2.68}$$

[2] While it appears to be physically inconsistent to have half of the electron states occupied at one energy and half of the electron states empty at a different energy, quasi-Fermi levels are defined mainly for mathematical convenience.

and

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - R_p + G_p, \tag{2.69}$$

where G_n and G_p are the electron and hole generation rates, R_n and R_p are the electron and hole recombination rates, and $\partial J_n/\partial x$ and $\partial J_p/\partial x$ are the net flux of mobile charges in and out of x. At thermal equilibrium, the generation rate is equal to the recombination rate and $np = n_i^2$. When excess minority carriers are injected by light or other means, the recombination rate exceeds the generation rate, which establishes a tendency to return to equilibrium.

In silicon, the probability of direct band-to-band recombination by a *radiative* (transfer of energy to a photon) or *Auger* (transfer of energy to another carrier) process is very low due to its indirect bandgap. Most of the recombination processes take place indirectly via a trap or a deep impurity level near the middle of the forbidden gap. This is often referred to as the *Shockley–Read* recombination (Shockley and Read, 1952). Under low-injection conditions, the recombination rate is inversely proportional to the *minority-carrier lifetime*, τ, which is in the range of 10^{-4} to 10^{-9} s, depending on the quality of the silicon crystal. The *minority-carrier diffusion length*, which is the average distance a minority carrier travels before it recombines with a majority carrier, is given by $L = (D\tau)^{1/2}$, where D is the diffusion coefficient. The diffusion length is typically a few microns to a few millimeters in silicon. (A discussion of the minority-carrier diffusion process can be found in Section 2.2.4.) ***Since L is much larger than the active dimensions of a VLSI device, generation–recombination in general plays very little role in device operation.*** Only in a few special circumstances, such as CMOS latch-up, the SOI floating-body effect, junction leakage current, and radiation-induced soft error, must the generation–recombination mechanism be taken into account.

More detailed discussions on generation and recombination can be found in Appendix 5. In the steady state, $\partial n/\partial t = \partial p/\partial t = 0$. Also, the net electron reduction rate must equal the net hole reduction rate, i.e., $R_n - G_n = R_p - G_p$, so that there is no buildup of net immobile charge with time at any point. Subtracting Eq. (2.69) from Eq. (2.68) then yields $\partial(J_n + J_p)/\partial x = 0$, or continuity of the total current, $J_n + J_p$. In a device region where generation and recombination are negligible, the continuity equations in the steady state are reduced to $dJ_n/dx = dJ_p/dx = 0$, which simply states the conservation of electron current and conservation of hole current, respectively.

2.1.4.8 Dielectric Relaxation Time

In contrast to the minority-carrier lifetime discussed above, the majority-carrier response time is very short in a semiconductor. It can be estimated for a one-dimensional (1-D) homogeneous n-type silicon as follows. Suppose there is a local perturbation in the majority carrier density, Δn. From Poisson's equation, the resulting charge imbalance sets up a field divergence, $\partial \mathscr{E}/\partial x = -q\Delta n/\varepsilon_{si}$, around the point of perturbation. This, in turn, leads to a divergent current according to Ohm's law, $J_n = \mathscr{E}/\rho_n$, which tends to restore the majority carrier concentration back to its equilibrium, charge neutral valuc. Neglecting R_n and G_n in the continuity equation, Eq. (2.68), one then obtains

$$\frac{\partial \Delta n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} = \frac{1}{q\rho_n}\frac{\partial \mathscr{E}}{\partial x} = -\frac{\Delta n}{\rho_n \varepsilon_{si}}. \tag{2.70}$$

The solution to this equation takes the form of $\Delta n(t) \propto \exp(-t/\rho_n\varepsilon_{si})$, where $\rho_n\varepsilon_{si}$ is the majority-carrier response time, or *dielectric relaxation time*. ***The majority-carrier response time in silicon is typically on the order of 10^{-12} s, which is shorter than most device switching times.***

Note that $\rho_n\varepsilon_{si}$ is the minimum response time for an ideal 1-D case without any parasitic capacitances. In practice, the majority-carrier response time may be limited by the *RC* delay of the specific silicon device structure and contacts.

2.2 p–n Junctions

p–n junctions, also called *p–n diodes*, are important devices as well as important components of all MOSFET and bipolar devices. The characteristics of p–n diodes are therefore important in determining the characteristics of VLSI devices and circuits. A p–n diode is formed when one region of a semiconductor substrate is doped n-type and an immediately adjacent region is doped p-type. In practice, a silicon p–n diode is usually formed by counterdoping a local region of a larger region of doped silicon. For instance, a region of a p-type silicon substrate or "well" can be counterdoped with n-type impurities to form the n-type region of a p–n diode. The n-type region thus formed has a donor concentration higher than its acceptor concentration.

A doped semiconductor region is called *compensated* if it contains both donor and acceptor impurities such that neither impurity concentration is negligible compared to the other. For a compensated semiconductor region, it is the *net* doping concentration, i.e., $N_d - N_a$ if it is n-type and $N_a - N_d$ if it is p-type, that determines its Fermi level and its mobile carrier concentration. However, for simplicity, we shall derive the characteristics and behavior of p–n diodes assuming none of the doped regions are compensated, i.e., the n-sides of the diodes have a net donor concentration of N_d and the p-sides have a net acceptor concentration of N_a. The resultant equations can be extended to diodes with compensated doped regions simply by replacing N_d by $N_d - N_a$ for the n-regions and replacing N_a by $N_a - N_d$ for the p-regions.

2.2.1 Energy-Band Diagrams for a p–n Diode

It was shown in Sections 2.1.1.3 and 2.1.4.5 that at thermal equilibrium or when there is no net electron or hole current, the Fermi level is spatially constant. Furthermore, when an external voltage V_{app} is connected to two contacts of a piece of silicon, the Fermi level at the lower voltage contact is shifted relative to the Fermi level at the higher voltage contact by qV_{app}. In this section, we apply these results to establish the energy-band diagrams for a p–n diode under various bias conditions.

Consider a p-silicon region and an n-silicon region physically separate from each other. As discussed in Section 2.1.2.2, the Fermi level for a p-type silicon lies close to its

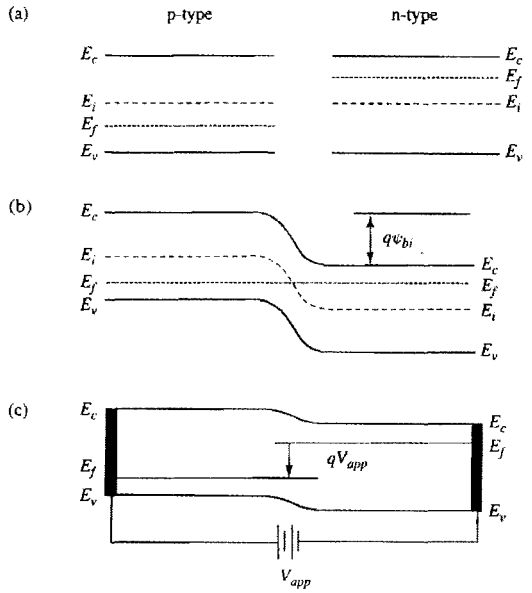


Figure 2.12. Energy-band diagrams for a p–n diode. (a) A p-silicon region and an n-silicon region physically separate from each other. (b) A p–n junction at thermal equilibrium. (c) A p–n diode connected to a battery, with the n-side connected to the negative end and the p-side connected to the positive end of the battery. The solid vertical bars represent the ohmic contacts of the p- and n-regions. For simplicity and clarity of the figure, E_i is not shown.

valence band, and that for an n-type silicon lies close to its conduction band. The energy-band diagrams for the two silicon pieces are illustrated schematically in Fig. 2.12(a).

If the p-silicon and the n-silicon are brought together to form a p–n diode, the resulting energy-band diagram is as shown in Fig. 2.12(b). At thermal equilibrium, the Fermi level must remain flat across the entire p–n diode structure, causing the energy bands of the p-region to lie higher than those of the n-region. Near the physical junction, the energy-bands are bent in order to maintain energy-band continuity between the p-region and the n-region. The band bending implies an electric field, $\mathscr{E} = -d\psi_i/dx$, in this transition region. This electric field causes a drift component of electron and hole currents to flow. At thermal equilibrium, this drift-current component is exactly balanced by a diffusion component of electron and hole currents flowing in the opposite direction caused by the large electron and hole concentration gradients across the junction. The net result is zero electron and hole currents across the p–n junction at thermal equilibrium. On both sides of the band-bending region, the energy bands are flat and there is no electric field. These regions are referred to as the *quasineutral regions*.

If a battery of voltage V_{app} is connected to the diode, with the p-side connected to the positive end of the battery and the n-side connected to the negative end of the battery, the

Fermi level at the n-side contact becomes shifted by qV_{app} relative to the Fermi level at the p-side contact. This is illustrated in Fig. 2.12(c).

2.2.1.1 Built-in Potential

Consider the energy-band diagram in Fig. 2.12(b). The difference between the energy bands on the p-side and the corresponding energy bands on the n-side is E_c(p-side) − E_c (n-side) = $q\psi_{bi}$, where ψ_{bi} is the *built-in potential* of the p–n junction. In this subsection, we want to establish the relationship between ψ_{bi} and the p- and n-side doping concentrations.

To facilitate description of both the n-side and the p-side of a diode simultaneously, when necessary for clarity, we shall distinguish the parameters on the n-side from the corresponding ones on the p-side by adding a subscript n to the symbols associated with the parameters on the n-side, and a subscript p to the symbols associated with the parameters on the p-side (Shockley, 1950). For example, E_{fn} and E_{in} denote the Fermi level and intrinsic Fermi level, respectively, on the n-side, and E_{fp} and E_{ip} denote the Fermi level and intrinsic Fermi level, respectively, on the p-side. Similarly, n_n and p_n denote the electron concentration and hole concentration, respectively, on the n-side, and n_p and p_p denote the electron concentration and hole concentration, respectively, on the p-side. Thus, n_n and p_p signify majority-carrier concentrations, while n_p and p_n signify minority-carrier concentrations.

Consider the n-side of a p–n diode at thermal equilibrium. If the n-side is non-degenerately doped to a concentraon of N_d, then the separation between its Fermi level, which is flat across the diode, and its intrinsic Fermi level is given by Eq. (2.24), namely

$$E_{fn} - E_{in} = kT\ln\left(\frac{n_{n0}}{n_i}\right) = kT\ln\left(\frac{N_d}{n_i}\right), \tag{2.71}$$

where n_{n0} denotes the n-side electron concentration at thermal equilibrium.

Similarly, for the nondegenerately doped p-side of a p–n diode at thermal equilibrium, with a doping concentration of N_a, we have

$$E_{ip} - E_{fp} = kT\ln\left(\frac{p_{p0}}{n_i}\right) = kT\ln\left(\frac{N_a}{n_i}\right), \tag{2.72}$$

where p_{p0} is the p-side hole concentration at thermal equilibrium.

The built-in potential across the p–n diode is

$$q\psi_{bi} = E_{ip} - E_{in} = kT\ln\left(\frac{n_{n0}p_{p0}}{n_i^2}\right). \tag{2.73}$$

Since Eq. (2.16) gives $n_{n0}p_{n0} = n_{p0}p_{p0} = n_i^2$, Eq. (2.73) can also be written as

$$q\psi_{bi} = kT\ln\left(\frac{p_{p0}}{p_{n0}}\right) = kT\ln\left(\frac{n_{n0}}{n_{p0}}\right), \tag{2.74}$$

which relates the built-in potential to the electron and hole densities on the two sides of the p–n diode.

2.2.2 Abrupt Junctions

Analysis of a p–n diode is much simpler if the junction is assumed to be abrupt, i.e., the doping impurities are assumed to change abruptly from p-type on one side to n-type on the other side of the junction. The abrupt-junction approximation is reasonable for modern VLSI devices, where the use of ion implantation for doping the junctions, followed by low-thermal-cycle diffusion and/or annealing, results in junctions that are fairly abrupt. Besides, the abrupt-junction approximation often leads to closed-form solutions which render the device physics much easier to understand.

2.2.2.1 Depletion Approximation

The spatial dependence of the electrostatic potential $\psi_i(x)$ is governed by Poisson's equation, i.e., Eq. (2.44). For a diode at thermal equilibrium, the electron and hole densities, $n(x)$ and $p(x)$, are given by Eqs. (2.49) and (2.50), respectively. As suggested in Fig. 2.12(b), $\psi_i(x)$ is independent of x in the uniformly doped quasineutral regions. Within the band-bending region, $\psi_i(x)$ changes from being $-E_{ip}/q$ at the p-region end of the band-bending region to being $-E_{in}/q$ at the n-region end of the band-bending region. Within the band-bending region, Eq. (2.49) suggests that the electron density drops very rapidly as $\psi_i(x)$ changes, being equal to the ionized donor density at the n-region end and dropping 10 × at room temperature for every 60 mV change in $\psi_i(x)$. Thus, the density of electrons within the band-bending region is negligible compared to the density of ionized donors except for a very narrow region adjacent the quasineutral n-region where $q(\psi_i - \psi_f)$ is less than about $3kT$. Similarly, Eq. (2.50) suggests that, within the band-bending region, the density of holes is negligible compared to the density of ionized acceptors except for a very narrow region adjacent to the quasineutral p-region.

A closed-form solution to Poisson's equation can be obtained if the electron and hole densities are assumed to be negligible in the entire band-bending region. This is called the *depletion approximation*. In this case, the abrupt junction is approximated by three regions as illustrated in Fig. 2.13(a). Both the quasineutral p-region, i.e., the region with $x < -x_p$ and the quasineutral n-region, i.e., the region with $x > x_n$, are assumed to be charge-neutral, while the transition region, i.e., the region with $-x_p < x < x_n$, is assumed to be depleted of mobile electrons and holes. As we shall show later, the depletion-layer widths, x_p and x_n, are dependent on the donor concentration N_d on the n-side and the acceptor concentration N_a on the p-side, as well as on the applied voltage V_{app} across the junction. The depletion approximation is quite accurate for all applied voltages except at large forward biases, where the mobile-charge densities are not negligible compared to the ionized impurity concentrations in the transition region. The transition region is often referred to as the *depletion region* or *depletion layer*. Since the transition region is not charge-neutral, it is also referred to as the *space-charge region* or *space-charge layer*.

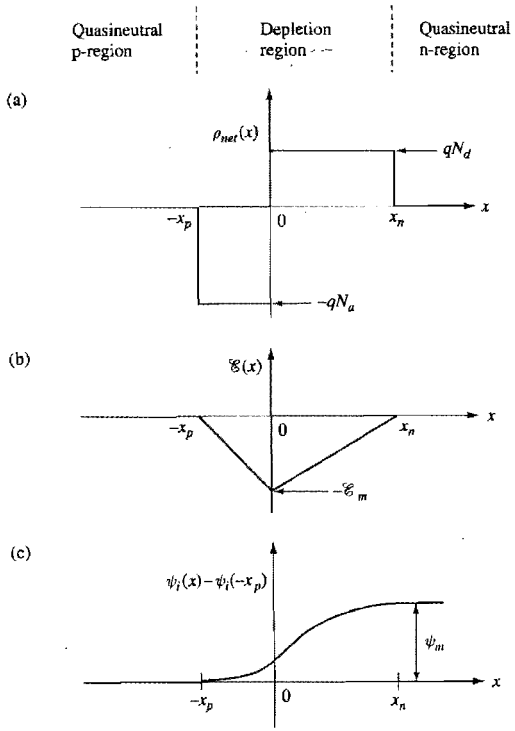


Figure 2.13. Depletion approximation of a p–n junction: (a) charge distribution, (b) electric field, and (c) electrostatic potential.

Poisson's equation, i.e., (2.44), for the depletion region is

$$-\frac{d^2\psi_i}{dx^2} = \frac{d\mathscr{E}}{dx} = \frac{q}{\varepsilon_{si}}[p(x) - n(x) + N_d^+(x) - N_a^-(x)] = \frac{q}{\varepsilon_{si}}[N_d^+(x) - N_a^-(x)], \tag{2.75}$$

where N_d^+ is the ionized-donor concentration and N_a^- is the ionized-acceptor concentration, and where the mobile-electron and -hole concentrations have been set to zero, consistent with the depletion approximation. For simplicity, we shall assume that all the donors and acceptors within the depletion region are ionized, and that the junction is abrupt and not compensated, i.e., there are no donor impurities on the p-side and no acceptor impurities on the n-side. With these assumptions, Eq. (2.75) becomes

$$-\frac{d^2\psi_i}{dx^2} = \frac{qN_d}{\varepsilon_{si}} \quad \text{for} \quad 0 \le x \le x_n \tag{2.76}$$

and

$$-\frac{d^2\psi_i}{dx^2} = -\frac{qN_a}{\varepsilon_{si}} \quad \text{for} \quad -x_p \le x \le 0. \tag{2.77}$$

Integrating Eq. (2.76) once from $x = 0$ to $x = x_n$, and Eq. (2.77) once from $x = -x_p$ to $x = 0$, subject to the boundary conditions of $d\psi_i/dx = 0$ at $x = -x_p$ and at $x = x_n$, we obtain the maximum electric field, $\mathscr{E}_m$, which is located at $x = 0$. That is,

$$\mathscr{E}_m \equiv \left|\frac{-d\psi_i}{dx}\right|_{x=0} = \frac{qN_d x_n}{\varepsilon_{si}} = \frac{qN_a x_p}{\varepsilon_{si}}. \tag{2.78}$$

It is clear from Eq. (2.78) that the total space charge inside the n-side of the depletion region is equal (but opposite in sign) to the total space charge inside the p-side of the depletion region. Thus, in Fig. 2.13(a), the two charge distribution plots have the same area. Equation (2.78) could have been obtained directly from Gauss's law, i.e., Eq. (2.43).

Let ψ_m be the total potential drop across the p–n junction, i.e., $\psi_m = [\psi_i(x_n) - \psi_i(-x_p)]$. The total potential drop can be obtained by integrating Eqs. (2.76) and (2.77) twice, the second time from $x = -x_p$ to $x = x_n$. That is,

$$\begin{aligned}\psi_m &= \int_{-x_p}^{x_n} d\psi_i(x) = -\int_{-x_p}^{x_n} \mathscr{E}(x)dx \\ &= \frac{\mathscr{E}_m(x_n + x_p)}{2} = \frac{\mathscr{E}_m W_d}{2},\end{aligned} \tag{2.79}$$

where $W_d = x_n + x_p$ is the total width of the depletion layer. It can be see from Eq. (2.79) that ψ_m is equal to the area in the $\mathscr{E}(x) - x$ plot, i.e., Fig. 2.13(b). Eliminating $\mathscr{E}_m$ from Eqs. (2.78) and (2.79) gives

$$W_d = \sqrt{\frac{2\varepsilon_{si}(N_a + N_d)\psi_m}{qN_a N_d}}. \tag{2.80}$$

This equation relates the total width of the depletion layer to the total potential drop across the junction and to the doping concentrations of the two sides of the diode.

2.2.2.2 Externally Biased Junctions

In the absence of any externally applied voltage, the total electrostatic potential drop ψ_m across a p–n diode is equal to the built-in potential ψ_{bi}, as indicated in Fig. 2.12(b). This built-in potential represents an energy barrier limiting the flow of electrons from the n-side to the p-side and the flow of holes from the p-side to the n-side. An externally applied voltage across a p–n diode shifts the Fermi level at the n-region contact relative to the Fermi level at the p-region contact. If the applied voltage causes ψ_m to be reduced, the diode is said to be *forward biased*. If the applied voltage causes ψ_m to be increased, the diode is said to be *reverse biased*. In considering a p–n diode in the context of VLSI devices, the forward-bias characteristics are more interesting than the reverse-bias characteristics. Therefore, we shall adopt the convention where a positive applied voltage also means a forward-bias voltage. Physically, this means the external voltage is connected such that

the p-side is biased positively relative to the n-side, as in the case illustrated in Fig. 2.12(c). The total potential drop ψ_m and the externally applied voltage V_{app} are related by

$$\psi_m = \psi_{bi} - V_{app}, \tag{2.81}$$

where $V_{app} > 0$ means the diode is forward biased and $V_{app} < 0$ means the diode is reverse biased. If Eq. (2.81) is used in Eq. (2.80), it gives the total depletion-layer width of a forward- or reverse-biased diode.

A quasineutral region has a finite resistivity determined by its dopant impurity concentration (see Fig. 2.9). When a current flows in a region of finite resistivity, there is a corresponding voltage drop, or *IR* drop, along the current path. In writing Eq. (2.81), the *IR* drops in the quasineutral regions are assumed to be negligible so that V_{app} is the same as the *voltage across the space-charge region, V'_{app}*. ***If IR drops in the quasineutral regions are not negligible, then V_{app} should be replaced by V'_{app} in Eq. (2.81).***

- *p–n diode as a rectifier.* When a diode is forward biased, the energy barrier limiting current flow is lowered, causing electrons to be injected from the n-side into the p-side and holes injected from the p-side into the n-side, resulting in a current flow through the diode. As we shall show in Section 2.2.4, the forward current increases exponentially with V'_{app} and hence can be very large. When a diode is reverse biased, the energy barrier limiting current flow is increased. There is no current flow due to electron and hole injection, only a relatively low background or leakage current. Thus a diode has rectifying current–voltage characteristics, being conducting when it is forward biased, and nonconducting when it is reverse biased. This is illustrated in Fig. 2.14. The equations governing the current–voltage characteristics of a diode will be derived in Sections 2.2.3 and 2.2.4.
- *Depletion-layer capacitance.* Consider a small change dV_{app} in the applied voltage. dV_{app} causes a charge per unit area dQ to flow into the p-side, which is equal to the change in the charge in the p-side depletion region. Since all mobile carriers are

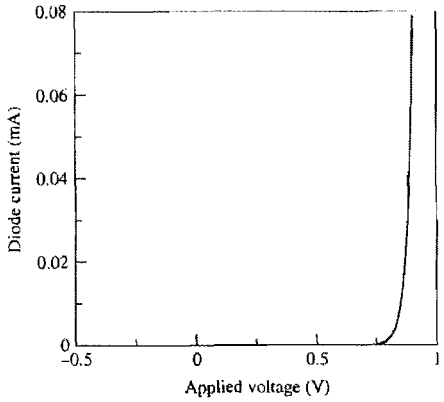


Figure 2.14. A schematic linear plot of the current of a typical silicon diode as a function of its applied voltage. On a linear plot, the reverse current is too low to be observable.

ignored in our depletion approximation, we can write the charge per unit area in the p-side depletion region as

$$Q_d(\text{p-side}) = -qN_a x_p(V_{app}), \tag{2.82}$$

where we have indicated that the p-side depletion-layer width, x_p, is a function of V_{app}. Notice that Q_d for the p-side is negative because ionized acceptors have a charge $-q$. The depletion-layer capacitance per unit area is

$$C_d \equiv \frac{dQ}{dV_{app}} = \frac{dQ_d(\text{p-side})}{dV_{app}} = \frac{\varepsilon_{si}}{W_d}. \tag{2.83}$$

That is, the depletion-layer capacitance of a diode is equivalent to a parallel-plate capacitor of separation W_d and dielectric constant ε_{si}. Physically, this is due to the fact that only the majority carriers at the edges of the depletion layer, not the space charge within the depletion region, respond to changes in the applied voltage.

- *Extending the depletion approximation to include injected current flows in the space-charge region.* When a diode is forward biased, the electrons flowing from the n-side to the p-side and the holes flowing from the p-side to the n-side add to the space charge in the transition region of the diode. To be accurate, we cannot assume the transition region to be depleted of mobile charge carriers. However, as long as the density of mobile carriers is small compared to the densities of ionized donors and acceptors, we have a well-defined space-charge region. (When the density of mobile carriers is comparable to or larger than the densities of ionized donors or acceptors, the boundaries of the space-charge region are no longer well defined. This situation will be discussed further in Section 6.3.3.2 in the context of base widening at high injection in a bipolar transistor.) For a well-defined space-charge layer of width W_d, the associated capacitance per unit area is the same as a parallel-plate capacitor, namely, Eq. (2.83). In this case, W_d can be obtained from integrating Poisson's equation, i.e., Eq. (2.44). An example of how mobile charge carriers flowing through a space-charge region affect the space-charge-region thickness is given in Section 6.3.3.1 in the context of base widening at low injection in a bipolar transistor.

2.2.2.3 One-Sided Junctions

In many applications, such as the source or drain junction of a MOSFET or the emitter–base diode of a bipolar transistor, one side of the p–n diode is degenerately doped while the other side is lightly to moderately doped. In this case, practically all the voltage drop and the depletion layer occur across the lightly doped side of the diode. That this is the case can be inferred readily from Eq. (2.78), which implies that $x_n = N_a W_d/(N_a + N_d)$ and $x_p = N_d W_d/(N_a + N_d)$. The characteristics of a one-sided p–n diode are therefore determined primarily by the properties of the lightly doped side alone. In this sub-subsection, we shall derive the equations for an n^+–p diode where the characteristics are determined by the p-side. The results can be extended straightforwardly to a p^+–n diode.

As discussed in Section 2.1.2, for a lightly to moderately doped p-type silicon, the Fermi level is given by Eq. (2.25), and for a heavily or degenerately doped n-type silicon,

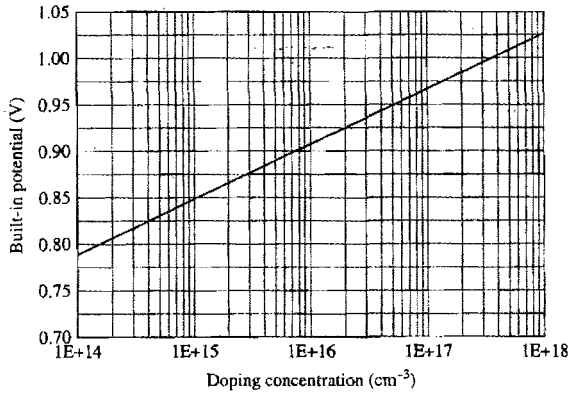


Figure 2.15. Built-in potential for a one-sided p–n junction versus the doping concentration of the lightly doped side.

it is a good approximation to assume its Fermi level to be at the conduction-band edge. Therefore, the built-in potential for an n^+–p diode, from Eqs. (2.72) and (2.73), is given by

$$\begin{aligned} q\psi_{bi} &= E_{fn} - E_{in} + kT\ln\left(\frac{N_a}{n_i}\right) \\ &\approx E_{cn} - E_{in} + kT\ln\left(\frac{N_a}{n_i}\right) \\ &\approx \frac{E_g}{2} + kT\ln\left(\frac{N_a}{n_i}\right), \end{aligned} \tag{2.84}$$

where we have made a further approximation that the intrinsic Fermi level is located half way between the conduction- and valence-band edges, E_{cn} and E_{vn}, on the n-side. [See Eq. (2.12) and the discussion that follows.] Figure 2.15 is a plot of ψ_{bi}, as approximated by Eq. (2.84), as a function of the doping concentration of the lightly doped side.

The depletion-layer width, from Eqs. (2.80) and (2.81), is

$$W_d = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} - V_{app})}{qN_a}}, \tag{2.85}$$

where $V_{app} > 0$ if the diode is forward biased and $V_{app} < 0$ if the diode is reverse biased. The depletion-layer capacitance per unit area is given by Eq. (2.83). Figure 2.16 is a plot of the depletion-layer width and capacitance as a function of doping concentration for $V_{app} = 0$. Again, V_{app} in Eq. (2.85) should be replaced by V'_{app} whenever the IR drops in the quasineutral regions are not negligible.

2.2.2.4 Thin-i-Layer p–i–n Diodes

Many modern VLSI devices operate at very high electric fields within the depletion regions of some of their p–n diodes. In fact, the junction fields are often so high that

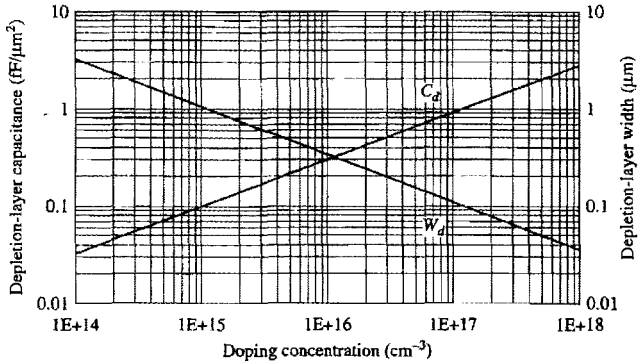


Figure 2.16. Depletion-layer width and depletion-layer capacitane, at zero bias, as a function of doping concentration of the lightly doped side of a one-sided p–n junction.

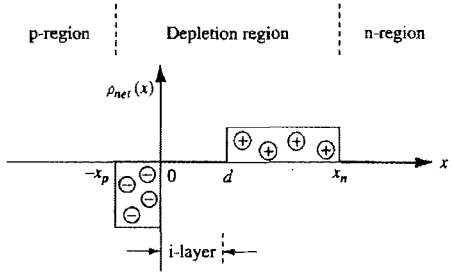


Figure 2.17. Charge distribution in a p–i–n diode.

detrimental high-field effects, such as avalanche multiplication and hot-carrier effects, limit the attainable device and circuit performance. To overcome the constraints imposed by high fields in a diode, device designers often introduce a thin but lightly doped region between the n- and the p-sides. In practice, this can be accomplished by sandwiching a lightly doped layer during epitaxial growth of the doped layers, or by grading the doping concentrations at or near the junction by ion implantation and/or diffusion. Analyses of such a diode structure become very simple if the lightly doped region is assumed to be intrinsic or undoped, i.e., if the lightly doped region is assumed to be an *i-layer*. This actually is not a bad approximation as long as the net charge concentration in the i-layer is at least several times smaller than the space-charge concentration on either side of the p–n junction, so that the contribution by the i-layer charge to the junction electric field is negligible. Figure 2.17 shows the charge distribution in such a p–i–n diode. The corresponding Poisson equation is

$$-\frac{d^2\psi_i}{dx^2} = \frac{qN_d}{\varepsilon_{si}} \quad \text{for} \quad d < x < x_n, \tag{2.86}$$

$$-\frac{d^2\psi_i}{dx^2} = 0 \quad \text{for} \quad 0 < x < d, \tag{2.87}$$

$$-\frac{d^2\psi_i}{dx^2} = -\frac{qN_a}{\varepsilon_{si}} \quad \text{for} \quad -x_p < x < 0. \tag{2.88}$$

These equations can be solved in the same way as Eqs. (2.76) and (2.77). Thus, integrating the equations once, subject to the boundary conditions that the electric field is zero at $x = -x_p$ and at $x = x_n$, gives

$$\mathscr{E}_m = \frac{qN_a x_p}{\varepsilon_{si}} = \frac{qN_d(x_n - d)}{\varepsilon_{si}}, \tag{2.89}$$

where $\mathscr{E}_m$ is the maximum electric field which exists in the region $0 \le x \le d$. Integrating the equations twice gives the total potential drop ψ_m across the junction as

$$\psi_m = \frac{\mathscr{E}_m(W_d + d)}{2}, \tag{2.90}$$

where $W_d = x_n + x_p$ is the total depletion-layer width. Eliminating $\mathscr{E}_m$ from Eqs. (2.89) and (2.90) gives

$$W_d = \sqrt{\frac{2\varepsilon_{si}(N_a + N_d)\psi_m}{qN_aN_d} + d^2}. \tag{2.91}$$

It is interesting to compare two diodes with the same externally applied voltage and the same p-side and n-side doping concentrations, one with an i-layer and one without. These two diodes have the same ψ_m. From Eq. (2.91), we can write

$$W_d = \sqrt{W_{d0}^2 + d^2} \tag{2.92}$$

for the diode with an i-layer, where W_{d0} is the depletion-layer width, given by Eq. (2.80), for the diode without an i-layer. Therefore,

$$\frac{W_d}{W_{d0}} = \sqrt{1 + \frac{d^2}{W_{d0}^2}}. \tag{2.93}$$

If we denote by $\mathscr{E}_{m0}$ the maximum electric field for the diode without an i-layer, then Eqs. (2.79) and (2.90) give the ratio of the electric fields as

$$\frac{\mathscr{E}_m}{\mathscr{E}_{m0}} = \frac{W_{d0}}{W_d + d} = \sqrt{1 + \frac{d^2}{W_{d0}^2}} - \frac{d}{W_{d0}}. \tag{2.94}$$

Thus, introduction of a lightly doped layer between the n- and p-regions of a diode reduces the maximum electric field in the junction. The depletion-layer charge ratio for the two diodes is, by Gauss's law,

$$\frac{Q_d}{Q_{d0}} = \frac{\mathscr{E}_m}{\mathscr{E}_{m0}} = \sqrt{1 + \frac{d^2}{W_{d0}^2}} - \frac{d}{W_{d0}}, \tag{2.95}$$

where Q_{d0} is the depletion-layer charge for the diode without an i-layer.

The depletion-layer capacitance per unit area can be calculated from Eq. (2.83), i.e., from $dQ_d(\text{p-side})/dV_{app}$, and the result is

$$C_d = \frac{\varepsilon_{si}}{W_d}. \tag{2.96}$$

The junction depletion-layer capacitance is related to the depletion-layer width in exactly the same way with or without an i-layer. This is expected from the physical picture of a parallel-plate capacitor where the capacitance is determined by the separation of the plates and not by any fixed charge distribution between the plates. The ratio of the capacitance with an i-layer to that without an i-layer is

$$\frac{C_d}{C_{d0}} = \frac{W_{d0}}{W_d} = \frac{1}{\sqrt{1 + d^2/W_{d0}^2}}, \tag{2.97}$$

where $C_{d0} = \varepsilon_{si}/W_{d0}$ is the depletion-layer capacitance for the diode without an i-layer.

2.2.3 The Diode Equation

In considering the current–voltage characteristics of a p–n diode, it is much more convenient to work with the quasi-Fermi potentials, instead of the intrinsic Fermi potential. The current densities and the quasi-Fermi potentials are given by Eqs. (2.63) to (2.66). These are repeated here for convenience:

$$J_n(x) = -qn(x)\mu_n(x)\frac{d\phi_n(x)}{dx}, \tag{2.98}$$

$$J_p(x) = -qp(x)\mu_p(x)\frac{d\phi_p(x)}{dx}, \tag{2.99}$$

where

$$\phi_n(x) \equiv \psi_i(x) - \frac{kT}{q}\ln\left[\frac{n(x)}{n_i}\right] \tag{2.100}$$

is the quasi-Fermi potential for electrons and

$$\phi_p(x) \equiv \psi_i(x) + \frac{kT}{q}\ln\left[\frac{p(x)}{n_i}\right] \tag{2.101}$$

is the quasi-Fermi potential for holes, and ψ_i is the electrostatic potential given by Eq. (2.40). In terms of the quasi-Fermi potentials, the pn product is

$$p(x)n(x) = n_i^2 \exp\left\{\frac{q[\phi_p(x) - \phi_n(x)]}{kT}\right\}. \tag{2.102}$$

In writing Eqs. (2.98) to (2.102), we have indicated explicitly the x dependence of the variables. In theory, the current–voltage characteristics of a diode can be obtained from these coupled equations. However, simple and close-form equations relating the electron and hole densities and currents to the applied voltage can be obtained if some approximations and assumptions are made. Here we discuss the physical bases for these approximations and assumptions, which will be used in later subsections to obtain equations describing the behavior of a diode in response to an applied voltage.

- *Quasineutrality.* As discussed in Section 2.1.4.8, the majority-carrier response time is on the order of 10^{-12} s. As we shall show later [see Fig. 2.24(b)], this time is extremely short compared to typical minority-carrier lifetimes. Therefore, as minority carriers are injected into a doped silicon region, the majority carriers respond practically instantaneously to maintain quasineutrality. For instance, let us consider the p-region of a forward-biased diode. As electrons are injected from the n-side, the change in electron concentration in the p-region instantaneously induces a change in the hole concentration in the region such that $\Delta p_p(x) = \Delta n_p(x)$ to maintain quasineutrality. Similarly, for the n-region of a forward-biased diode, we have $\Delta p_n(x) = \Delta n_n(x)$.
- *High-level injection.* A forward-bias voltage is said to cause a high-level injection of minority carriers in a p–n diode if it results in the injected minority-carrier density being comparable to or larger than the majority-carrier density. For a one-sided n^+–p diode, high-level injection means $\Delta n_p(x)$ is comparable to or larger than $p_{p0}(x)$ [$= N_a(x)$ for silicon at room temperature]. When high injection occurs, there is no simple relationship between the current components and the applied voltage. Usually, Eqs. (2.98) to (2.101) are used in numerical simulation to determine the electron and hole currents at any point in the diode.
- *Low-level injection.* A forward-bias voltage is said to cause only low-level injection of minority carriers in a p–n diode if the injected minority-carrier density is small compared to the majority-carrier density. For a one-sided n^+–p diode, this means $\Delta n_p(x) \ll p_{p0}(x)$. Unless stated otherwise, ***low-level injection is assumed*** in all cases discussed in this book. In the low-level-injection approximation, the majority-carrier density is simply equal to the density of ionized dopant impurities, and Eq. (2.102) suggests that the minority-carrier density is proportional to $\exp\left\{q\left[\phi_p(x) - \phi_n(x)\right]/kT\right\}$. Therefore, we need to establish the relationship between $\phi_p(x) - \phi_n(x)$ and the externally applied voltage V_{app} in order to derive the equations governing the current–voltage characteristics of a diode. This is done in Section 2.2.3.1 below.

2.2.3.1 Spatial Variaton of Majority-Carrier Quasi-Fermi Potential and *IR* Drop in a Quasineutral Region

Consider a p-type silicon having an equilibrium hole density of $p_{p0}(x)$ connected as a resistor. A current will flow in the resistor when a voltage is applied across it. The current flow causes an incremental voltage drop dV between x and $x + dx$. From Eq. (2.60), this incremental voltage drop is

$$dV(x) = d\psi_f(x) = -\frac{J_p(x) + J_n(x)}{qp_{p0}(x)\mu_p(x) + qn_{p0}(x)\mu_n(x)}dx \quad \text{(p-resistor)}$$
$$\approx -\frac{J_p(x)}{qp_{p0}(x)\mu_p(x)}dx, \tag{2.103}$$

where we have used the fact that $J_n(x) \ll J_p(x)$ and $n_{p0} \ll p_{p0}$ in a p-type resistor. Integrating Eq. (2.103) from contact to contact we have the voltage drop across the p-type resistor.

Next, let us consider the quasineutral p-region of a forward-biased p–n diode. The forward bias causes a current to flow in the diode, and hence through the quasineutral p- and n-regions. As the hole current flows in the quasineutral p-region, it causes a drop in the hole quasi-Fermi potential according to Eq. (2.98), i.e.,

$$d\phi_p(x) = -\frac{J_p(x)}{qp_p(x)\mu_p(x)}dx \quad \text{(p-region of forward-biased diode)}. \tag{2.104}$$

If we assume the quasineutral p-region extends from $x = 0$ (depletion-layer edge) to $x = W_p$ (p-region contact), we have

$$\phi_p(W_p) - \phi_p(0) = -\int_0^{W_p} \frac{J_p(x)}{qp_p(x)\mu_p(x)}dx \quad \text{(p-region of forward-biased diode)}$$
$$= -\int_0^{W_p} \frac{J_p(x)}{q[p_{p0}(x) + \Delta p_p(x)]\mu_p(x)}dx$$
$$\approx -\int_0^{W_p} \frac{J_p(x)}{qp_{p0}(x)\mu_p(x)}dx, \tag{2.105}$$

where we have used the fact that $p_p(x) = p_{p0}(x) + \Delta p_p(x)$ and that $\Delta p_p(x) = \Delta n_p(x) \ll p_{p0}(x)$ at low injection. Comparing with Eq. (2.103), we see that Eq. (2.105) has the physical meaning of being the voltage drop caused by the hole current flowing through the p-region. Similarly, the change in ϕ_n in the quasineutral n-region between the depletion-layer edge and the n-contact is equal to the *IR* drop caused by the electron current flowing through the quasineutral n-region.

2.2.3.2 Change of Quasi-Fermi Potentials across the Space-Charge Region

In theory, as long as there are electrons and holes flowing across the space-change region, there are ϕ_n and ϕ_p drops across the space-charge region governed by Eqs. (2.98) and (2.99). It is shown in Appendix 4 that the behavior of ϕ_n and ϕ_p in the space-charge region in forward bias is quite different than in reverse bias.

- *Forward-biased diode.* In forward bias, the results in Appendix 4 show that the drops in ϕ_n and ϕ_p across the space-charge region are small compared to kT/q. As a result, ϕ_n and ϕ_p are essentially constant across the space-charge region, as illustrated schematically in Fig. 2.18(b).
- *Reverse-biased diode.* In the case of reverse bias, the results in Appendix 4 show that the drops in ϕ_n and ϕ_p across the space-charge layer are small compared to kT/q only for small reverse bias ($|V_{app}|$ less than about $4kT/q$). For larger reverse bias, the drops in ϕ_n and ϕ_p across the space-charge layer increase approximately linearly with increase in reverse bias. Therefore, ϕ_n and ϕ_p are also relatively constant across the space-charge region for the case of small reverse bias, as illustrated schematically in Fig. 2.18(c).

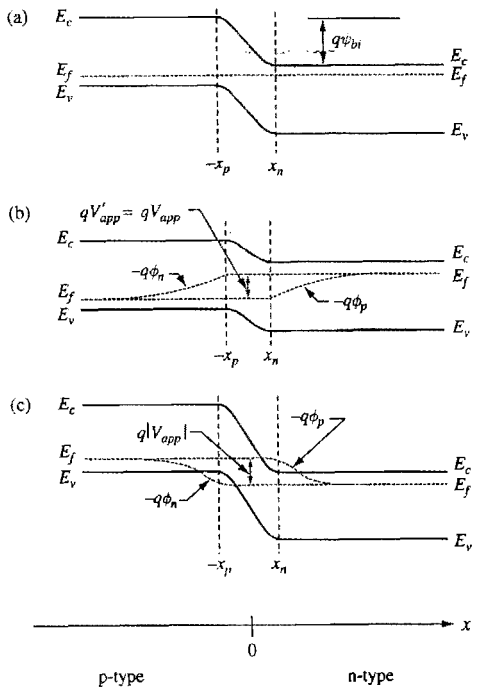


Figure 2.18. Schematics showing the variations of the quasi-Fermi potentials, ϕ_p for holes and ϕ_n for electrons, as a function of distance in a p–n diode. (a) A diode at thermal equilibrium with $V_{app} = 0$. (b) A forward-biased diode ($V_{app} > 0$) with negligible *IR* drops in the quasineutral regions. (c) A reverse-biased diode ($V_{app} < 0$). In the case of reverse bias, the drops in ϕ_n and ϕ_p across the space-charge region are small only at small $|V_{app}|$. At large $|V_{app}|$, the drops in ϕ_n and ϕ_p across the space-charge region increase with $|V_{app}|$.

2.2.3.3 Relationship Between Minority-Carrier Density and Applied Voltage

The relationship between the voltage across the space-charge region, V'_{app}, and the majority-carrier quasi-Fermi potentials at the space-charge-region boundaries, $\phi_p(-x_p)$ and $\phi_n(x_n)$ is

$$
\begin{aligned}
V'_{app} &\equiv V_{app} - IR(\text{p-side}) - IR(\text{n-side}) \\
&= V_{app} - [\phi_p(\text{p-contact}) - \phi_p(-x_p)] - [\phi_n(x_n) - \phi_n(\text{n-contact})] \\
&= \phi_p(-x_p) - \phi_n(x_n).
\end{aligned} \tag{2.106}
$$

In Eq. (2.106), we have used the results discussed in Sections 2.1.4.5 and 2.1.4.6 which state that $V_{app} = [E_{fn}(\text{n-contact}) - E_{fp}(\text{p-contact})]/q = \phi_p(\text{p-contact}) - \phi_n(\text{n-contact})$. For forward bias and small reverse bias, the drops in the quasi-Fermi potentials across the space-charge region are small compared to kT/q, i.e., $\phi_p(-x_p) \approx \phi_p(x_n)$ and $\phi_n(-x_p) \approx \phi_n(x_n)$. Therefore, Eqs. (2.102) and (2.106) can be combined, for forward bias and small reverse bias, to give the electron density on the p-side at the space-charge-layer edge as

$$
\begin{aligned}
n_p(-x_p) &= \frac{n_i^2}{p_p(-x_p)} \exp\{q[\phi_p(-x_p) - \phi_n(-x_p)]/kT\} \\
&\approx \frac{n_i^2}{p_p(-x_p)} \exp\{q[\phi_p(-x_p) - \phi_n(x_n)]/kT\} \\
&= \frac{n_i^2}{p_p(-x_p)} \exp(qV'_{app}/kT) \\
&= \frac{n_{p0}(-x_p)p_{p0}(-x_p)}{p_p(-x_p)} \exp(qV'_{app}/kT) \\
&\approx n_{p0}(-x_p) \exp(qV'_{app}/kT),
\end{aligned} \tag{2.107}
$$

where we have used the low-injection approximation to write $p_p \approx p_{p0}$. (The case of high injection will be discussed later.) Similarly, we have

$$
p_n(x_n) \approx p_{n0}(x_n) \exp(qV'_{app}/kT) \quad \text{forward bias and small reverse bias} \tag{2.108}
$$

is the hole density at the space-charge-layer edge on the n-side. ***Equations (2.107) and (2.108) are the most important boundary conditions governing a p–n diode.*** They relate the minority-carrier concentrations at the space-charge-region boundaries of the quasi-neutral regions to their thermal-equilibrium values and to the voltage across the space-charge region. For a forward-biased diode ($V'_{app} > 0$), we have an excess of minority carriers at the boundaries of the quasineutral regions. For a reverse-biased diode ($V'_{app} < 0$), we have a depletion of minority carriers at the boundaries of the quasineutral regions. Equations (2.107) and (2.108) are often referred to as the *Shockley diode equations* (Shockley, 1950).

The fact that Eqs. (2.107) and (2.108) are valid only for small reverse bias is often overlooked in the literature. It is shown in Appendix 4 that for reverse biases more negative than about $-4kT/q$, Eqs. (2.107) and (2.108) overestimate the degree of minority-carrier depletion at the quasineutral-region boundaries. However, it is also shown in Appendix 4 that once the depletion of minority carriers at the boundaries has reached 90%, corresponding to $|V_{app}| \sim 3\,kT/q$, further depletion of minority carriers has little effect on the diode current. In other words, using Eqs. (2.107) and (2.108) for $|V_{app}| > 3kT/q$ does not lead to any significant error in the calculated reverse-biased diode currents. Therefore, Eqs. (2.107) and (2.108) can be used to describe the transport properties in a reverse-biased diode as if they are valid for arbitrary reverse biases.

The distinction between V'_{app} and V_{app} is important whenever there is significant parasitic series resistance in a forward-biased diode, for instance, in the forward-biased emitter–base diode of a bipolar transistor. In most cases, the parasitic resistance can be modeled as a lump resistor in series with the diode, allowing us to quantify the difference between V'_{app} and V_{app} readily. For simplicity in writing the equations, we shall not make the distinction between V_{app} and V'_{app} when we use Eqs. (2.107) and (2.108) to derive the equations for the current–voltage characteristics. The distinction between V_{app} and V'_{app} will be pointed out wherever it is important to do so.

2.2.3.4 Diode Equation at High Minority-Carrier Injection

As stated in the derivation of Eqs. (2.107) and (2.108), these equations are valid at low injection. If the low-injection condition is not met, these equations are not valid and Eq. (2.102) should be used instead. At sufficiently large forward biases, the injected minority-carrier concentration, particularly on the lightly doped side of the diode, can be so large that, in order to maintain quasineutrality, the electron and hole concentrations become approximately equal. In this case, Eq. (2.102) gives $n \approx p \approx n_i \exp(qV'_{app}/2kT)$. At such high levels of minority-carrier injection, the concept of a well-defined transition region is no longer valid, and the quasi-Fermi potentials do not have simple behavior in any region of the diode (Gummel, 1967). The effect of high minority-carrier injection on the measured current–voltage characteristics of a diode will be discussed further in Section 2.2.4.10. An example of how the "boundary" of a p–n junction can be "relocated" at high minority-carrier injection can be found in Section 6.3.3 in connection with the discussion of base-widening effects in a bipolar transistor.

2.2.4 Current–Voltage Characteristics

As discussed in Section 2.2.1, at thermal equilibrium, the drift component of the current caused by the electric field in the space-charge region is exactly balanced out by the diffusion component of the current caused by the electron and hole concentration gradients across the junction, resulting in zero current flow in the diode. When an external voltage is applied, this current component balance is upset, and current will flow in the diode. If carriers are generated by light or some other means, thermal equilibrium is disturbed, and current can also flow in the diode. Here only the current flow in a diode as a result of an externally applied voltage is discussed. We first consider the current–voltage characteristics of an ideal diode governed by the Shockley diode equations (2.107) and (2.108). The space-charge-region current will be added later in Section 2.2.4.10 when we consider the deviation of a practical diode from ideal behavior.

Consider a forward-biased p–n diode. Electrons are injected from the n-side into the p-side, and holes are injected from the p-side into the n-side. Since space-charge-region current is ignored, the hole current leaving the p-side is the same as the hole current

entering the n-side. Similarly, the electron current leaving the n-side is equal to the electron current entering the p-side. To determine the total current flowing in the diode, all we need to do is to determine the hole current entering the n-side and the electron current entering the p-side.

The starting point for describing the current–voltage characteristics is the continuity equations. For electrons, it is given by Eq. (2.68) which is repeated here:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - R_n + G_n, \tag{2.109}$$

where R_n and G_n are the electron recombination and generation rates, respectively. (A detailed discussion of generation and recombination processes is given in Appendix 5.) Equation (2.109) can be rewritten as

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{n - n_0}{\tau_n}, \tag{2.110}$$

where

$$\tau_n \equiv \frac{n - n_0}{R_n - G_n} \tag{2.111}$$

is the *electron lifetime*, and n_0 is the electron concentration at thermal equilibrium. Substituting Eq. (2.54) for J_n into Eq. (2.110) gives

$$\frac{\partial n}{\partial t} = n\mu_n\frac{\partial \mathscr{E}}{\partial x} + \mu_n\mathscr{E}\frac{\partial n}{\partial x} + D_n\frac{\partial^2 n}{\partial x^2} - \frac{n - n_0}{\tau_n}. \tag{2.112}$$

2.2.4.1 Diodes with Uniformly Doped Regions

Let us consider electrons in the p-region of a p–n diode. For simplicity, we assume the p-region to be uniformly doped so that at low electron injection currents the hole density is uniform in the p-region. As will be shown in Section 6.1.2, the electric field is zero for a region where the majority-carrier concentration is uniform. Thus, for the p-region under discussion, $\partial\mathscr{E}/\partial x = 0$ and $\mathscr{E} = 0$. For electrons in this p-region, Eq. (2.112) reduces to

$$\frac{\partial n_p}{\partial t} = D_n\frac{\partial^2 n_p}{\partial x^2} - \frac{n_p - n_{p0}}{\tau_n}. \tag{2.113}$$

At steady state, Eq. (2.113) becomes

$$\frac{\partial n_p}{\partial t} = 0 = D_n\frac{\partial^2 n_p}{\partial x^2} - \frac{n_p - n_{p0}}{\tau_n}, \tag{2.114}$$

which can be rewritten as

$$\frac{d^2 n_p}{dx^2} - \frac{n_p - n_{p0}}{L_n^2} = 0, \tag{2.115}$$

where

$$L_n \equiv \sqrt{\tau_n D_n} = \sqrt{\frac{kT\mu_n\tau_n}{q}} \tag{2.116}$$

is the *electron diffusion length* in the p-region. It should be noted that ***the quantities in Eq. (2.116) are all for minority carriers, not majority carriers.***

In deriving the equations for minority-carrier transport, we can focus on minority electrons or minority holes. As we shall show later, the current–voltage characteristics of a one-sided diode are determined primarily by the transport of minority carriers in the lightly doped side. Forward-biased diodes are usually found in the operation of bipolar transistors. High-speed bipolar transistors are n–p–n type, instead of p–n–p type. That is, most commonly encountered one-sided forward-biased diodes are of the n^+–p type, instead of p^+–n type. Therefore, we choose to focus on minority electrons in deriving the transport equations. Also, we like to make some rearrangement to simplify the algebra in deriving the current equations. Earlier in this chapter, the physical junction of a p–n diode is assumed to be located at $x = 0$ with the p-silicon to the left side of the junction and the p-side depletion-layer edge located at $x = -x_p$. The n-silicon is located to the right side of the junction. The excess electrons in the p-region of the diode are injected from the n-side. These excess electrons will then move further into the p-region, contributing to electron current and becoming recombined along the way. That is, the p-side space-charge-region boundary is really the starting location for considering the distribution and transport of the excess electrons in the p-region. For considering the transport of electrons in the p-region, the algebra is simpler if we flip the p–n diode in Fig. 2.18 such that the n-region is on the left and the p-region is on the right, resulting in electrons flowing in the x-direction. The algebra can be further simplified if we shift the origin such that the quasineutral region of the p-side starts at $x = 0$ and ends at $x = W_p$. Note that in this arrangement, the electron current in the p-region has a negative sign (negative charges flowing in the x-direction). This is illustrated in Fig. 2.19 for an n^+–p diode.

In this rearranged coordinate system, the electron density at $x = 0$ is given by the Shockley diode equation, i.e., Eq. (2.107), while the electron density at $x = W_p$ is equal to n_{p0}, i.e.,

$$n_p(0) = n_{p0}\exp(qV_{app}/kT) \tag{2.117}$$

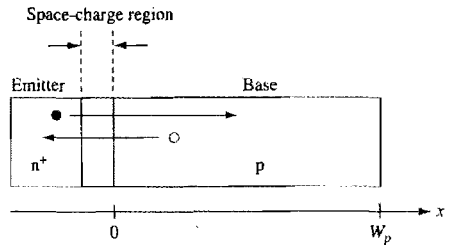


Figure 2.19. Schematic showing the coordinates used to develop the transport equations for a p–n diode. An n^+–p diode is assumed with the quasineutral p-region starting at $x = 0$ and ending at $x = W_p$.

and

$$n_p(W_p) = n_{p0}. \tag{2.118}$$

[To be accurate, V_{app} should be replaced by V'_{app} in Eq. (2.117). For simplicity in writing the equations, we are not making the distinction between V_{app} and V'_{app} unless there is confusion.] Solving Eq. (2.115) subject to these boundary conditions gives

$$n_p(x) - n_{p0} = n_{p0}\left[\exp\left(\frac{qV_{app}}{kT}\right) - 1\right]\frac{\sinh[(W_p - x)/L_n]}{\sinh(W_p/L_n)}. \tag{2.119}$$

Since there is no electric field in the quasineutral p-region, there is no electron drift-current component, only an electron diffusion-current component. The electron current density entering the p-region is

$$\begin{aligned} J_n(0) &= qD_n\left(\frac{dn_p}{dx}\right)_{x=0} \\ &= -\frac{qD_n n_{p0}[\exp(qV_{app}/kT) - 1]}{L_n \tanh(W_p/L_n)} \\ &= -\frac{qD_n n_i^2[\exp(qV_{app}/kT) - 1]}{p_{p0}L_n \tanh(W_p/L_n)}, \end{aligned} \tag{2.120}$$

where in writing the last equation we have used the fact that $n_{p0}p_{p0} = n_i^2$. Equations (2.119) and (2.120) are valid for a p-region of arbitrary width W_p. Note that J_n is negative in sign because electrons have a charge $-q$ and are flowing in the x-direction.

The hole density in the n-region and the hole current density entering the n-side have the same forms as Eq. (2.119) and Eq. (2.120), respectively, and can be derived in an analogous manner (cf. Exercise 2.16). ***The total current flowing through a p–n diode is the sum of the electron current on the p-side and the hole current on the n-side.*** That is, the diode current density is

$$\begin{aligned} J_{diode} &= -\frac{qD_n n_i^2[\exp(qV_{app}/kT) - 1]}{p_{p0}L_n \tanh(W_p/L_n)} - \frac{qD_p n_i^2[\exp(qV_{app}/kT) - 1]}{n_{n0}L_p \tanh(W_n/L_p)} \\ &= -\left[\frac{qL_n n_i^2}{N_a\tau_n \tanh(W_p/L_n)} + \frac{qL_p n_i^2}{N_d\tau_p \tanh(W_n/L_p)}\right]\left[\exp\left(\frac{qV_{app}}{kT}\right) - 1\right], \end{aligned} \tag{2.121}$$

where we have assumed that all the dopants are ionized so that $p_{p0} = N_a$ and $n_{n0} = N_d$. [The diode current represented by Eq. (2.121) is due to the diffusion of minority carriers in the quasineutral regions. It does not include the generation-recombination current in the space-charge region, which will be discussed in Section 2.2.4.10. The total diode current is the sum of the diffusion current and the generation-recombination current.] The negative sign in Eq. (2.121) is due to the fact that we placed the p-region to the right of the n-region, causing electrons to flow in the $+x$ direction and holes to flow in the $-x$ direction. The negative sign will not be there if we place the p-region to the left of the n-region. Ignoring the sign, Eq. (2.121) is often referred to as the *Shockley diode current equation*, or simply the Shockley diode equation. It is applicable to both forward bias ($V_{app} > 0$) and reverse bias ($V_{app} < 0$). Figure 2.20 is a semi-log plot of the diode current

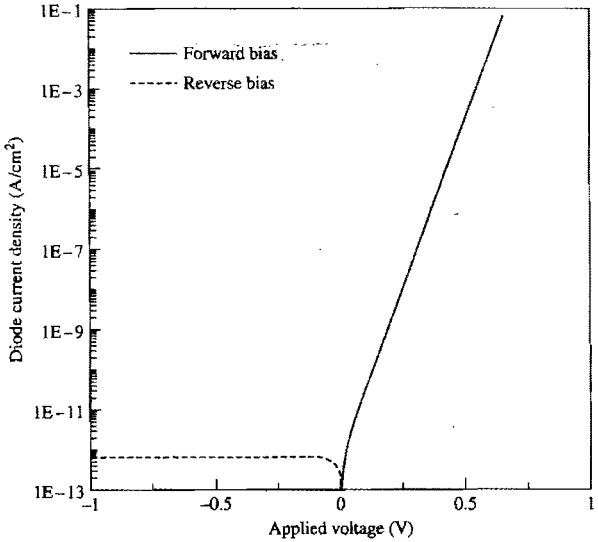


Figure 2.20. The current density of an ideal diode as given by Eq. (2.121). We assume $N_a = N_d = 10^{17}$ cm^{-3}, and use the corresponding values for L and τ in Fig. 2.24. W/L is assumed to be large so that $\tanh(W/L) = 1$.

density given by Eq. (2.121) as a function applied voltage. It represents the I–V characteristics of an ideal diode. It is of interest to contrast Fig. 2.20 with Fig. 2.14. On a linear plot (Fig. 2.14), the rectifying characteristics of the diode current are evident, with a turn-on voltage of about 0.8 V. On a semi-log plot (Fig. 2.20), only the exponential dependence of the forward-bias current on voltage and a low-level reverse-bias background current are obvious. Deviations of a practical diode from an ideal case are usually observable in a semi-log plot, but not in a linear plot, and will be discussed later in Section 2.2.4.10. At sufficiently large reverse bias, the diode will break down (not shown in Fig. 2.20). High-field effects, including avalanche breakdown of a p–n diode, will be covered in Section 2.5.

2.2.4.2 Emitter and Base of a Diode

Equation (2.121) shows that the minority-carrier current is inversely proportional to the doping concentration. Thus, in a one-sided diode, the minority-carrier current in the lightly doped side is much larger than that in the heavily doped side. The diode current is dominated by the flow of minority carriers in the lightly doped side of the diode, while minority-carrier current in the heavily doped side usually can be neglected in comparison. (The effect of heavy doping can increase the minority-current flowing in the heavily doped region substantially. Heavy-doping effect is particularly important in bipolar devices, and will be covered in Chapter 6. The effect of heavy doping on the magnitudes of the currents in a diode will be discussed as exercises.) The lightly doped side is often referred to as the *base* of the diode. The heavily doped side is often referred

to as the *emitter* of the diode, since the minority carriers entering the base are emitted from it.

In discussing the current–voltage characteristics of a diode, often only the minority-carrier current flow in the base is considered, since the minority-carrier current flow in the emitter is small in comparison. (However, if the width of an emitter is not larger than its minority-carrier diffusion length, the minority-carrier current flow in the emitter may not be negligible. Diodes with such emitters will be discussed further in Section 2.2.4.9.) As a result, unless stated explicitly, the region of the diode under discussion is assumed to be the base. That is, only the term in Eq. (2.121) corresponding to the base is kept. Whenever the emitter term is not negligible, both terms in Eq. (2.121) should be kept. In the following subsections, we examine in detail the current–voltage characteristics of one-sided n^+–p diodes. The equations derived can be modified readily to describe p^+–n diodes by changing the parameters for electrons in p-silicon to parameters for holes in n-silicon.

2.2.4.3 Forward-Biased n^+–p Diodes

We first consider the case where the n^+–p diode is moderately forward biased, i.e., $V_{app} > 0$, and $qV_{app}/kT \gg 1$. In this case, Eqs. (2.119) and (2.120) become

$$n_p(x) - n_{p0} = n_{p0}\exp(qV_{app}/kT)\frac{\sinh[(W_p - x)/L_n]}{\sinh(W_p/L_n)} \quad \text{(forward biased)} \tag{2.122}$$

and

$$J_n(0) = -\frac{qD_n n_i^2 \exp(qV_{app}/kT)}{P_{p0}L_n \tanh(W_p/L_n)} \quad \text{(forward biased)}. \tag{2.123}$$

That is, both ***the excess minority-carrier concentration and the minority-carrier current increase exponentially with the applied voltage*** (see Fig. 2.20).

2.2.4.4 Reverse-Biased n^+–p Diodes

Next we consider the case where the n^+–p diode is reverse-biased, i.e., $V_{app} < 0$, and $|qV_{app}| \gg kT$. In this case, Eqs. (2.119) and (2.120) become

$$n_p(x) - n_{p0} = -n_{p0}\frac{\sinh[(W_p - x)/L_n]}{\sinh(W_p/L_n)} \quad \text{(reverse biased)} \tag{2.124}$$

and

$$J_n(0) = \frac{qD_n n_i^2}{p_{p0}L_n \tanh(W_p/L_n)} \quad \text{(reverse biased)}. \tag{2.125}$$

Notice that $n_p(x) - n_{p0}$ is negative, and J_n is positive. The reverse bias causes a gradual depletion of electrons in the p-region near the depletion-region boundary, and this electron concentration gradient causes an electron current to flow from the quasineutral p-region towards the depletion region (in $-x$ direction according to our coordinates). This

is the electron diffusion component of the leakage current in a reverse-biased diode. It is also referred to as the electron *saturation current* of a diode.

The hole saturation current can be inferred from Eq. (2.121). The total diffusion leakage current in a diode is the sum of the electron and hole saturation currents. Notice that ***the diffusion leakage current is independent of the applied voltage.***

2.2.4.5 Wide-Base n^+–p Diodes

A diode is *wide-base* if its base width is large compared to the minority-carrier diffusion length in the base. For an n^+–p diode, this means $W_p/L_n \gg 1$. For a forward-biased wide-base diode, Eqs. (2.122) and (2.123) reduce to

$$n_p(x) - n_{p0} = n_{p0}\exp(qV_{app}/kT)\exp(-x/L_n) \quad \text{(forward, wide base)} \tag{2.126}$$

and

$$J_n(0) = -\frac{qD_n n_i^2}{p_{p0}L_n}\exp\left(\frac{qV_{app}}{kT}\right) \quad \text{(forward, wide base)}. \tag{2.127}$$

Thus, for a forward-biased wide-base diode, the excess minority-carrier concentration decreases exponentially with distance from the depletion-region boundary, and the minority-carrier current is independent of the base width.

For a reverse-biased wide-base diode, Eqs. (2.124) and (2.125) reduce to

$$n_p(x) - n_{p0} = -n_{p0}\exp(-x/L_n) \quad \text{(reverse, wide base)} \tag{2.128}$$

and

$$J_n(0) = \frac{qD_n n_i^2}{p_{p0}L_n} \quad \text{(reverse, wide base)}. \tag{2.129}$$

That is, the minority-carrier electrons in the base within a diffusion length of the depletion-region boundary diffuse towards the depletion region, with a saturation current density given by Eq. (2.129) which is independent of the base width.

2.2.4.6 Narrow-Base n^+–p Diodes

A diode is called *narrow-base* if its base width is small compared to the minority-carrier diffusion length in the base. In this case, this means $W_p/L_n \ll 1$. For a forward-biased narrow-base diode, Eqs. (2.122) and (2.123) reduce to

$$n_p(x) - n_{p0} = n_{p0}\exp\left(\frac{qV_{app}}{kT}\right)\left(1 - \frac{x}{W_p}\right) \quad \text{(forward, narrow base)} \tag{2.130}$$

and

$$J_n(0) = -\frac{qD_n n_i^2}{p_{p0}W_p}\exp\left(\frac{qV_{app}}{kT}\right) \quad \text{(forward, narrow base)}. \tag{2.131}$$

For a reverse-biased narrow-base diode, the corresponding equations are

$$n_p(x) - n_{p0} = -n_{p0}\left(1 - \frac{x}{W_p}\right) \quad \text{(reverse, narrow base)} \tag{2.132}$$

and

$$J_n(0) = \frac{qD_n n_i^2}{p_{p0} W_p} \quad \text{(reverse, narrow base)}. \tag{2.133}$$

For both forward and reverse biases, the minority-carrier current density in a narrow-base n^+–p diode increases as $1/W_p$. That is, for a narrow-base diode, the base current increases rapidly as the base width is reduced.

2.2.4.7 Spatial Distribution of Excess Minority Carriers

It can be seen from Eqs. (2.122) and (2.124) that both a forward-biased diode and a reverse-biased diode have the same sinh $[(W-x)/L]$ spatial dependence for the distribution of excess minority carriers (actually depletion of minority carriers in a reverse-biased diode). Figure 2.21 is a plot of the relative magnitude of the excess minority-carrier density as a function of x/L with W/L as a parameter. The $\exp(-x/L)$ distribution is for the case of $W/L = \infty$. It shows that a diode behaves like a wide-base diode for $W/L > 2$. For $W/L < 2$, the diode behavior depends strongly on W. For $W/L < 1$, the distribution can be approximated by the $1 - x/W$ dependence of a narrow-base diode.

2.2.4.8 Dependence of Minority-Carrier Current on Base Width

Figure 2.22 is a plot of the minority-carrier current density given by Eq. (2.120), normalized to its wide-base value. It shows that when $W/L < 1$, the minority-carrier current increases very reapidly as the diode base width decreases.

2.2.4.9 Shallow-Junction or Shallow-Emitter Diodes

Thus far, we have assumed the minority-carrier current in the emitter to be negligible compared to that in the base. A diode has a *shallow emitter* if the minority-carrier diffusion length in the emitter is comparable to or smaller than the width of the emitter

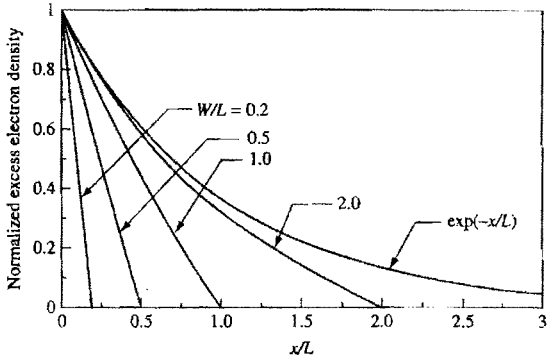


Figure 2.21. Relative magnitude of the excess minority-carrier concentration in the base of a diode as a function of distance from the base depletion-layer edge, with W/L as a parameter, where L is the minority-carrier diffusion length in the base and W is the base-region width. The case of $W/L = \infty$ is given by $\exp(-x/L)$.

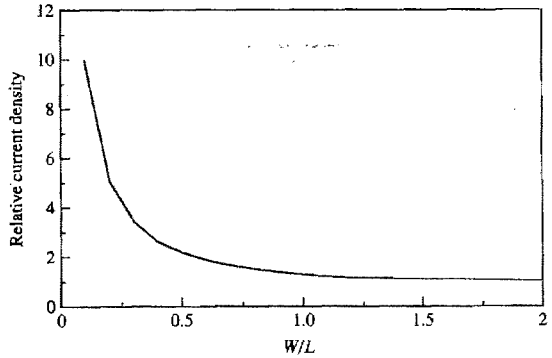


Figure 2.22. Relative maganitude of the minority-carrier current density in the base region of a diode as a function of W/L, normalized to the current at $W/L = \infty$. Here L is the minority-carrier diffusion length in the base, and W is the width of the base region.

region. The width of the emitter region of a p–n diode is also referred to as the junction depth. Therefore, a shallow-emitter diode is also a p–n junction having an electrically *shallow junction*. Figure 2.22 applies to the emitter region as well. Thus, we see from Fig. 2.22 that when $W/L < 1$ in the emitter, the minority-carrier current in the emitter increases very rapidly as the emitter depth decreases.

As can be inferred from Fig. 2.24(c), to be developed later in Section 2.2.4.12, the minority-carrier diffusion length is about 0.3 μm for a doping concentration of $1 \times 10^{20}\ \text{cm}^{-3}$, and much larger for lower doping concentrations. This length is larger than the emitter depth of a typical one-sided p–n diode in a modern VLSI device (e.g., the emitter of a bipolar transistor and the source and/or drain of a CMOS device). That is, ***typical p–n diodes in modern VLSI devices should be treated as shallow-junction diodes***.

There are effective means for reducing the minority-carrier current in a shallow-emitter diode. For instance, a shallow emitter can be contacted using a doped polysilicon layer instead of a metal or metal silicide layer. The physics of minority-carrier transport in a shallow emitter will be covered in detail in Chapters 6 and 7 in the context of modern bipolar transistors.

2.2.4.10 Space-Charge-Region Current and Ideality Factor of a Diode

Thus far, we have neglected the current originating from the generation and recombination of electrons and holes within the space-charge region. In practical silicon diodes, the space-charge region current can be larger than the Shockley diode current at reverse bias and at low forward bias. It is shown in Appendix 5 that the space-charge-region current can be written in the form

$$I_{SC}(V_{app}) = I_{SC0}[\exp(qV_{app}/2kT) - 1], \tag{2.134}$$

with

$$I_{SC0} = \frac{A_{diode} q n_i W_d}{\tau_n + \tau_p}, \tag{2.135}$$

where W_d is the width of the space-charge region, A_{diode} is the cross-sectional area of the diode, and τ_n and τ_p are the electron and hole lifetimes, respectively. Equation (2.134) is often referred to as the *Sah–Noyce–Shockley diode equation* (Sah *et al.*, 1957; Sah, 1991).

From Eq. (2.121), we can write the Shockley diode current in the form

$$I_{diode} = I_0[\exp(qV_{app}/kT) - 1], \tag{2.136}$$

with

$$I_0 = A_{diode} q n_i^2 \left[\frac{D_n}{p_{p0} L_n \tanh(W_p/L_n)} + \frac{D_p}{n_{p0} L_p \tanh(W_n/L_p)}\right]. \tag{2.137}$$

As discussed in Section 2.2.3.4, ***Eq. (2.136) is valid only at low injection levels***. For an n^+–p diode, high injection occurs when n_p approaches N_a where N_a is the acceptor concentration of the p-side. At high injection, IR drops in the quasineutral regions can be significant. Also, I_{diode} changes to an $\exp(qV'_{app}/2kT)$ dependence (see the discussion in Section 2.2.3.4). The onset of high injection can be pushed to higher voltage by increasing N_a.

The current measured at the diode terminals is

$$I_{total} = I_{diode} + I_{SC}. \tag{2.138}$$

Figure 2.23 is a schematic semi-log plot of a diode current as a function of its forward-bias terminal voltage, with series resistances neglected. A semi-log current–voltage plot

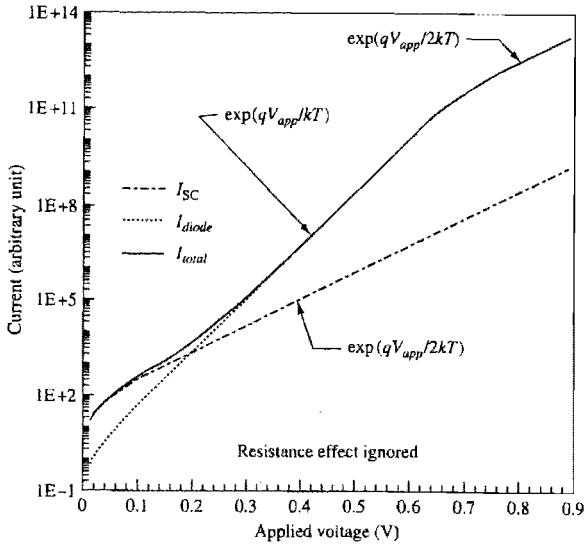


Figure 2.23. A schematic Gummel plot of the forward-bias current of a p–n diode. Series resistance effects are ignored. I_{diode} is the Shockley diode current. I_{SC} is the space-charge-region current.

for a diode or a bipolar transistor is called a *Gummel plot*. The slope in a Gummel plot is often used to infer the ideality of a diode. That is, the forward diode current is often expressed in the form

$$I_{total}(\text{forward}) \sim \exp(qV_{app}/mkT), \tag{2.139}$$

where m is called the *ideality factor*. Note that ***it is the diode terminal voltage V_{app}, not V'_{app} across the space-charge region that is in Eq. (2.139). The difference between V_{app} and V'_{app} is contained in the ideality factor.*** When m is unity, the current is considered "ideal." Figure 2.23 suggests that a forward diode current is ideal except at very small and very large forward biases. The nonideality at small forward bias is caused by the space-charge-region current. Space-charge-region current leads to $m \sim 2$ [see Eq. (2.134)]. The nonideality with $m \sim 2$ at very large forward bias is due to high-injection effect in the Shockley diode current (see the discussion in Section 2.2.3.4). At intermediate voltages, we have $1 < m < 2$.

Finite resistivity of the p- and n-regions results in voltage drops between the ohmic contacts and the junction. Finite resistivity effect is important only at very large forward biases. On a Gummel plot, finite resistivity effect can lead to m being very large. In general, when $1 < m < 2$ at large forward bias, it is not easy to clearly tell if the nonideality is caused by series resistance or by high injection. It may be a combination of both. However, when $m > 2$, we know that the series resistance effect dominates because the high injection effect by itself has an ideality factor of no larger than 2.

Series resistance effects can be reduced by increasing the diode doping concentrations, particularly the doping concentration of the base side of the diode. As discussed earlier, increasing doping concentration also delays the onset of high injection. Practical silicon diodes usually can be designed such that it appears quite ideal for forward biases of up to 0.8 V or slightly higher. Degradation in ideality factor is usually observable only at low forward biases and only in diodes having significant amounts of generation-recombination centers in the space-charge region. (An example of how the ideality factor changes with forward bias can be seen in the base current of a modern bipolar transistor shown in Fig. 6.13.)

2.2.4.11 Temperature Dependence and Magnitude of Diode Leakage Currents

For a reverse-biased diode, the total leakage current is the sum of the space-charge-region saturation current I_{SC0} and the diffusion saturation current I_0 given by Eqs. (2.135) and (2.137), respectively. The temperature dependence of I_0 is dominated by the temperature dependence of the n_i^2 factor, which, as shown in Eq. (2.13), is proportional to $\exp(-E_g/kT)$ where E_g is the bandgap energy. The space-charge-region leakage current I_{SC0}, being proportional to n_i, has a temperature dependence of $\exp(-E_g/2kT)$. In other words, the diffusion leakage current has an activation energy of about 1.1 eV while the generation-recombination leakage current has an activation energy of about 0.5 eV. This difference in activation energy can be used to distinguish the sources of the observed leakage current (Grove and Fitzgerald, 1966). [The diffusion leakage current is independent of reverse-bias voltage. The space-charge-region current is proportional to the space-charge-layer width which increases with reverse-bias voltage. In some devices,

this difference in voltage dependence may be used to distinguish the two leakage current sources.]

For an n^+–p diode with a base doping concentration of 1×10^{17} cm^{-3}, using the values of $D_n/L_n = L_n/\tau_n$ in Fig. 2.24 (to be derived later), Eq. (2.129) gives an electron diffusion leakage current density of about 8×10^{-13} A/cm^2 at room temperature. The observed diffusion leakage current in a typical n^+–p diode (which is the sum of the electron and hole diffusion currents) is comparable to the space-charge-region generation current, both being on the order of 10^{-13} A/cm^2 at room temperature (Kircher, 1975). However, the diffusion leakage current, due to its larger activation energy, is usually larger than the generation-recombination leakage current at elevated temperatures.

2.2.4.12 Minority-Carrier Mobility, Lifetime, and Diffusion Length

There have been many attempts to measure the minority-carrier lifetimes, mobilities, and diffusion lengths. For doping concentrations greater than about 1×10^{19} cm^{-3}, the experiments are quite difficult, since the minority-carrier concentrations are too small, and as a result there is quite a bit of spread in the reported data (Dziewior and Silber, 1979: Dziewior and Schmid, 1977: del Alamo *et al.*, 1985a, b). For purposes of device modeling, the following empirical equations have been proposed for minority-carrier electrons (Swirhun *et al.*, 1986) and minority-carrier holes (del Alamo *et al.*, 1985a, b):

$$\mu_n = 232 + \frac{1180}{1 + (N_a/8 \times 10^{16})^{0.9}} \text{cm}^2\text{-V-s}^{-1} \tag{2.140}$$

$$\mu_p = 130 + \frac{370}{1 + (N_d/8 \times 10^{17})^{1.25}} \text{cm}^2\text{-V-s}^{-1} \tag{2.141}$$

$$\frac{1}{\tau_n} = 3.45 \times 10^{-12} N_a + 0.95 \times 10^{-31} N_a^2 \text{ s}^{-1} \tag{2.142}$$

$$\frac{1}{\tau_p} = 7.8 \times 10^{-13} N_d + 1.8 \times 10^{-31} N_d^2 \text{ s}^{-1}. \tag{2.143}$$

The minority-carrier mobilities, lifetimes, and diffusion lengths are plotted as a function of doping concentration in Fig. 2.24(a), (b), and (c), respectively. The diffusion lengths are calculated from the mobilities and lifetimes using the relation $L = (kT\mu\tau/q)^{1/2}$.

There are more recent models of minority-carrier mobilities. In particular, there is a physics-based model that describes the mobilities of both majority and minority carriers in a consistent manner (Klaassen, 1990). For minority electrons, the Klaassen model is about the same as that in Fig. 2.24(a). For minority holes, Klaassen's model gives about the same mobilities as in Fig. 2.24(a) for high ($>2 \times 10^{18}$cm^{-3}) doping concentrations and about 30% lower mobilities at low ($<1 \times 10^{18}$cm^{-3}) doping concentrations (Klaassen *et al.*, 1992).

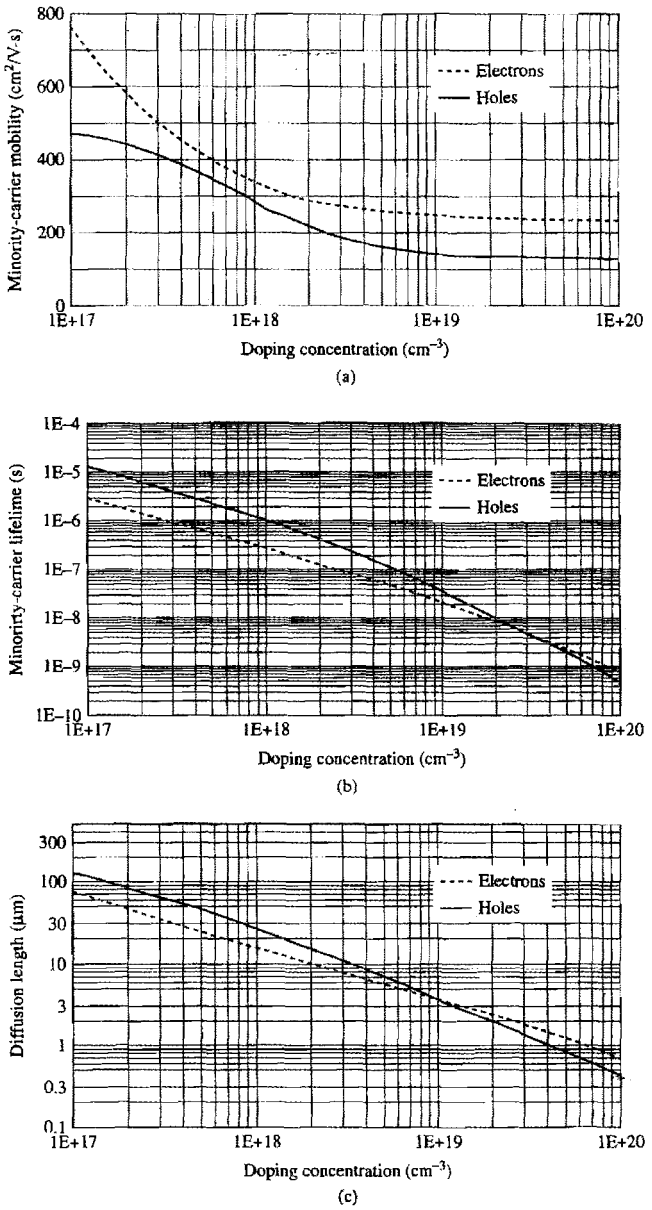


Figure 2.24. Minority-carrier (a) mobilities, (b) lifetimes, and (c) diffusion lengths as a function of doping concentration, calculated using the empirical equations (2.140) to (2.143).

2.2.5 Time-Dependent and Switching Characteristics

As discussed in Section 2.2.2, there is a capacitance associated with the depletion layer of a diode. As the diode is switched from off (zero-biased or reverse-biased) to on (forward-biased), it takes some time before the diode is turned on and reaches the steady state. This time is associated with charging up the depletion-layer capacitor and filling up the p- and n-regions with excess minority carriers. Similarly, when a diode is switched from the on state to the off state, it takes some time before the diode is turned off. This time is associated with discharging the depletion-layer capacitor and discharging the excess minority carriers stored in the p- and n-regions. The majority-carrier response time, or dielectric relaxation time, is negligibly short, on the order of 10^{-12} s, as shown in Section 2.1.4.8.

Consider the time needed to charge and discharge the depletion-layer capacitor. From Fig. 2.16, the depletion-layer capacitance C_d is typically on the order of 1 fF/μm^2. To turn a diode from off to on, and from on to off, the voltage swing V is typically about 1 V. If the diode is connected so that it carries a current density J of 1 mA/μm^2, then the time associated with charging and discharging the deplection-layer capacitor is on the order of C_dV/J, which is on the order of 10^{-12} s. Of course, this time changes in proportion to the current density J. However, as we shall show below, the time needed to charge and discharge the depletion-layer capacitor is usually very short compared with the time associated with charging and discharging the p- and n-regions of their minority carriers.

2.2.5.1 Excess Minority Carriers in the Base and Base Charging Time

Consider an n$^+$–p diode with a p-region base width W. When a forward bias is applied to it, minority-carrier electrons are injected into the base. As discussed in Section 2.2.4, for a wide-base diode, the minority-carrier density decreases exponentially with increasing distance, and practically all the minority carriers recombine before they reach the minority-carrier sink at $x = W$. For a narrow-base diode, on the other hand, practically all the minority carriers can travel across the base region without recombining.

The total *excess minority-carrier charge* (electrons) per unit area in the p-type base region is

$$Q_B = -q\int_0^W (n_p - n_{p0})\,dx. \tag{2.144}$$

For a wide-base diode, substituting Eqs. (2.126) and (2.127) into Eq. (2.144), we obtain

$$\begin{aligned} Q_B\,(\text{wide base}) &= -q(n_p - n_{p0})_{x=0}\,L_n \\ &= J_n(x=0)\tau_n, \end{aligned} \tag{2.145}$$

where we have used $\tau_n = L_n^2/D_n$ from Eq. (2.116).

For a narrow-base diode, substituting Eqs. (2.130) and (2.131) into Eq. (2.144), we obtain

$$\begin{aligned} Q_B\,(\text{narrow base}) &= -q(n_p - n_{p0})_{x=0}\left(\frac{W}{2}\right) \\ &= J_n\,(x=0)t_B, \end{aligned} \tag{2.146}$$

where the *base-transit time* t_B is defined by

$$\begin{aligned} t_B &\equiv \frac{Q_B\,(\text{narrow base})}{J_n\,(x=0)} \\ &= \frac{W^2}{2D_n}. \end{aligned} \tag{2.147}$$

As will be shown below, t_B is also equal to the average time for the minority carriers to traverse the narrow base region.

In a wide-base diode, it takes a time equal to the minority-carrier lifetime to fill the base with minority carriers. In a narrow-base diode, it takes a time equal to the base-transit time to fill the base with minority carriers. It should be noted that the charging current, $J_n(x=0)$, is different for wide-base and narrow-base diodes. The dependence of $J_n(x=0)$ on base width is shown in Fig. 2.22.

2.2.5.2 Average Time for Traversing a Narrow Base

From Eq. (2.130), the excess electron concentration at any point x in the narrow p-type base region is

$$n_p - n_{p0} = (n_p - n_{p0})_{x=0}\left(1 - \frac{x}{W}\right). \tag{2.148}$$

Let $v(x)$ be the *apparent velocity* of these excess carriers at point x. The current density due to those excess carriers at x is then

$$J_n(x) = -qv(x)(n_p - n_{p0})_x = -qv(x)(n_p - n_{p0})_{x=0}\left(1 - \frac{x}{W}\right). \tag{2.149}$$

The electron current density at $x=0$ is given by Eq. (2.131), i.e.,

$$J_n\,(x=0) = -\frac{q\,D_n}{W}\,(n_p - n_{p0})_{x=0}. \tag{2.150}$$

Assuming negligible recombination in the narrow base region, then current continuity requires $J_n(x)$ to be independent of x, i.e.,

$$v(x) = \frac{D_n}{W - x}. \tag{2.151}$$

The average time for traversing the base is thus given by

$$t_{avg} \equiv \int_0^W \frac{dx}{v(x)} = \frac{W^2}{2\,D_n}. \tag{2.152}$$

Comparison of Eqs. (2.147) and (2.152) shows that the base-transit time is equal to the average time for the minority carriers to traverse the narrow base.

It is instructive to esimate the magnitude of t_B. Modern n–p–n bipolar transistors typically have base widths of about 0.1μm, and a peak base doping concentration of about 2×10^{18} cm^{-3} (Nakamura and Nishizawa, 1995). The corresponding minority electron mobility, from Fig. 2.24, is about 300 cm^2/V-s. The base-transit time is therefore less than 1×10^{-11}s, which is extremely short compared with the corresponding

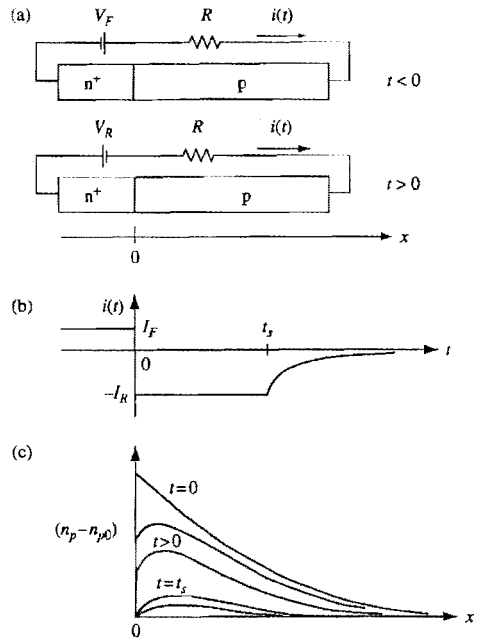


Figure 2.25. Schematics showing the switching of an n^+–p diode from forward bias to reverse bias: (a) the circuit schematics, (b) the diode current as a function of time, and (c) the excess-electron distribution in the base for different times.

minority-carrier lifetime on the order of 1×10^{-7}s. ***Recombination is negligible in the base layers of modern bipolar transistors.***

2.2.5.3 Discharge Time of a Forward-Biased Diode

Consider an n^+–p diode in a circuit configuration shown in Fig. 2.25(a). For simplicity, let us assume the external voltage, V_F or V_R, driving the circuit to be large compared to the internal junction voltage, i.e., the voltage immediately across the diode depletion layer, which is typically less than 1.0 V. At $t < 0$, there is a forward current of $I_F \approx V_F / R$ as illustrated in Fig. 2.25(b), and an excess electron distribution in the base region as illustrated in Fig. 2.25(c).

At time $t = 0$, the external bias is switched to a reverse voltage of V_R. The excess electrons in the base start to diffuse back towards the depletion region of the diode. Those electrons at the edge of the depletion region are swept away by the electric field in the depletion region towards the n^+ emitter at a saturated velocity of about 10^7 cm/s. As shown in Fig. 2.16, the depletion-layer width is typically on the order of 0.1 μm. The transit time across the depletion region is typically on the order of 10^{-12}s. As we shall see later, except for diodes of very narrow base widths, this time is extremely short compared to the total time for emptying the excess electrons out of the base region. Thus, as long as

there are sufficient excess electrons in the base region, the reverse current is limited not by the diffusion of excess electrons but by the external resistor and has a value of $I_R \approx V_R / R$, and the slope $(dn_p / dx)_{x=0}$, being proportional to I_R, is approximately constant.

As the excess electrons are discharged, part of the external voltage starts to appear across the p–n junction, and the junction becomes less forward biased. However, as long as there is still an appreciable amount of excess electrons stored in the base, the amount of external voltage appearing across the p–n junction remains very small. This is evident from Eq. (2.107), which indicates that even after the excess-electron concentration at the edge of the depletion layer has decreased by a factor of 10, the junction voltage has changed by only $2.3kT/q$, or 60mV. This is consistent with our assumptions that the reverse current remains essentially constant. During this time, the diode remains in the on condition.

At time $t = t_s$, the excess electrons have been depleted to the point that the reverse current is limited by the diffusion of electrons instead of by the external resistor. The rate of voltage change across the junction increases. Finally, when all the excess electrons are removed, the p–n diode is completely off. The external reverse-bias voltage appears entirely across the junction, and the reverse current is limited by the diode leakage current.

The time needed to switch off a forward-biased diode can be estimated from a charge-control analysis (Kuno, 1964). For simplicity, we shall estimate only the time during which the reverse current is approximately constant, and during which the diode remains in the on condition. Since the junction voltage remains approximately constant during this time, charging and discharging of the depletion-layer capacitance of the junction can be ignored. Let us consider the change in the amount of charge within the p-type base region. From Eq. (2.110), the continuity equation governing the electron concentration in the base region is

$$A_{diode}\frac{\partial(n_p - n_{p0})}{\partial t} = \frac{1}{q}\frac{\partial i_n(t)}{\partial x} - A_{diode}\frac{n_p - n_{p0}}{\tau_n}, \tag{2.153}$$

where $i_n(t)$ is the time-dependent electron current in the base region and A_{diode} is the cross-sectional area of the diode. Multiplying Eq. (2.153) by $-q$ and integrating over the base region, we have

$$\begin{aligned} -A_{diode}\,q\frac{\partial}{\partial t}\int_0^W (n_p - n_{p0})\,dx &= -\int_0^W \frac{\partial i_n(t)}{\partial x}\,dx \\ &\quad + A_{diode}\frac{q}{\tau_n}\int_0^W (n_p - n_{p0})\,dx, \end{aligned} \tag{2.154}$$

or

$$i_n(0,t) = A_{diode}\frac{dQ_B(t)}{dt} + A_{diode}\frac{Q_B(t)}{\tau_n} + i_n(W,t), \tag{2.155}$$

where $Q_B(t)$ is the excess minority charge per unit area stored in the base region, given by Eq. (2.144). Equation (2.155) is simply the continuity equation for the base region stated in the charge-control form. $i_n(0, t)$ is the electron current entering the base region, and

$i_n(W, t)$ is the electron current leaving the base region. At $x = W$, the electrons represented by $i_n(W, t)$ can simply exit the base region and continue on as an electron current outside the base, which is the case for electrons exiting the base of an n–p–n bipolar transistor (to be discussed in Chapter 6). Alternatively, the electrons represented by $i_n(W, t)$ can recombine with holes at the base ohmic contact located at $x = W$. The recombination gives rise to a current equal to $i_n(W, t)$ outside the base region. In either case, the current $i_n(W, t)$ is continuous across the base boundary at $x = W$, as required by charge conservation. The current flowing through the external resistor R is $i_n(0, t)$.

It is tempting to equate the current difference $i_n(0, t) - i_n(W, t)$ to the resistor current, but that is an inaccurate picture of current continuity. To see this, let us consider the steady-state situation when $dQ_B(t)/dt = 0$, and the special situation where recombination within the base is negligible ($\tau_n \to \infty$). In this case, Eq. (2.155) gives $i_n(0) - i_n(W) = 0$. However, the resistor current is not zero. The resistor current is $i_n(0)$, which is equal to $i_n(W)$ in this special case.

Consider a forward-biased diode being discharged. In this case, $i_n(0, t)$ is due to electrons diffusing back towards the n^+ emitter. For the coordinates system used here (see Section 2.2.4.1), these electrons travel in the $-x$ direction. Therefore, $i_n(0, t)$ is a positive quantity, and Eq. (2.155) gives

$$A_{diode}\frac{dQ_B(t)}{dt} + A_{diode}\frac{Q_B(t)}{\tau_n} + i_n(W, t) = I_R. \tag{2.156}$$

Equation (2.156) is the continuity equation stated in the charge-control form for the base region of a diode at the initial stage of being discharged.

- *Discharge time for a wide-base diode.* For a wide-base diode, $i_n(W, t) = 0$ and the solution for Eq. (2.156) is

$$A_{diode}Q_B(t) = I_R\tau_n + [A_{diode}Q_B(0) - I_R\tau_n]\exp(-t/\tau_n), \tag{2.157}$$

or

$$\frac{Q_B(t)}{Q_B(0)} = \frac{I_R\tau_n}{A_{diode}Q_B(0)}[1 - \exp(-t/\tau_n)] + \exp(-t/\tau_n), \tag{2.158}$$

where $Q_B(0)$ is the excess minority charge per unit area just after the diode is switched from forward bias to reverse bias. For a wide-base diode, Eq. (2.145) gives $A_{diode}Q_B(0) = -\tau_n I_F$. (Note the negative sign in Q_B, since Q_B is negative for electrons.) Therefore, Eq. (2.158) gives

$$\frac{Q_B(t)}{Q_B(0)} = -\frac{I_R}{I_F}[1 - \exp(-t/\tau_n)] + \exp(-t/\tau_n). \tag{2.159}$$

Figure 2.26 is a plot of the charge ratio $Q_B(t)/Q_B(0)$ as a function of t/τ_n with the current ratio I_R/I_F as a parameter. It shows that a forward-biased wide-base diode discharges with a time constant approximately equal to the minority-carrier lifetime, unless the reverse discharge current is much larger than the forward charging current. Even for $I_R/I_F = 10$, the diode discharges in a time of approximately $\tau_n/10$ which, as can be seen

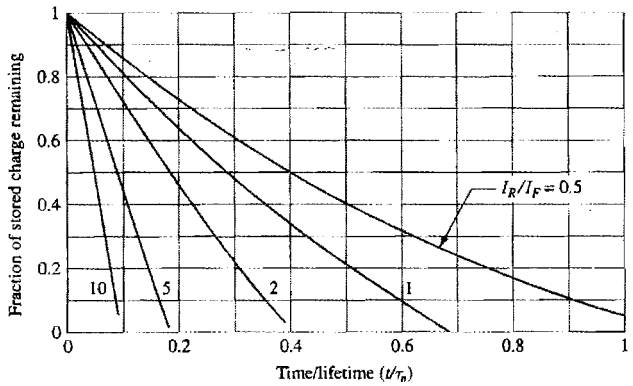


Figure 2.26. Plot of charges ratio $Q_B(t)/Q_B(0)$ as a function of t/τ_n during the discharge of a forward-biased diode, with the ratio of discharge current to charging current, I_R/I_F, as a parameter.

from Fig. 2.24(b), is larger than 10^{-8} s for most diodes of practical doping concentrations. This time is very long compared to the typical switching delays of VLSI circuits. The important point is that ***it takes a long time to drain off the excess minority carriers stored in a wide-base diode and turn it off.*** It is important to minimize excess minority carriers stored in forward-biased diodes if these diodes are to be switched off fast.

- *Discharge time for a narrow-base diode.* For a narrow-base diode, recombination can be ignored. Therefore, we have $|i_n(0)| = |i_n(W)| = I_F$ while the diode is in forward bias, and the distribution of excess electrons in the base has a constant gradient given by Eq. (2.130). At $t > 0$, after the diode has been switched from forward bias to reverse bias, electrons continue to flow towards and recombine at the base contact, i.e., $|i_n(W, t)| > 0$. As we shall show below, a narrow-base diode discharges in a time very small compared to τ_n. That is, during the discharge of a narrow-base diode, the recombination term $Q_B(t)/\tau_n$ can be neglected, and the minority electrons are discharged only through back diffusion towards the n^+ emitter and recombination at the base contact. With this approximation, Eq. (2.156) reduces to

$$A_{diode}\frac{dQ_B(t)}{dt} = I_R - i_n(W, t). \tag{2.160}$$

To get an idea of how fast a narrow-base diode can be discharged, let us assume that the gradient of the electron distribution at $x = W$ remains about the same for a short time immediately after the diode is switched to reverse bias as during forward bias. That is, for a short time after switching from forward bias to reverse bias, we have $i_n(W, t) \sim -I_F$. [Note the negative sign for $i_n(W, t)$. Electrons flowing in the x-direction lead to a negative current.] Also, we note that Eq. (2.146) gives $A_{diode}Q_B(0) = -t_B I_F$ for a narrow-base diode, where t_B is the base transit time. With these assumptions, Eq. (2.160) gives

$$\frac{Q_B(t)}{Q_B(0)} \approx 1 - \frac{(I_R + I_F)t}{t_B I_F} \tag{2.161}$$

for a short time after switching from on to off. Equation (2.161) shows that the discharge time for a narrow-base diode lasts approximately $t_B I_F/(I_R + I_F)$ which, for a large I_R/I_F ratio, can be much shorter than the base transit time. A complex but closed-form solution can be obtained in the large I_R/I_F limit (Lindmayer and Wrigley, 1965), which shows that most of the charge has come out by about $t_B/3$. The important point is that ***a forward-biased narrow-base diode can be switched off fast.***

2.2.6 Diffusion Capacitance

For a forward-biased diode, in addition to the capacitance associated with the space-charge layer, there is an important capacitance component associated with the rearrangement of the excess minority carriers in the diode in response to a change in the applied voltage. This minority-carrier capacitance is called *diffusion capacitance* C_D.

Consider an n^+–p wide-emitter narrow-base diode, i.e., a diode where the depth or width of the n^+ emitter region is large compared to its hole diffusion length and the width of the p-type base region is small compared to its electron diffusion length. (This diode is of interest because it represents the emitter–base diode of an n–p–n bipolar transistor.) When a voltage V_{app} is applied across the diode, an electron current of magnitude I_n is injected from the emitter into the base and a hole current of magnitude I_p is injected from the base into the emitter. The diode current is $I_n + I_p$. Both I_n and I_p are proportional to $\exp(qV_{app}/kT)$.

- *Quasisteady state.* In a quasisteady state, the voltage is assumed to vary slowly in time such that the *minority charge distribution can respond to the applied voltage fully* without any delay. In this case, the excess electron charge in the base is given by Eq. (2.146), i.e.,

$$A_{diode}\,|Q_n(V_{app})| = I_n(V_{app})t_B \quad \text{(narrow base)}, \tag{2.162}$$

which in turn gives

$$A_{diode}\frac{\Delta|Q_n(V_{app})|}{\Delta V_{app}} = \frac{q}{kT}I_n(V_{app})t_B = \frac{q}{kT}I_n(V_{app})\left(\frac{W_B^2}{2D_{nB}}\right) \quad \text{(narrow base)}, \tag{2.163}$$

where we have used Eq. (2.147) for the base transit time t_B. In Eq. (2.163), W_B is the base width and D_{nB} is the electron diffusion coefficient in the base. Similarly, using Eq. (2.145) and Eq. (2.116), we have

$$A_{diode}\frac{\Delta|Q_p(V_{app})|}{\Delta V_{app}} = \frac{q}{kT}I_p(V_{app})\tau_{pE} = \frac{q}{kT}I_p(V_{app})\left(\frac{L_{pE}^2}{D_{pE}}\right) \quad \text{(wide emitter)} \tag{2.164}$$

for the stored holes in the n^+ emitter, where L_{pE}, D_{pE}, and τ_{pE} are the diffusion length, diffusion coefficient, and lifetime, respectively, of holes in the emitter. Equations (2.163) and (2.164) relate the change in the stored charge caused by a change in the voltage across the diode in a quasisteady state. However they do *not* represent the true diffusion capacitance components of a forward-biased diode, which we shall discuss next.

- *Diffusion capacitance components.* Consider the discharge of a forward-biased base region illustrated schematically in Fig. 2.25(c). When the diode is forward biased, the electron distribution is represented by the $t=0$ curve. When the forward bias is reduced, or when the diode is switched to reverse bias, the electron distribution evolves as a function of time, as indicated by the $t>0$ curves. Part of the excess electrons diffuses to the left (back towards the emitter) and part of them diffuses to the right. The opposing electron currents suggest that the *net* charge moved through the external circuit in the discharge process is less than the total stored charge represented by the $t=0$ curve in Fig. 2.25(c). When an ac voltage is applied across the diode, only those electrons located sufficiently close to the depletion-region boundaries can keep up with the signal and get into and out of the base. The exact amount depends on the signal frequency. These signal-following electrons give rise to $i_n(0, t)$, the time-dependent electron current at the emitter end of the quasineutral base region. As discussed in Section 2.2.5.3, $i_n(0, t)$ is the electron component of the current in the external circuit. Similarly, if we consider the stored holes in the emitter, the signal-following holes in the n^+ emitter give rise to a hole current component $i_p(0, t)$ at the base end of the emitter region and in the external circuit. These signal-following stored charges are responsible for the diffusion capacitance.

The exact diffusion capacitance components can be derived from a frequency-dependent small-signal analysis of the current through a diode starting from the differential equations governing the transport of minority carriers (Shockley, 1949; Lindmayer and Wrigley, 1965; Pritchard, 1967). This is done for a wide-emitter narrow-base diode in Appendix 6. Here we simply state the results.

For a wide-emitter and narrow-base n^+–p diode, the low-frequency diffusion capacitance due to the excess electrons in the base is

$$C_{Dn} = \frac{qI_n(V_{app})}{kT}\left(\frac{W_B^2}{3D_{nB}}\right) = \frac{2}{3}\frac{q}{kT}I_n(V_{app})t_B \quad \text{(narrow base)}, \tag{2.165}$$

and that due to the excess holes in the emitter is

$$C_{Dp} = \frac{qI_p(V_{app})}{kT}\left(\frac{L_{pE}^2}{2D_{pE}}\right) = \frac{1}{2}\frac{q}{kT}I_p(V_{app})\tau_{pE} \quad \text{(wide emitter)}. \tag{2.166}$$

The total diffusion capacitance is

$$C_D = C_{Dn} + C_{Dp}. \tag{2.167}$$

Comparison with Eqs. (2.163) and (2.164) shows that 2/3 of the stored charge in the narrow base and 1/2 of the stored charge in the wide emitter contribute to the diffusion capacitance of a forward-biased diode. [In the case of a narrow base, a closed-form

solution can be obtained in the large-discharge-current limit for the transient discharge current. Integration of the transient discharge current shows that 2/3 of the total stored charge in the narrow base diffuses back to the emitter when the base region is discharged. This fraction is the same as the fraction of total stored charge in the base contributing to the diffusion capacitance. In other words, one can think of the diffusion capacitance as coming from the portion of the stored minority charge that is "reclaimable" in the form of an ac current as the diode responds to an ac signal (Lindmayer and Wrigley, 1965).]

It is instructive to examine the relative magnitude of the two capacitance components C_{Dn} and C_{Dp}. Using the hole equivalent of Eq. (2.127) for hole current and Eq. (2.131) for electron current, and the relationship in Eq. (2.116), we have

$$\frac{C_{Dn}\,(\text{narrow base})}{C_{Dp}\,(\text{wide emitter})} = \frac{2}{3}\frac{N_E}{N_B}\frac{W_B}{L_{pE}}. \tag{2.168}$$

The ratio N_E/N_B is typically about 100 for an n^+–p diode. For an emitter with $N_E = 1 \times 10^{20}\text{cm}^{-3}$, L_{pE} is about 0.3 μm (see Fig. 2.24). Therefore, for practical one-sided diodes where the base width is larger than 0.03 μm, the ratio C_{Dn}/C_{Dp} is much larger than unity. That is, the diffusion capacitance of a one-sided p–n diode is dominated by the minority charge stored in the base of the diode. The diffusion capacitance due to the minority charge stored in the emitter is small in comparison. The effect of heavy doping, when included, will increase the amount of stored charge and hence the diffusion capacitance. Since the heavy-doping effect is larger in the more heavily doped emitter than in the base, it will make the ratio C_{Dn}/C_{Dp} smaller than that given by Eq. (2.168) (See Exercise 2.18).

2.3 MOS Capacitors

The metal–oxide–semiconductor (MOS) structure is the basis of CMOS technology. The Si–SiO_2 MOS system has been studied extensively (Nicollian and Brews, 1982) because it is directly related to most planar devices and integrated circuits. In this section, we review the fundamental properties of MOS capacitors and the basic equations that govern their operation. The effects of charges in the oxide layer and at the oxide–silicon interface are discussed in Section 2.3.6.

2.3.1 Surface Potential: Accumulation, Depletion, and Inversion

2.3.1.1 Energy-Band Diagram of an MOS System

The cross section of an MOS capacitor is shown in Fig. 2.27. It consists of a conducting gate electrode (metal or heavily doped polysilicon) on top of a thin layer of silicon dioxide grown on a silicon substrate. The energy band diagrams of the three components when separate are shown in Figure. 2.28. Before we discuss the energy band diagram of an MOS device, it is necessary to first introduce the concept of *free electron level* and *work function* which play key roles in the relative energy band placement when two different materials are brought into contact. Figure 2.28(c) shows the band diagram of a

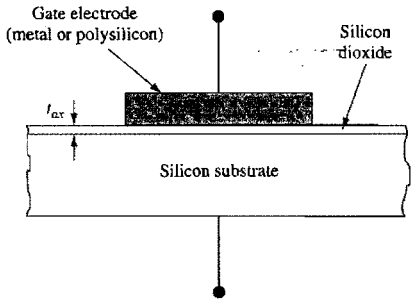


Figure 2.27. Schematic cross section of an MOS capacitor.

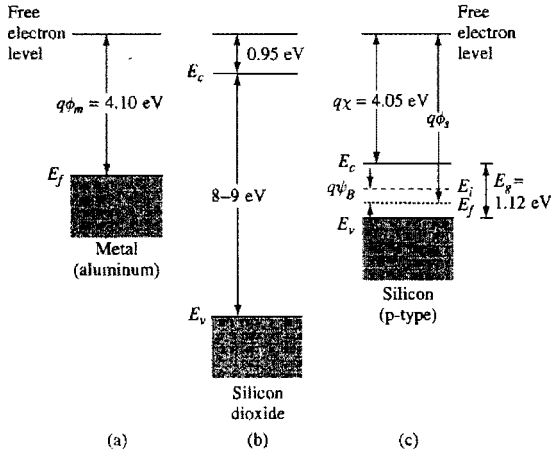


Figure 2.28. Energy-band diagram of the three components of an MOS capacitor: (a) metal (aluminum), (b) silicon dioxide, and (c) p-type silicon.

p-type silicon with the addition of the free electron level at some energy above the conduction band. The free electron level is defined as the energy level above which the electron is free, i.e., no longer bonded to the lattice.[3] In silicon, the free electron level is 4.05 eV above the conduction band edge, as shown in Figure. 2.28(c). In other words, an electron at the conduction band edge must gain an additional energy of 4.05 eV (called the *electron affinity*, $q\chi$) in order to break loose from the crystal field of silicon. Figure 2.28(b) shows the band diagram of silicon dioxide – an insulator with a large energy gap in the range of 8–9 eV. The free electron level in silicon dioxide is 0.95 eV above its conduction band.

[3] In other texts, the free electron level is often referred to as the *vacuum level*. Here we use a different term to avoid the implication that the vacuum level is universal.

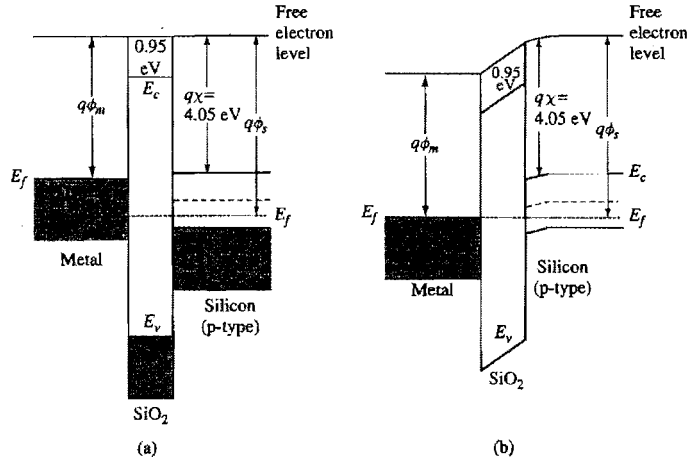


Figure 2.29. Band diagrams of an MOS system under (a) the flatband condition, and (b) zero gate-voltage condition.

Work function is defined as the energy difference between the free electron level and the Fermi level. For the p-type silicon example in Fig. 2.28(c), the work function, $q\phi_s$, can be expressed as:

$$q\phi_s = q\chi + \frac{E_g}{2} + q\psi_B. \tag{2.169}$$

Here ψ_B is the difference between the Fermi potential and the intrinsic potential given by Eq. (2.48). The same definition of work function, $q\phi_m$, applies to metals (remember that the conduction band is half filled in metals), as shown in Fig. 2.28(a). It means that an electron at the Fermi level needs to receive an energy equal to $q\phi_m$ to be free from the metal. Different metals have different work functions.

When two different materials are brought into contact, they must share the same free electron level at the interface, i.e., the free electron level is continuous from one material to the next. This is because at the interface of two materials, an electron that is free from the crystal field of one material is also free from the crystal field of the other material. Figure 2.29(a) shows the band diagram of an MOS system under the *flatband condition* in which there is no field in all three materials. Since for this example the metal work function is less than the silicon work function, the flatband condition is reached by applying a negative gate voltage, $-(\phi_s - \phi_m) \equiv \phi_{ms}$, called the *flatband voltage*, with respect to the silicon substrate. This is seen in Fig. 2.29(a) as the displacement between the two Fermi levels. In general, the flatband voltage of an MOS device is given by

$$V_{fb} = (\phi_m - \phi_s) - \frac{Q_{ox}}{C_{ox}} = \phi_{ms} - \frac{Q_{ox}}{C_{ox}}, \tag{2.170}$$

where Q_{ox} is the equivalent oxide charge per unit area at the oxide–silicon interface (defined in Section 2.3.7), and C_{ox} is the oxide capacitance per unit area,

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{2.171}$$

for an oxide film of thickness t_{ox} and permittivity ε_{ox}. In modern VLSI technologies, Q_{ox}/q at the Si–SiO_2 interface can be controlled to below 10^{10} cm^{-2} (positive) for ⟨100⟩-oriented surfaces. Its contribution to the flatband voltage is less than 50 mV for thin gate oxides used in 1-μm technology and below ($t_{ox} \leq 20$ nm). Therefore, the flatband voltage is mainly determined by the work function difference ϕ_{ms}.

At zero gate voltage when the Fermi levels line up, electric fields are developed in both the oxide and silicon, as shown in Fig. 2.29(b). One should note that the electron affinity, $q\chi$, is a material property which depends only on the type of the semiconductor and does not change with either the location or the doping type. When there is band bending as in Fig. 2.29(b) or, e.g., as in the depletion region of a p–n junction, the free electron level would bend in parallel with the conduction band such that the distance between the free electron level and the conduction band remains constant, much like the energy gap. While in this case the free electron level is not flat within one material, it must still be continuous between adjacent materials. Because of the common free electron level at the interface, ***the electron energy barrier is $q\phi_{ox}$ = 4.05 eV − 0.95 eV = 3.1 eV between the conduction bands of silicon and silicon dioxide***. This figure has important significance when discussing the reliability of Si–SiO_2 systems (Section 2.5.3). For simplicity, the free electron level is mostly omitted in subsequent MOS band diagrams. One should keep in mind its essential role in setting the relative band placement between different materials.

The same principle of continuity of free-electron level at interface applies to metal–semiconductor junctions (Schottky diodes discussed in Section 2.4.1) and heterojunction devices with different bandgaps in establishing their band alignment relationships.

2.3.1.2 Gate Voltage and Surface Potential

Figure 2.30 shows the band diagrams of an MOS device when different gate voltages are applied. Free electron levels are omitted. The top level in the oxide region represents the

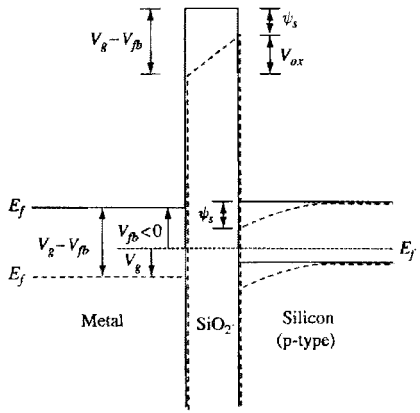


Figure 2.30. Energy band and potential diagrams of an MOS capacitor showing how the bands change under different gate bias conditions. The solid lines represent the flatband condition. The dashed lines represent the condition when a positive gate voltage is applied. Note that in the diagram the electron

conduction band. The same energy band diagrams are also used as potential diagrams to show the relationship between the applied voltage and the potential drop in different regions of the device. The applied gate voltage is referenced with respect to the Fermi level of the p-type substrate. In the flatband condition depicted by the solid lines in Fig. 2.30, $V_{fb} < 0$ is applied to the gate, as in the case of Fig. 2.29(a). When a positive gate voltage, V_g, is applied, as depicted by the dashed lines in Fig. 2.30, the metal Fermi level is displaced downward from that of the flatband condition by a total amount of $V_g - V_{fb}$. Because of the fixed band relationship between the metal and oxide, the oxide conduction band on the metal side is also displaced downward by the same amount. This causes a field to develop in the oxide and, at the same time, a downward bending of the bands in the p-type silicon near the surface. The amount of band bending in silicon is defined as the *surface potential*, ψ_s, i.e., the potential at the silicon surface relative to that in the bulk substrate. Because of the fixed band relationship between the oxide and silicon, it is clear that

$$V_g - V_{fb} = \psi_s + V_{ox}, \tag{2.172}$$

where V_{ox} is the potential drop across the oxide, as indicated in Fig. 2.30.

How $V_g - V_{fb}$ is partitioned into ψ_s and V_{ox} depends on both the oxide thickness and the doping concentration of the p-type silicon. Based on the dielectric boundary conditions discussed in Section 2.1.4.2, a field relationship exists at the silicon–oxide interface,

$$\varepsilon_{ox} \mathscr{E}_{ox} = \varepsilon_{si} \mathscr{E}_s, \tag{2.173}$$

or $\mathscr{E}_{ox} \approx 3\mathscr{E}_s$, assuming negligible trapped charge at the interface. Note that the above equation applies to both the magnitude and the direction of the fields. In most cases, there is negligible net charge in the oxide and Poisson's equation becomes $d\mathscr{E}/dx = 0$. Therefore, the field in the oxide is constant,[4] and $V_{ox} = \mathscr{E}_{ox} t_{ox}$.

2.3.1.3 Accumulation, Depletion, and Inversion

Figure 2.31 shows the band diagrams of p-type ((a)–(d)) and n-type ((e)–(h)) MOS capacitors under different gate bias voltages with respect to the flatband voltage. For simplicity, the flatband voltage is taken to be zero for all cases. The flatband condition for p-type MOS discussed before is shown in Fig. 2.31(a). There is no charge, no field, and the carrier concentration equals the ionized acceptor concentration throughout the silicon. Now consider the case when a negative voltage is applied to the gate of a p-type MOS capacitor, as shown in Fig. 2.31(b). This raises the metal Fermi level (i.e., electron energy) with respect to the silicon Fermi level and creates an electric field in the oxide that would accelerate a negative charge toward the silicon substrate. A field is also induced at the silicon surface in the same direction as the oxide field. Because of the low carrier concentration in silicon (compared with metal), the bands bend upward toward the oxide interface. ***The Fermi level stays flat within the silicon, since there is no net flow of conduction current***, as was discussed in Section 2.1.4.5. Due to the band bending, the

[4] If the field in the oxide is not constant, $\mathscr{E}_{ox}$ in Eq. (2.173) is defined as the oxide field at the oxide–silicon interface.

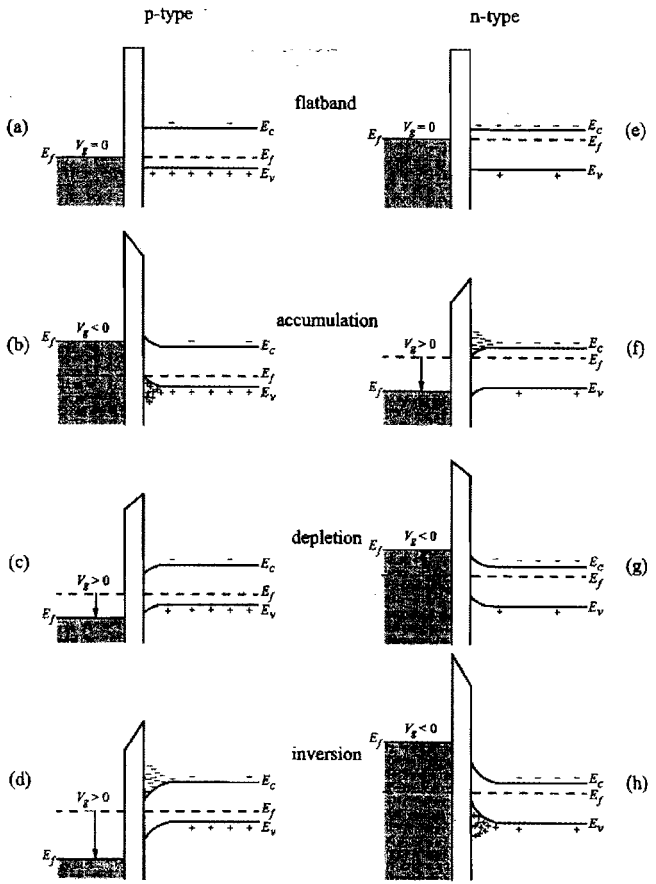


Figure 2.31. Energy-band diagrams for ideal (zero flatband voltage) (a)–(d) p-type and (e)–(h) n-type MOS capacitors under different bias conditions: (a), (e), flat band; (b), (f), accumulation; (c), (g), depletion; (d), (h), inversion. (After Sze, 1981.)

valence band at the surface is much closer to the Fermi level than is the valence band in the bulk silicon. This results in a hole concentration much higher at the surface than the equilibrium hole concentration in the bulk. Since excess holes are accumulated at the surface, this is referred to as the *accumulation* condition. One can think of the excess holes as being attracted toward the surface by the negative gate voltage. An equal amount of negative charge appears on the metal side of the MOS capacitor, as required for charge neutrality.

On the other hand, if a positive voltage is applied to the gate of a p-type MOS capacitor, the metal Fermi level moves downward, which creates an oxide field in the direction of accelerating a negative charge toward the metal electrode. A similar field is induced in the silicon, which causes the bands to bend downward toward the surface, as

shown in Fig. 2.31(c). Since the valence band at the surface is now farther away from the Fermi level than is the valence band in the bulk, the hole concentration at the surface is lower than the concentration in the bulk. This is referred to as the *depletion* condition. One can think of the holes as being repelled away from the surface by the positive gate voltage. The situation is similar to the depletion layer in a p–n junction discussed in Section 2.2.2. The depletion of holes at the surface leaves the region with a net negative charge arising from the unbalanced acceptor ions. An equal amount of positive charge appears on the metal side of the capacitor.

As the positive gate voltage increases, the band bending also increases, resulting in a wider depletion region and more (negative) depletion charge. This goes on until the bands bend downward so much that at the surface, the conduction band is closer to the Fermi level than the valence band is, as shown in Fig. 2.31(d). When this happens, not only are the holes depleted from the surface, but the surface potential is such that it is energetically favorable for electrons to populate the conduction band. In other words, the surface behaves like n-type material with an electron concentration given by Eq. (2.49). ***Note that this n-type surface is formed not by doping, but instead by inverting the original p-type substrate with an applied electric field.*** This condition is called *inversion*. The negative charge in the silicon consists of both the ionized acceptors and the thermally generated electrons in the conduction band. Again, it is balanced by an equal amount of positive charge on the metal gate. The surface is inverted as soon as $E_i = (E_c + E_v)/2$ crosses E_f. This is called *weak inversion* because the electron concentration remains small until E_i is considerably below E_f. If the gate voltage is increased further, the concentration of electrons at the surface will be equal to, and then exceed, the hole concentration in the substrate. This condition is called *strong inversion*.

So far we have discussed the band bending for accumulation, depletion, and inversion of silicon surface in a p-type MOS capacitor. Similar conditions hold true in an n-type MOS capacitor, except that the polarities of voltage, charge, and band bending are reversed, and the roles of electrons and holes are interchanged. The band diagrams for flatband, accumulation, depletion, and inversion conditions of an n-type MOS capacitor are shown in Fig. 2.31(e)–(h), where the metal work function per electron charge ϕ_m is assumed to be equal to that of the n-type silicon, given by

$$\phi_s = \chi + \frac{E_g}{2q} - \psi_B, \tag{2.174}$$

instead of Eq. (2.169). Accumulation occurs when a positive voltage is applied to the metal gate and the silicon bands bend downward at the surface. Depletion and inversion occur when the gate voltage is negative and the bands bend upward toward the surface.

2.3.2 Electrostatic Potential and Charge Distribution in Silicon

2.3.2.1 Solving Poisson's Equation

In this sub-subsection, the relations among the surface potential, charge, and electric field are derived by solving Poisson's equation in the surface region of silicon. A more detailed band diagram at the surface of a p-type silicon is shown in Fig. 2.32. The potential $\psi(x) =$

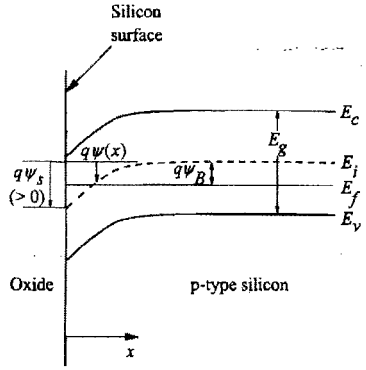


Figure 2.32. Energy-band diagram near the silicon surface of a p-type MOS device. The band bending ψ is defined as positive when the bands bend downward with respect to the bulk. Accumulation occurs when $\psi_s < 0$. Depletion and inversion occur when $\psi_s > 0$.

$\psi_i(x) - \psi_i(x = \infty)$ is defined as the amount of band bending at position x, where $x = 0$ is at the silicon surface and $\psi_i(x = \infty)$ is the intrinsic potential in the bulk silicon. Remember that $\psi(x)$ is positive when the bands bend downward. The boundary conditions are $\psi = 0$ in the bulk silicon, and $\psi = \psi(0) = \psi_s$ at the surface. The surface potential ψ_s depends on the applied gate voltage, as discussed in Section 2.3.1.2. Poisson's equation, Eq. (2.44), is

$$\frac{d^2\psi}{dx^2} = -\frac{d\mathscr{E}}{dx} = -\frac{q}{\varepsilon_{si}}[p(x) - n(x) + N_d^+(x) - N_a^-(x)]. \tag{2.175}$$

For a uniformly doped p-type substrate of acceptor concentration N_a with complete ionization, $N_d^+(x) - N_a^-(x) = -N_a$, independent of x. Charge neutrality condition deep in the bulk substrate requires

$$N_d^+(x) - N_a^-(x) = -N_a = -p_0 + \frac{n_i^2}{p_0}, \tag{2.176}$$

where $p_0 = p(x = \infty)$ and $n_i^2/p_0 = n(x = \infty)$ are the majority (holes) and minority (electrons) carrier densities in the bulk substrate, respectively. In general, $p(x)$ and $n(x)$ are given by Eq. (2.50) and Eq. (2.49), which can be expressed in terms of $\psi(x)$ using $\psi(x) = \psi_i(x) - \psi_i(x = \infty)$ and $\psi_B = \psi_f - \psi_i(x = \infty)$ as defined in Fig. 2.32:

$$p(x) = n_i e^{q[\psi_f - \psi_i(x)]/kT} = n_i e^{q[\psi_B - \psi(x)]/kT} = p_0 e^{-q\psi(x)/kT} \tag{2.177}$$

and

$$n(x) = n_i e^{q[\psi_i(x) - \psi_f]/kT} = n_i e^{q[\psi(x) - \psi_B]/kT} = \frac{n_i^2}{p_0} e^{q\psi(x)/kT}. \tag{2.178}$$

Note that ψ_f is independent of x because there is no net current flow perpendicular to the surface and the Fermi level stays flat. Also note that $p_0 = p(x = \infty) = n_i \exp(q\psi_B/kT)$ and $n_i^2/p_0 = n(x = \infty) = n_i \exp(-q\psi_B/kT)$ in the last step of the above equations.

In practice, $N_a \gg n_i$, and $p_0 \approx N_a$ from Eq. (2.176). Substituting the last three equations into Eq. (2.175) and replacing p_0 by N_a yield

$$\frac{d^2\psi}{dx^2} = -\frac{q}{\varepsilon_{si}}\left[N_a\left(e^{-q\psi/kT} - 1\right) - \frac{n_i^2}{N_a}\left(e^{q\psi/kT} - 1\right)\right]. \tag{2.179}$$

Multiplying $(d\psi/dx)\,dx$ on both sides of Eq. (2.179) and integrating from the bulk ($\psi = 0$, $d\psi/dx = 0$) toward the surface, one obtains

$$\begin{aligned}\int_0^{d\psi/dx} \frac{d\psi}{dx}\, d\left(\frac{d\psi}{dx}\right) \\ = -\frac{q}{\varepsilon_{si}}\int_0^{\psi}\left[N_a\left(e^{-q\psi/kT} - 1\right) - \frac{n_i^2}{N_a}\left(e^{q\psi/kT} - 1\right)\right]d\psi,\end{aligned} \tag{2.180}$$

which gives the electric field at x, $\mathscr{E} = -d\psi/dx$, in terms of ψ:

$$\begin{aligned}\mathscr{E}^2(x) = \left(\frac{d\psi}{dx}\right)^2 = \frac{2kTN_a}{\varepsilon_{si}}\left[\left(e^{-q\psi/kT} + \frac{q\psi}{kT} - 1\right)\right. \\ \left. + \frac{n_i^2}{N_a^2}\left(e^{q\psi/kT} - \frac{q\psi}{kT} - 1\right)\right].\end{aligned} \tag{2.181}$$

At $x = 0$, we let $\psi = \psi_s$ and $\mathscr{E} = \mathscr{E}_s$. From Gauss's law, Eq. (2.43), the total charge per unit area induced in the silicon (equal and opposite to the charge on the metal gate) is

$$\begin{aligned}Q_s = -\varepsilon_{si}\mathscr{E}_s = \pm\sqrt{2\varepsilon_{si}kTN_a}\left[\left(e^{-q\psi_s/kT} + \frac{q\psi_s}{kT} - 1\right)\right. \\ \left. + \frac{n_i^2}{N_a^2}\left(e^{q\psi_s/kT} - \frac{q\psi_s}{kT} - 1\right)\right]^{1/2}.\end{aligned} \tag{2.182}$$

This function is plotted in Fig. 2.33. At the flat-band condition, $\psi_s = 0$ and $Q_s = 0$. In accumulation, $\psi_s < 0$ (bands bending upward) and the first term in the square brackets dominates once $-q\psi_s/kT > 1$. The accumulation charge density is then proportional to $\exp(-q\psi_s/2kT)$ as indicated in the Fig 2.33.[5] In depletion, $\psi_s > 0$ and $q\psi_s/kT > 1$, but $\exp(q\psi_s/kT)$ is not large enough to make the n_i^2/N_a^2 term appreciable. Therefore, the $q\psi_s/kT$ term in the square brackets dominates and the negative depletion charge density (from ionized acceptor atoms) is proportional to $\psi_s^{1/2}$. When ψ_s increases further, the $(n_i^2/N_a^2)\exp(q\psi_s/kT)$ term eventually becomes larger than the $q\psi_s/kT$ term and dominates the square bracket. This is when inversion occurs. The negative inversion charge density is proportional to $\exp(q\psi_s/2kT)$ as indicated in Fig. 2.33.

A popular criterion for the onset of strong inversion is for the surface potential to reach a value such that $(n_i^2/N_a^2)\exp(q\psi_s/kT) = 1$, i.e.,

$$\psi_s(\text{inv}) = 2\psi_B = 2\frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right). \tag{2.183}$$

[5] These results stem from the Maxwell–Boltzmann approximation made in Section 2.1.1.3, which overestimates the occupancy of electron states near and below the Fermi level. For accumulation and inversion layers with carrier densities in the degenerate range, i.e., when $\psi_s < -0.2$ V or > 0.9 V where the Fermi level goes into the valence or the conduction band, the more exact Fermi–Dirac distribution gives a less steep rise of the sheet charge density with the surface potential.

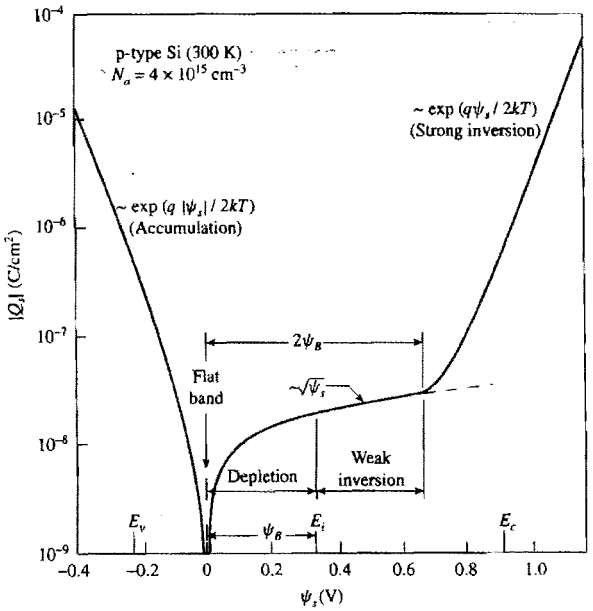


Figure 2.33. Variation of total charge density (fixed plus mobile) in silicon as a function of surface potential ψ_s for a p-type MOS device. The labels E_v, E_i, E_c indicate the surface potential values where the valence band, the intrinsic level, and the conduction band cross the Fermi level. (After Sze, 1981.)

Under this condition, the electron concentration given by Eq. (2.178) at the surface becomes equal to the depletion charge density N_a. ***After inversion takes place, even a slight increase in the surface potential results in a large buildup of electron density at the surface.*** The inversion layer effectively shields the silicon from further penetration of the gate field. Since almost all of the incremental charge is taken up by electrons, there is no further increase of either the depletion charge or the depletion-layer width. The expression in Eq. (2.183) is a rather weak function of the substrate doping concentration. For typical values of $N_a = 10^{16}$–10^{18} cm^{-3}, $2\psi_B$ varies only slightly, from 0.70 to 0.94 V.

2.3.2.2 Depletion Approximation

In general, Eq. (2.181) must be solved numerically to obtain $\psi(x)$. In particular cases, approximations can be made to allow the integral to be carried out analytically. For example, in the depletion region where $2\psi_B > \psi > kT/q$, only the $q\psi/kT$ term in the square bracket needs to be kept and

$$\frac{d\psi}{dx} = -\sqrt{\frac{2qN_a\psi}{\varepsilon_{si}}}. \tag{2.184}$$

One can then rearrange the factors and integrate:

$$\int_{\psi_s}^{\psi} \frac{d\psi}{\sqrt{\psi}} = -\int_0^x \sqrt{\frac{2qN_a}{\varepsilon_{si}}}\, dx, \tag{2.185}$$

where ψ_s is the surface potential at $x = 0$ as assumed before. Therefore,

$$\psi = \psi_s \left(1 - \sqrt{\frac{qN_a}{2\varepsilon_{si}\psi_s}}\, x\right)^2, \tag{2.186}$$

which can be written as

$$\psi = \psi_s \left(1 - \frac{x}{W_d}\right)^2. \tag{2.187}$$

This is a parabolic equation with the vertex at $\psi = 0$, $x = W_d$, where

$$W_d = \sqrt{\frac{2\varepsilon_{si}\psi_s}{qN_a}} \tag{2.188}$$

is the depletion-layer width defined as the distance to which the band bending extends. The total depletion charge density in silicon, Q_d, is equal to the charge per unit area of ionized acceptors in the depletion region:

$$Q_d = -qN_aW_d = -\sqrt{2\varepsilon_{si}qN_a\psi_s}. \tag{2.189}$$

These results are very similar to those of the one-sided abrupt p–n junction under the depletion approximation, discussed in Section 2.2.2. ***In the MOS case, however, W_d reaches a maximum value W_{dm} at the onset of strong inversion when $\psi_s = 2\psi_B$.*** Substituting Eq. (2.183) into Eq. (2.188) gives the maximum depletion width:

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a/n_i)}{q^2N_a}}. \tag{2.190}$$

2.3.2.3 Strong Inversion

Beyond strong inversion, the $(n_i^2/N_a^2)\exp(q\psi/kT)$ term representing the inversion charge in Eq. (2.181) becomes appreciable and must be kept, together with the depletion charge term:

$$\frac{d\psi}{dx} = -\sqrt{\frac{2kTN_a}{\varepsilon_{si}}\left(\frac{q\psi}{kT} + \frac{n_i^2}{N_a^2}e^{q\psi/kT}\right)}. \tag{2.191}$$

This equation can only be integrated numerically. The boundary condition is $\psi = \psi_s$ at $x = 0$. After $\psi(x)$ is solved, the electron distribution $n(x)$ in the inversion layer can be calculated from Eq. (2.178). Examples of numerically calculated $n(x)$ are plotted in Fig. 2.34 for two values of ψ_s with $N_a = 10^{16}$ cm^{-3}. ***The electrons are distributed extremely close to the surface with an inversion-layer width less than 50 Å.*** A higher surface potential or field tends to confine the electrons even closer to the surface. In

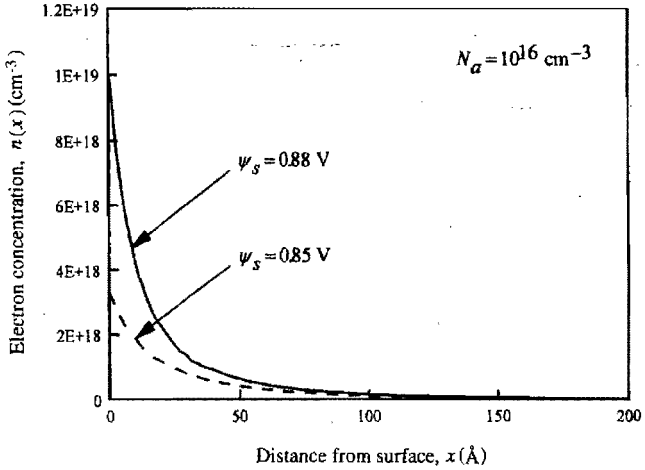


Figure 2.34. Electron concentration versus distance in the inversion layer of a p-type MOS device.

general, electrons in the inversion layer must be treated quantum-mechanically as a 2-D gas (Stern and Howard, 1967). According to the quantum-mechanical model, inversion-layer electrons occupy discrete energy bands and have a peak distribution 10–20 Å away from the surface. More details will be discussed in Section 4.2.4.

When the inversion charge density per unit area, Q_i, is much greater than the depletion charge density, Eq. (2.182) can be approximated by

$$Q_i = -\sqrt{\frac{2\varepsilon_{si}kTn_i^2}{N_a}}\, e^{q\psi_s/2kT}. \tag{2.192}$$

Since the electron concentration at the surface is

$$n(0) = \frac{n_i^2}{N_a}e^{q\psi_s/kT}, \tag{2.193}$$

one can write

$$|Q_i| = \sqrt{2\varepsilon_{si}kTn(0)}. \tag{2.194}$$

The effective inversion-layer thickness (classical model) can be estimated from $Q_i/qn(0) = 2\varepsilon_{si}kT/qQ_i$, which is inversely proportional to Q_i. Similar expressions also hold true for the surface charge density of extra holes under accumulation, except that the factor n_i^2/N_a is replaced by N_a.

2.3.2.4 Surface Potential and Charge Density as a Function of Gate Voltage

In Section 2.3.2.1, charge and potential distributions in silicon were solved in terms of the surface potential ψ_s as a boundary condition. ψ_s is not directly measurable, but is controlled by and can be determined from the applied gate voltage. The gate voltage

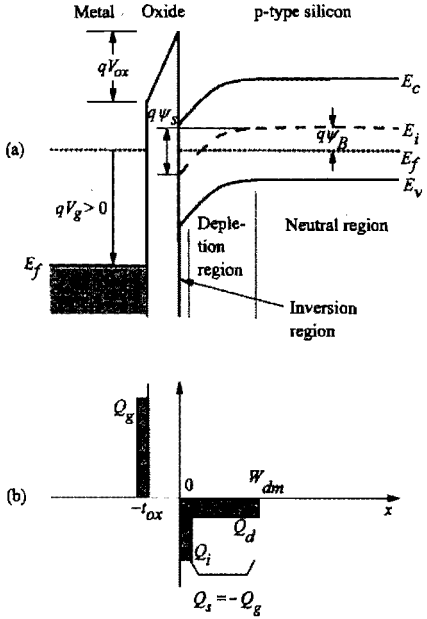


Figure 2.35. (a) Band diagram of a p-type MOS capacitor with a positive voltage applied to the gate ($V_{fb} = 0$). (b) Charge distribution under inversion condition.

equation, Eq. (2.172), relates the potential drop V_{ox} across the oxide and the band bending ψ_s in silicon to the departure from the flatband condition due to the applied gate voltage V_g (Fig. 2.35(a)). Assuming negligible fixed charges in the oxide, the potential drop V_{ox} can be expressed as $\mathscr{E}_{ox}t_{ox}$, which equals $(\varepsilon_{si}/\varepsilon_{ox})\mathscr{E}_s t_{ox}$ based on the boundary condition, Eq. (2.173). Applying Gauss's law, $Q_s = -\varepsilon_{si}\mathscr{E}_s$, ***the gate bias equation becomes***

$$V_g - V_{fb} = V_{ox} + \psi_s = \frac{-Q_s}{C_{ox}} + \psi_s, \tag{2.195}$$

where Q_s is the total charge per unit area induced in the silicon, and $C_{ox} = \varepsilon_{ox}/t_{ox}$ is the oxide capacitance per unit area for an oxide of thickness t_{ox}. There is a negative sign in front of Q_s in Eq. (2.195) because the charge on the metal gate is always equal but opposite to the charge in silicon, i.e., Q_s is negative when $V_g - V_{fb}$ is positive and vice versa. The charge distribution in an MOS capacitor is shown schematically in Fig. 2.35 (b), where the total charge Q_s may include both depletion and inversion components. For simplicity of discussion, oxide and interface trapped charges are ignored here. They will be discussed in detail in Sections 2.3.6 and 2.3.7.

In general, Q_s is a function of ψ_s given by Eq. (2.182), and plotted in Fig. 2.33. Equation (2.195) is then an implicit equation that can be solved for ψ_s as a function of V_g. An example of the numerical solution is shown in Fig. 2.36. Below the condition for strong inversion, $\psi_s = 2\psi_B$, ψ_s increases more or less linearly with V_g. Beyond $\psi_s = 2\psi_B$, ψ_s nearly saturates –

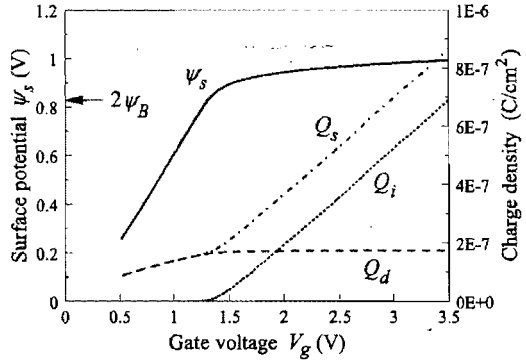


Figure 2.36. Numerical solutions of surface potential, total silicon charge density, inversion charge density, and depletion charge density from the gate voltage equation (2.195) coupled to Eq. (2.182). The MOS device parameters are $N_a = 10^{17}$ cm^{-3}, $t_{ox} = 10$ nm, and $V_{fb} = 0$.

increasing by less than 0.2 V while V_g increases by 2 V. After ψ_s is solved, Q_s is calculated and plotted as a function of V_g in Fig. 2.36. By numerically evaluating the integrals in Exercise 2.6, Q_s is separated into its two components, the depletion charge density Q_d and the inversion charge density Q_i, which are also plotted in Fig. 2.36. ***It is clear that before the $\psi_s = 2\psi_B$ condition, the charge in the silicon is predominantly of the depletion type.*** Under such depletion conditions, $Q_s(\psi_s) = Q_d(\psi_s)$, an analytical expression for $\psi_s(V_g)$ can be derived by solving a quadratic equation (see Eq. (2.202)). After $\psi_s = 2\psi_B$, the depletion charge no longer increases with V_g because of shielding by the inversion layer discussed before. ***Almost all of the increase of Q_s beyond $\psi_s = 2\psi_B$ is taken up by Q_i with a slope $dQ_i/dV_g \approx C_{ox}$.*** While on the linear scale it appears that Q_i is zero below the $\psi_s = 2\psi_B$ threshold, on the log scale it can be seen that Q_i actually remains finite and decreases exponentially with V_g. It is the source of the subthreshold leakage current in MOSFETs – an important design consideration further addressed in detail in Section 3.1.3.2. Under extreme accumulation and inversion conditions, $-Q_s \approx C_{ox}(V_g - V_{fb})$, since both V_g and V_{ox} can be much larger than the silicon bandgap, $E_g/q = 1.12$ V (for CMOS technologies with $V_{dd} \gg 1$ V), while ψ_s is at most comparable to E_g/q (surface potential pinned to either the valence band or the conduction band edge).

2.3.3 Capacitances in an MOS Structure

2.3.3.1 Definition of Small-Signal Capacitances

We now consider the capacitances in an MOS structure. In most cases, MOS capacitances are defined as small-signal differential of charge with respect to voltage or potential. They can easily be measured by applying a small ac voltage on top of a dc bias across the device and sensing the out-of-phase ac current at the same frequency (the in-phase component gives the small-signal conductance). The total MOS capacitance per unit area is

$$C_g = \frac{d(-Q_s)}{dV_g}. \tag{2.196}$$

If we differentiate Eq. (2.195) with respect to $-Q_s$ and define the silicon part of the capacitance as

$$C_{si} = \frac{d(-Q_s)}{d\psi_s}, \tag{2.197}$$

we obtain

$$\frac{1}{C_g} = \frac{1}{C_{ox}} + \frac{d\psi_s}{d(-Q_s)} = \frac{1}{C_{ox}} + \frac{1}{C_{si}}. \tag{2.198}$$

In other words, the total capacitance equals the oxide capacitance and the silicon capacitance connected in series. The capacitances are defined in such a way that they are all positive quantities. An equivalent circuit is shown in Fig. 2.37(a). In reality, there is also an interface trap capacitance in parallel with C_{si}. It arises from charging and discharging of Si–SiO_2 interface traps and will be discussed in more detail in Section 2.3.7.

2.3.3.2 Capacitance–Voltage Characteristics: Accumulation

A typical capacitance-versus-gate-voltage (C–V) curve of a p-type MOS capacitor is plotted in Fig. 2.38, assuming zero flatband voltage. In fact, there are several different curves, depending on the frequency of the applied ac signal. We start with the "low-frequency" or *quasistatic* C–V curve. When the gate voltage is negative (by more than a few kT/q) with respect to the flatband voltage, the p-type MOS capacitor is in accumulation and $Q_s \propto \exp(-q\psi_s/2kT)$, as shown in Fig. 2.33. Therefore, $C_{si} = -dQ_s/d\psi_s = (q/2kT)Q_s = (q/2kT)C_{ox}|V_g - V_{fb} - \psi_s|$, and the MOS capacitance per unit area is given by

$$\frac{1}{C_g} = \frac{1}{C_{ox}}\left[1 + \frac{2kT/q}{|V_g - V_{fb} - \psi_s|}\right]. \tag{2.199}$$

Since $2kT/q \approx 0.052$ V and ψ_s is limited to 0.1 to 0.3 V in accumulation, the MOS capacitance rapidly approaches C_{ox} when the gate voltage is 1–2 V more negative than the flat-band voltage.[6]

2.3.3.3 Capacitance at Flat Band

When the gate bias is zero in Fig. 2.38, the MOS is near the flat-band condition; therefore, $q\psi_s/kT \ll 1$. The inversion charge term in Eq. (2.182) can be neglected and the first exponential term can be expanded into a power series. Keeping only the first three terms of the series, one obtains $Q_s = -(\varepsilon_{si}q^2N_a/kT)^{1/2}\psi_s$. From Eq. (2.198), the flatband capacitance per unit area is given by

$$\frac{1}{C_{fb}} = \frac{1}{C_{ox}} + \sqrt{\frac{kT}{\varepsilon_{si}q^2N_a}} = \frac{1}{C_{ox}} + \frac{L_D}{\varepsilon_{si}}, \tag{2.200}$$

[6] Actually, C_g approaches C_{ox} slower than that depicted by Eq. (2.199) because of the Fermi–Dirac distribution at degenerate carrier densities.

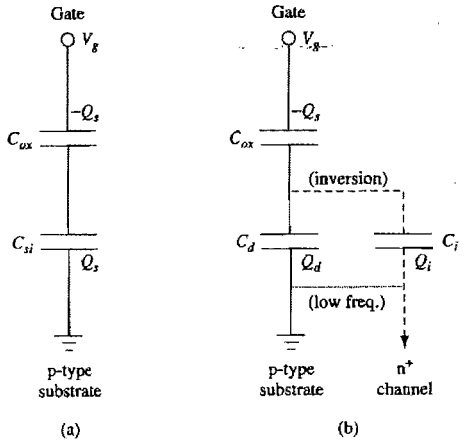


Figure 2.37. Equivalent circuits of an MOS capacitor. (a) All the silicon capacitances are lumped into C_{si}. (b) C_{si} is broken up into a depletion charge capacitance C_d and an inversion-layer capacitance C_i. C_d arises from the majority carriers, which can respond to high-frequency as well as low-frequency signals. C_i arises from the minority carriers, which can only respond to low-frequency signals, unless the surface inversion channel is connected to a reservoir of minority carriers as in a gated diode configuration. The thin dotted connection in (b) is effective only at low frequencies where minority carriers can respond.

where L_D is the Debye length defined in Eq. (2.53). In most cases, C_{fb} is somewhat less than C_{ox}. For very thin oxides and low substrate doping, C_{fb} can be much smaller than C_{ox}.

2.3.3.4 Capacitance–Voltage Characteristics: Depletion

When the gate voltage is slightly higher than the flatband voltage in a p-type MOS capacitor, the surface starts to be depleted of holes; $1/C_{si}$ becomes appreciable and the capacitance decreases. Using the depletion approximation, one can find an analytical expression for C_g in this case. From Eq. (2.188) and Eq. (2.189),

$$C_d = \frac{d(-Q_d)}{d\psi_s} = \sqrt{\frac{\varepsilon_{si}qN_a}{2\psi_s}} = \frac{\varepsilon_{si}}{W_d}. \tag{2.201}$$

The last expression is identical to the depletion-layer capacitance per unit area in the p–n junction case discussed in Section 2.2.2. The bias equation (2.195) becomes

$$V_g - V_{fb} = \frac{qN_aW_d}{C_{ox}} + \psi_s = \frac{\sqrt{2\varepsilon_{si}qN_a\psi_s}}{C_{ox}} + \psi_s. \tag{2.202}$$

Substituting C_d from Eq. (2.201) for C_{si} in Eq. (2.198), and eliminating ψ_s using Eq. (2.202), one obtains

$$C_g = \frac{C_{ox}}{\sqrt{1 + \left[2C_{ox}^2(V_g - V_{fb})/(\varepsilon_{si}qN_a)\right]}}. \tag{2.203}$$

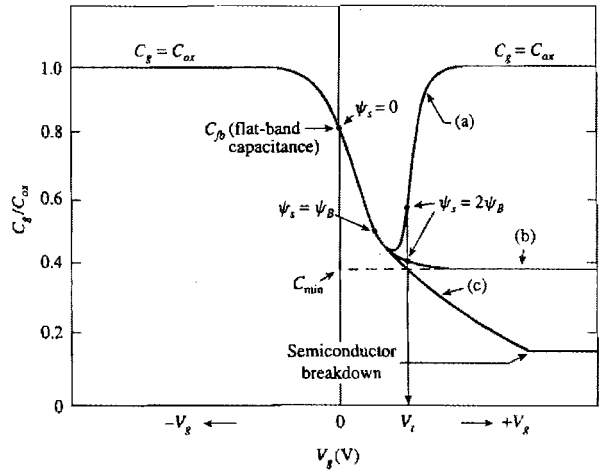


Figure 2.38. MOS capacitance-voltage curves: (a) low frequency, (b) high frequency, (c) deep depletion. $V_{fb} = 0$ is assumed. (After Size, 1981.)

This equation shows how the MOS capacitance decreases with increasing V_g under the depletion condition. It serves as a good approximation to the middle portions of the C–V curves in Fig. 2.38, provided that the MOS capacitor is not biased near the flat-band or the inversion condition.

2.3.3.5 Low-Frequency C–V Characteristics: Inversion

As the gate voltage increases further, however, the capacitance stops decreasing when $\psi_s = 2\psi_B$ [Eq. (2.183)] is reached and inversion occurs. ***Once the inversion layer forms, the capacitance starts to increase, since C_{si} is now given by the variation of the inversion charge with respect to ψ_s, which is much larger than the depletion capacitance.*** Assuming that the silicon charge is dominated by the inversion charge, one can carry out an approximation as in the accumulation case and show that the MOS capacitance in strong inversion is also given by Eq. (2.199). One difference is that ψ_s at inversion is in the range of 0.7 to 1.0 V, significantly higher than that at accumulation. In any case, the capacitance rapidly increases back to C_{ox} when the gate voltage is more than 2 to 3 V beyond the flat-band voltage, as shown in the low-frequency C–V curve (a) in Fig. 2.38.

2.3.3.6 High-Frequency Capacitance–Voltage Characteristics

The above discussion of the low-frequency MOS capacitance assumes that the minority carrier, i.e., the inversion charge, is able to follow the applied ac signal. This is true only if the frequency of the applied signal is lower than the reciprocal of the minority-carrier response time. The minority-carrier response time can be estimated from the generation–recombination current density, $J_R = qn_iW_d/\tau$, where τ is the minority-carrier lifetime discussed in Section 2.1.4. The time it takes to generate enough minority carriers to

replace something comparable to the depletion charge, $Q_d = qN_aW_d$, is on the order of $Q_d/J_R = (N_a/n_i)\tau$ (Jund and Poirier, 1966). This is typically 0.1–10 s. Therefore, ***for frequencies higher than 100 Hz or so, the inversion charge cannot respond to the applied ac signal.*** Only the depletion charge (majority carriers) can respond to the signal, which means that the silicon capacitance is given by C_d of Eq. (2.201) with W_d equal to its maximum value, W_{dm}, in Eq. (2.190). The high-frequency capacitance per unit area thus approaches a constant minimum value, C_{min}, at inversion given by

$$\frac{1}{C_{min}} = \frac{1}{C_{ox}} + \sqrt{\frac{4kT\ln(N_a/n_i)}{\varepsilon_{si}q^2N_a}}. \tag{2.204}$$

This is shown in the high-frequency C–V curve (b) in Fig. 2.38.

Typically, C–V curves are traced by applying a slow-varying ramp voltage to the gate with a small ac signal superimposed on it. However, if the ramp rate is fast enough that the ramping time is shorter than the minority-carrier response time, then there is insufficient time for the inversion layer to form, and the MOS capacitor is biased into deep depletion as shown by curve (c) in Fig. 2.38. In this case, the depletion width can exceed the maximum value given by Eq. (2.190), and the MOS capacitance decreases further below C_{min} until impact ionization takes place (Sze, 1981). Note that deep depletion is not a steady-state condition. If an MOS capacitor is held under such bias conditions, its capacitance will gradually increase toward C_{min} as the thermally generated minority charge builds up in the inversion layer until an equilibrium state is established. The time it takes for an MOS capacitor to recover from deep depletion and return to equilibrium is referred to as the *retention time*. It is a good indicator of the defect density in the silicon wafer and is often used to qualify processing tools in a facility.

It is possible to obtain *low-frequency-like* C–V curves at high measurement frequencies. One way is to expose the MOS capacitor to intense illumination, which generates a large number of minority carriers in the silicon. Another commonly used technique is to form an n^+ region adjacent to the MOS device and connect it electrically to the p-type substrate (Grove, 1967). The n^+ region then acts like a reservoir of electrons which can exchange minority carriers freely with the inversion layer. In other words, the n^+ region is connected to the *surface channel* of the inverted MOS device. This structure is similar to that of a gated diode, to be discussed in Section 2.3.5. Based on the equivalent circuit in Fig. 2.37(b), the total MOS capacitance per unit area is given by

$$C_g = \frac{C_{ox}(C_d + C_i)}{C_{ox} + C_d + C_i}. \tag{2.205}$$

When the MOS device is biased well into strong inversion, the *inversion-layer capacitance* C_i can be approximated by

$$C_i = \frac{d(-Q_i)}{d\psi_s} = \frac{|Q_i|}{2kT/q} \tag{2.206}$$

using Eq. (2.192).

The majority and minority carrier contributions to the total capacitance can be separately measured in a *split* C–V setup shown in Fig. 2.39(a) (Sodini *et al.*, 1982).

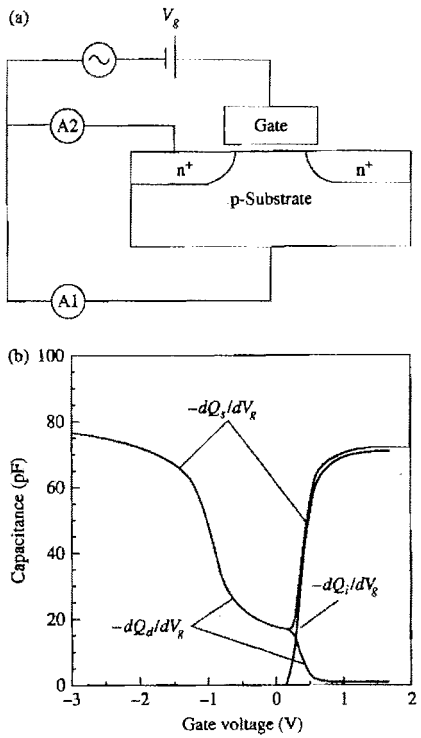


Figure 2.39. (a) Setup of the split C–V measurement. Both the dc bias and the small-signal ac voltage are applied to the gate. Small signal ac currents are measured by two ammeters, A1 and A2, connected separately as shown. (b) Measured C–V curves where the $-dQ_d/dV_g$ component is obtained from A1, and the $-dQ_i/dV_g$ component is obtained from A2. The sum is the total capacitance per unit area, $-dQ_s/dV_g$.

With a small signal ac voltage applied to the gate, the out-of-phase ac currents are sensed by two ammeters: one (A1) connected to the p-type substrate for the hole current, and another (A2) connected to the n^+ region for the electron current. Typical measured results are shown in Fig. 2.39(b). The hole contribution to the capacitance measured by A1 is

$$-\frac{dQ_d}{dV_g} = \frac{C_{ox}C_d}{C_{ox}+C_d+C_i}. \tag{2.207}$$

And the electron contribution to the capacitance measured by A2 is

$$-\frac{dQ_i}{dV_g} = \frac{C_{ox}C_i}{C_{ox}+C_d+C_i}. \tag{2.208}$$

They add up to the total capacitance per unit area, $C_g = -dQ_s/dV_g$. Note that the $-dQ_d/dV_g$ curve decreases to zero soon after strong inversion when C_i (Eq. (2.206)) becomes

dominant ($\gg C_{ox}$). To put it in another way, the highly conductive inversion channel shields the majority carriers in the bulk silicon so they do not respond to the modulation of gate field. The $-dQ_i/dV_g$ curve can be integrated to yield the inversion charge density as a function of the gate voltage. It is used, for example, in channel mobility measurements where the inversion charge density must be determined accurately.

2.3.4 Polysilicon-Gate Work Function and Depletion Effects

2.3.4.1 Work Function and Flatband Voltage of Polysilicon Gates

In the mainstream CMOS VLSI technology thus far, n^+-polysilicon gate has been used for nMOSFET and p^+-polysilicon gate for pMOSFET to obtain threshold voltages of low magnitude in both devices. The Fermi level of heavily doped n^+ polysilicon is near the conduction band edge, so its work function is given by the electron affinity, $q\chi$. From Eq. (2.169), the work function difference for an n^+ polysilicon gate on a p-type substrate of doping concentration N_a is

$$\phi_{ms} = -\frac{E_g}{2q} - \psi_B = -0.56 - \frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right) \tag{2.209}$$

in volts. Similarly, the work function difference for a p^+ polysilicon gate on an n-type substrate of doping concentration N_d is

$$\phi_{ms} = \frac{E_g}{2q} + \psi_B = 0.56 + \frac{kT}{q}\ln\left(\frac{N_d}{n_i}\right), \tag{2.210}$$

which is symmetric to Eq. (2.209). These relations give rise to flatband voltages with key implications on the scalability of MOSFET devices, as will be discussed in Chapter 4.

The band diagram of an n^+-polysilicon-gated p-type MOS capacitor at zero gate voltage is shown in Fig. 2.40(a), where the Fermi levels line up and the free electron level of the bulk p-type silicon is higher in electron energy than the free electron level of the n^+ polysilicon gate. This sets up an oxide field in the direction of accelerating electrons toward the gate, and at the same time a downward bending of the silicon bands (depletion) toward the surface to produce a field in the same direction. The flatband condition is reached by applying a negative voltage equal to the work function difference to the gate, as shown in Fig. 2.40(b).

2.3.4.2 Polysilicon-Gate Depletion Effects

The use of polysilicon gates is a key advance in modern CMOS technology, since it allows the source and drain regions to be self-aligned to the gate, thus eliminating parasitics from overlay errors (Kerwin *et al.*, 1969). However, if the polysilicon gate is not doped heavily enough, problems can arise from depletion of the gate itself. This is especially a concern with the dual n^+–p^+polysilicon-gate process in which the gates are doped by ion implantation (Wong *et al.*, 1988). Gate depletion results in an additional capacitance in series with the oxide capacitance, which in turn leads to a reduced inversion-layer charge density and degradation of the MOSFET transconductance.

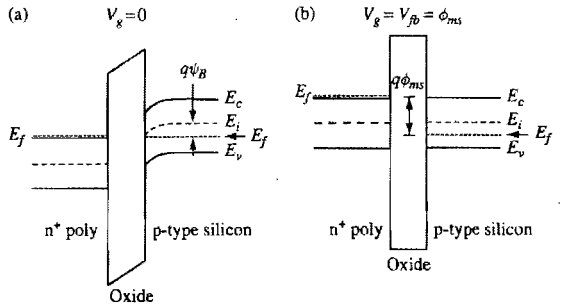


Figure 2.40. Band diagram of an n^+-polysilicon-gated p-type MOS capacitor biased at (a) zero gate voltage and (b) flatband condition.

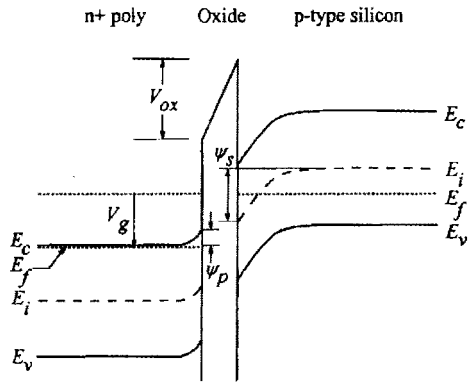


Figure 2.41. Band and potential diagram showing polysilicon-gate depletion effects when a positive voltage is applied to the n^+ polysilicon gate of a p-type MOS capacitor.

The analysis of MOS capacitance in the last subsection can be extended to include the polysilicon depletion effect and quantify certain observed features in the *C–V* characteristics. Consider the band diagram of an n^+-polysilicon-gated p-type MOS capacitor biased into inversion as shown in Fig. 2.41. Since the oxide field points in the direction of accelerating a negative charge toward the gate, the bands in the n^+ polysilicon bend slightly upward toward the oxide interface. This depletes the surface of electrons and forms a thin space-charge region in the polysilicon layer, which lowers the total capacitance.

2.3.4.3 Effect of Polysilicon Doping Concentration on *C–V* Characteristics

Typical low-frequency *C–V* curves in the presence of gate depletion effects are shown in Fig. 2.42 (Rios and Arora, 1994). A distinct feature is that the capacitance at inversion does not return to the full oxide capacitance as in Fig. 2.38. Instead, the inversion capacitance exhibits a maximum value somewhat less than C_{ox}, depending on the

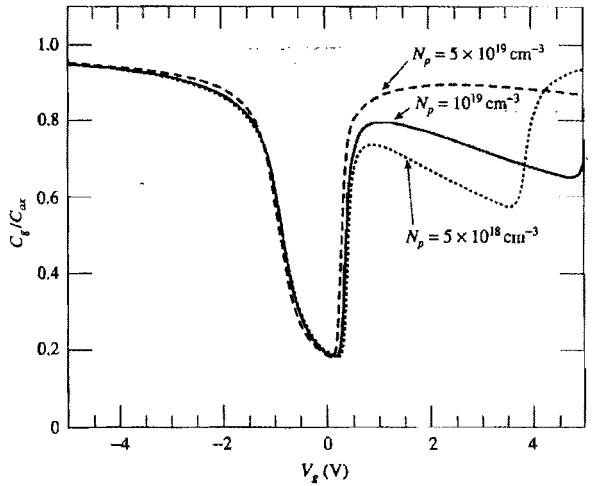


Figure 2.42. Low-frequency *C–V* curves of a p-type MOS capacitor with n^+ polysilicon gate doped at several different concentrations. (After Rios and Arora, 1994.)

effective doping concentration of the polysilicon gate. The higher the doping concentration, the less the gate depletion effect is and the closer the maximum capacitance is to the oxide capacitance.

The existence of a local maximum in the low-frequency *C–V* curve can be understood semiquantitatively as follows. In Fig. 2.41, we assume ψ_s to be the amount of band bending in the bulk silicon and ψ_p to be that in the n^+polysilicon. From charge neutrality, the total charge density Q_g of the ionized donors in the depletion region of the n^+ polysilicon gate is equal and opposite to the combined inversion and depletion charge density Q_s in the silicon substrate, i.e., $Q_g = -Q_s$. The gate bias equation for an applied voltage V_g is obtained by adding an additional term, ψ_p, for the band bending in the polysilicon gate to Eq. (2.195):

$$V_g = V_{fb} + \psi_s + \psi_p - \frac{Q_s}{C_{ox}}. \tag{2.211}$$

Differentiating Eq. (2.211) with respect to $-Q_s$ and using the capacitance definitions (2.196) and (2.197), we obtain

$$\frac{1}{C_g} = \frac{1}{C_{ox}} + \frac{1}{C_{si}} + \frac{1}{C_p}, \tag{2.212}$$

where $C_p = -dQ_s/d\psi_p = dQ_g/d\psi_p$ is the capacitance of the polysilicon depletion region.

When a p-type MOS device is biased well into strong inversion such that Q_s is dominated by the inversion charge, the low-frequency capacitance is approximately given by Eq. (2.206), $C_{si} = |Q_s|/(2kT/q) = Q_g/(2kT/q)$. Based on the depletion

approximation, Eqs. (2.189) and (2.201), the polysilicon depletion capacitance can be expressed in terms of the depletion charge density Q_g as $C_p = \varepsilon_{si} q N_p / Q_g$, where N_p is the doping concentration of the polysilicon gate. Substituting these expressions into Eq. (2.212) yields

$$\frac{1}{C_g} = \frac{1}{C_{ox}} + \frac{2kT/q}{Q_g} + \frac{Q_g}{\varepsilon_{si} q N_p}. \tag{2.213}$$

As V_g becomes more positive, C_{si} ($\propto Q_g$) increases but C_p ($\propto 1/Q_g$) decreases. This results in a local maximum of the low-frequency capacitance as observed in Fig. 2.42. For $N_p < 10^{19}$ cm^{-3}, another abrupt rise in the MOS capacitance may be observed at a much higher gate voltage, as shown in Fig. 2.42. This is due to the onset of inversion (to p^+) at the n^+ polysilicon surface.

In practice, polysilicon gates cannot be doped much higher than 1×10^{20} cm^{-3}. Modern CMOS devices often operate at a maximum oxide field of 5 MV/cm, corresponding to a sheet electron density of $Q_g/q = 10^{13}$ cm^{-2}. For these values, the polysilicon depletion capacitance has the equivalent effect as adding 3–4 Å to the gate oxide thickness. The above discussion applies to p^+ polysilicon gates on n-type silicon as well as to n^+ polysilicon gates on p-type silicon. Note that for n^+ polysilicon gates on n-type silicon and p^+ polysilicon gates on p-type silicon, gate depletion occurs when the substrate is accumulated.

2.3.5 MOS under Nonequilibrium and Gated Diodes

An important building block of VLSI devices is the gated diode, or gate-controlled diode. Consider an MOS capacitor where there is an n^+ region adjacent to the gated p-type region (Grove, 1967). The n^+ region and the p-type region form an n^+–p diode. This structure is shown schematically in Fig. 2.43 and was mentioned briefly at the end of Section 2.3.3.

2.3.5.1 Inversion Condition of an MOS Under Nonequilibrium

As discussed in Section 2.2.1, when both the n^+ region and the p-type substrate are connected to the same potential (grounded), the p–n junction is in equilibrium and the Fermi level is constant across the p–n junction. If the gate voltage is large enough to invert the p-type surface, which occurs for a surface potential bending of $\psi_s(\text{inv}) = 2\psi_B$, the inverted channel is connected to the n^+ region and has the same potential as the n^+ region. In other words, ***the electron quasi-Fermi level in the channel region is the same as the Fermi level in the n^+ region*** as well as the Fermi level in the p-type substrate. The depletion region now extends from the p–n junction to the region under the gate between the inverted channel and the substrate, as shown in Fig. 2.43(b).

If, on the other hand, the p–n junction is reverse-biased at a voltage V_R as shown in Fig. 2.44, the MOS is in a nonequilibrium condition in which $np \neq n_i^2$. From Eq. (2.107), the electron concentration on the p-type side of the junction is

$$n = \frac{n_i^2}{N_a} e^{-qV_R/kT}, \tag{2.214}$$

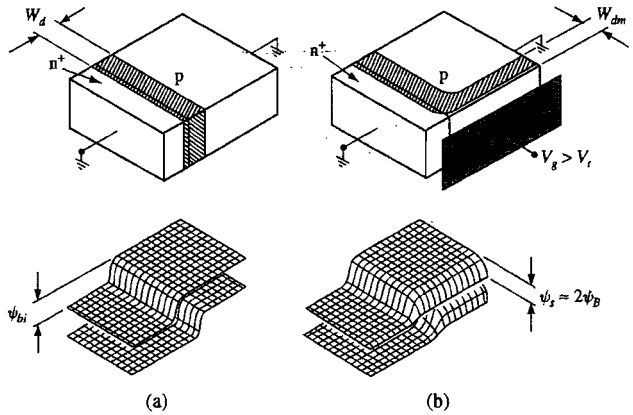


Figure 2.43. Gated diode or p-type MOS with adjacent n^+ region in equilibrium (zero voltage across the p–n junction). The gate is biased at (a) flatband and (b) inversion conditions. (After Grove, 1967.)

since the Fermi level of the n^+ region is now qV_R lower than the Fermi level in the p-type substrate.[7] Now consider the case when a positive voltage large enough to bend the bands by $2\psi_B$ is applied to the gate, as shown in Fig. 2.44(b). This brings the conduction band at the surface $2\psi_B$ closer to the electron quasi-Fermi level. From Eq. (2.178), the electron concentration at the surface is increased by a factor of $\exp(2q\psi_B / kT) = (N_a/n_i)^2$ over that of Eq. (2.214), i.e.,

$$n = \frac{n_i^2}{N_a} e^{2q\psi_B/kT} e^{-qV_R/kT} = N_a e^{-qV_R/kT}. \tag{2.215}$$

Since this is much lower than the depletion charge density N_a, the surface remains depleted. Even though the positive voltage is sufficient to invert the surface in the equilibrium case, it is not enough to cause inversion in the reverse-biased case. This is because ***the reverse bias lowers the quasi-Fermi level of electrons so that even if the bands at the surface are bent as much as in the equilibrium case in Fig. 2.43(b), the conduction band is still not close enough to the quasi-Fermi level of electrons for inversion to occur.***

To reach inversion in the nonequilibrium case, a much larger gate voltage, sufficient to bend the bands by $2\psi_B + V_R$, must be applied. This is the case shown in Fig. 2.44(c), where the electron concentration at the surface is now

$$n = \frac{n_i^2}{N_a} e^{q(2\psi_B + V_R)/kT} e^{-qV_R/kT} = N_a, \tag{2.216}$$

the same as the condition for inversion introduced in Section 2.3.2. Notice that the surface depletion layer is much wider than in the equilibrium case, just as in a reverse-biased p–n junction.

[7] While this is not exactly true, it will not affect later results. Further discussions follow Eq. (2.216).

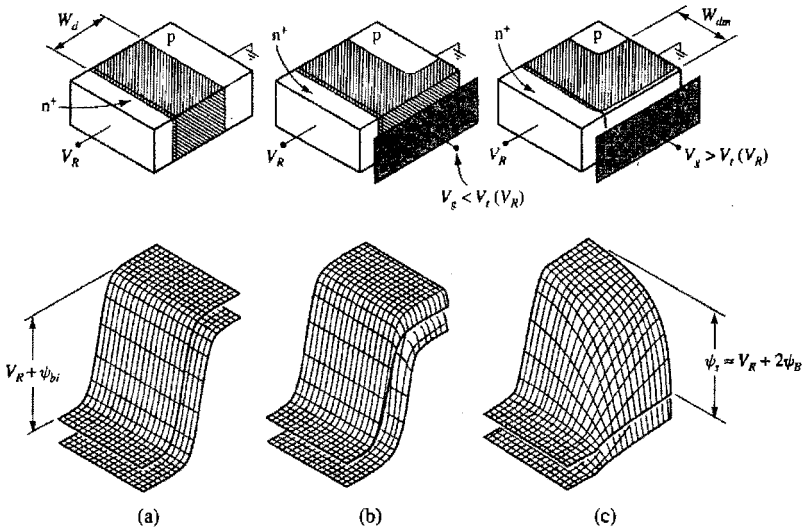


Figure 2.44. Gated diode or p-type MOS with adjacent n^+ region under nonequilibrium (reverse bias across the p–n junction). The gate is biased at (a) flat-band, (b) depletion, and (c) inversion conditions. (After Grove, 1967.)

To be more exact, Eq. (2.107), on which Eq. (2.214) is based, is not valid under a large reverse bias V_R. It is discussed in Appendix 4 that at large reverse biases, the electron quasi-Fermi level on the p-type side of the depletion region (at $x=0$ in Fig. A4.5) is higher than the Fermi level of the n-type region (to the left of $x=-W_d$). In other words, the electron quasi-Fermi level on the p-type side of the depletion region is displaced downward from the Fermi level of the p-type region by less than qV_R. While this is true at the flatband condition in the gated diode depicted in Fig. 2.44(a), once a positive gate voltage is applied to bend down the p-type bands as in Fig. 2.44(c), the electron concentration at the p-type surface increases and the electron quasi-Fermi level there moves down toward the Fermi level of the n-type region. As far as the bands and the electron concentration at the p-type surface are concerned, the positive gate voltage has a similar effect as moving from the p-type side of the depletion region (at $x=0$) toward the n-type side of the depletion region (at $x=-W_d$) in Fig. A4.5. When the threshold condition is reached, the electron quasi-Fermi level at the p-type surface would be the same as the Fermi level of the n-type region, i.e., displaced downward from the Fermi level of the p-type region by exactly qV_R. Another way to see this is that, in a region where the electron concentration is appreciable, electron quasi-Fermi level must remain essentially flat since the leakage current in a reverse-biased diode, $J_n \propto n\, d\,\phi_n/dx$, is negligibly small. In any case, the band bending requirement for threshold depicted in Fig. 2.44(c), $\psi_s = V_R + 2\psi_B$, is always valid.[8]

[8] This condition applies to the neutral p-type region. Less band bending is needed to reach the threshold for a point at the surface but within the depletion region of the reverse biased p–n junction, as is evident in Fig. 2.44(c).

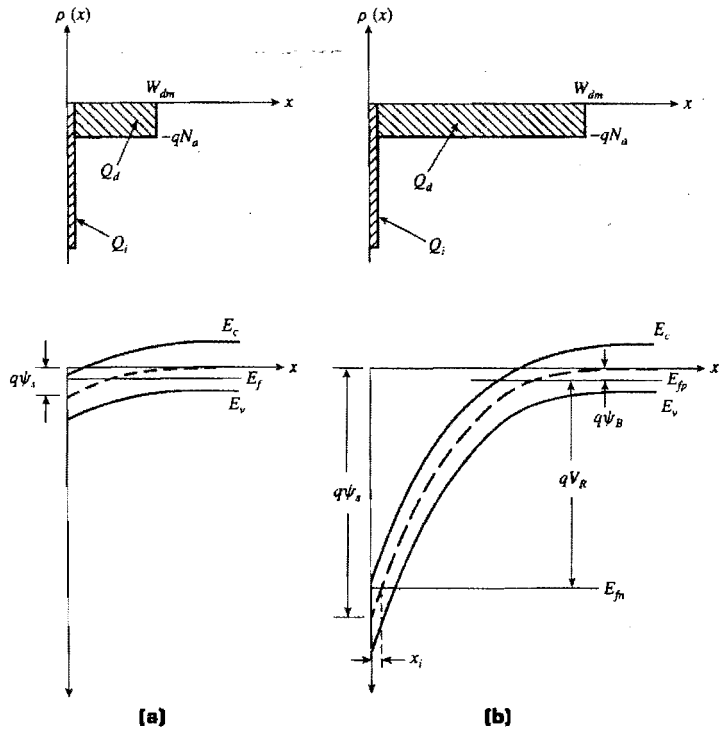


Figure 2.45. Comparison of charge distribution and energy-band variation of an inverted p-type region for (a) the equillibrium case and (b) the nonequillibrium case. (After Grove, 1967.)

2.3.5.2 Band Bending and Charge Distribution of an MOS Under Nonequilibrium

The above discussions are further illustrated in Fig. 2.45, where the charge distribution and band bending in a cross section perpendicular to the gate through the neutral p-type region are shown for both the equilibrium and the nonequilibrium cases. The equilibrium case is the same as that discussed in Section 2.3.2. In the nonequilibrium case, the hole quasi-Fermi level is the same as the Fermi level in the bulk p-type silicon, but the electron quasi-Fermi level is dictated by the Fermi level in the n^+ region (not shown in Fig. 2.45), which is qV_R lower than the p-type Fermi level. As a result, surface inversion occurs at a band bending

$$\psi_s(\text{inv}) = V_R + 2\psi_B, \tag{2.217}$$

and the maximum depletion width is a function of the reverse bias V_R,

$$W_{dm} = \sqrt{\frac{2\varepsilon_{si}(V_R + 2\psi_B)}{qN_a}}, \tag{2.218}$$

from Eq. (2.188).

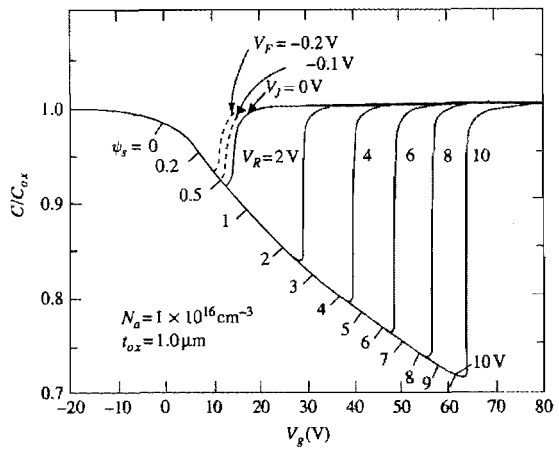


Figure 2.46. Normalized C–V characteristics of a p-type MOS capacitor with an adjacent n^+ region as that shown in Fig. 2.44(c). V_g is the voltage applied to the gate with respect to the p-type substrate. A series of C–V curves are shown for a range of V_R: the reverse bias voltage applied to the n^+/p junction.

When the surface is depleted, the gated diode behaves like an n^+–p diode with depletion of the p-region extending to underneath the gate electrode. When the surface is inverted, the gated diode behaves like an n^+–p diode with both the n^+ region and depletion of the p-region extending to underneath the gate electrode. High-field effects in gated diodes are discussed in Section 2.5.5

For a p-type MOS capacitor, the effect of an adjacent reverse biased n^+ region on the C–V characteristics is shown in Fig. 2.46. With increasing V_g, the surface potential ψ_s also increases, as labeled under the curve. At $V_R = 0$, the C–V curve resembles a regular low-frequency C–V curve (curve (a) in Fig. 2.38) with inversion (sharp rise of the capacitance to C_{ox}) taking place at $V_g \approx 13$ V where $\psi_s \approx 0.7$ V ($2\psi_B$). Note that as V_R increases, the onset of inversion shifts to increasingly more positive gate voltages as the MOS goes into deeper and deeper depletion (lower C_{min}). It can be seen that the value of surface potential at inversion increases by approximately V_R, consistent with Eq. (2.217). The decrease of C_{min}, the serial combination of C_{ox} and $C_d = \varepsilon_{si}/W_{dm}$, with V_R follows from Eq. (2.218).

2.3.6 Charge in Silicon Dioxide and at the Silicon–Oxide Interface

It is often said that the real magic in silicon technology lies not in the silicon crystalline material but in silicon dioxide. Silicon dioxide forms critical components of silicon devices, serves as insulation and passivation layers, and is often used as an effective masking and/or diffusion-barrier layer in device fabrication.

Thus far we have treated silicon dioxide as an ideal insulator, with no space charge in or associated with it, and no charge exchange between it and the silicon it covers. The

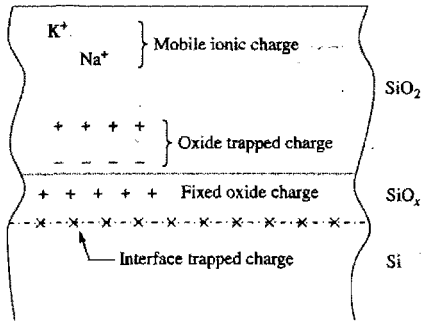


Figure 2.47. Charges and their location in thermally oxidized silicon. (After Deal, 1980.)

silicon dioxide and the oxide–silicon interface in real devices are never completely electrically neutral. There can be mobile ionic charges, electrons, or holes trapped in the oxide layer. There can also be fabrication-process-induced fixed oxide charges near the oxide–silicon interface, and charges trapped at the so-called *surface states* at the oxide–silicon interface. Electrons and holes can make transitions from the crystalline states near the oxide–silicon interface to the surface states, and vice versa. Since every device has some regions that are covered by silicon dioxide, the electrical characteristics of a device are very sensitive to the density and properties of the charges inside its oxide regions and at its silicon–oxide interface.

The nomenclature for describing the charges associated with the silicon dioxide in real devices was standardized in 1978 (Deal, 1980). The net charge per unit area is denoted by Q. Thus, Q_m denotes the *mobile charge* per unit area, Q_{ot} denotes the *oxide trapped charge* per unit area, Q_f denotes the *fixed oxide charge* per unit area, and Q_{it} denotes the *interface trapped charge* per unit area. The names and locations of these charges are illustrated in Fig. 2.47. The properties and characteristics of these charges are discussed further below.

2.3.6.1 Surface States and Interface Trapped Charge

At the Si–SiO_2 interface, the lattice of bulk silicon and all the properties associated with its periodicity terminate. As a result, localized states with energy in the forbidden energy gap of silicon are introduced at or very near the Si–SiO_2 interface (Many *et al.*, 1965). These localized surface states are illustrated schematically in Fig. 2.48. Interface trapped charges are electrons or holes trapped in these states.

Just like impurity energy levels in bulk silicon discussed in Section 2.1.2, the probability of occupation of a surface state by an electron or by a hole is determined by the surface-state energy relative to the Fermi level. Thus, as the surface potential is changed, the energy level of a surface state, which is fixed relative to the energy-band edges at the surface, moves with it. This change relative to the Fermi level causes a change in the probability of occupation of the surface state by an electron. For instance, referring to Fig. 2.32, as the bands are bent downward, or as the surface potential is increased,

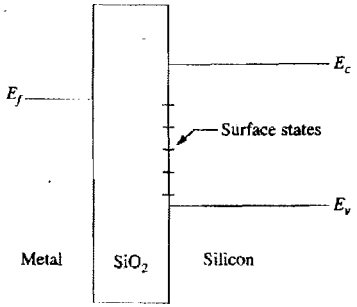


Figure 2.48. Schematic energy-band diagram of an MOS structure, illustrating the presence of surface states.

more surface states move below the Fermi level and hence become occupied by electrons. This change of interface trapped charge with a change in the surface potential gives rise to an additional silicon capacitance component, which will be discussed further in Section 2.3.7.

Electrons in silicon but near an oxide–silicon interface can make transitions between the conduction-band states and the surface states. An electron in the conduction band can contribute readily to electrical conduction current, while an electron in a surface state, i.e., an interface trapped electron, does not contribute readily to electrical conduction current, except by *hopping* among the surface states or by first making a transition to the conduction band. Similarly, holes in silicon but near an oxide–silicon interface can make transitions between the valence-band states and the surface states, and trapped interface holes do not contribute readily to electrical conduction. By trapping electrons and holes, surface states can reduce the conduction current in MOSFETs. Furthermore, the trapped electrons and holes can act like charged scattering centers, located at the interface, for the mobile carriers in a surface channel, and thus lower their mobility (Sah *et al.*, 1972).

Surface states can also act like localized generation–recombination centers. Depending on the surface potential, a surface state can first capture an electron from the conduction band, or a hole from the valence band. This captured electron can subsequently recombine with a hole from the valence band, or the captured hole can recombine with an electron from the conduction band. In this way, the surface state acts like a recombination center. Similarly, a surface state can act like a generation center by first emitting an electron followed by emitting a hole, or by first emitting a hole followed by emitting an electron. Thus, the presence of surface states can lead to surface generation–recombination leakage currents.

The density of surface states, and hence the density of interface traps, is a function of silicon substrate orientation and a strong function of the device fabrication process (EMIS, 1988; Razouk and Deal, 1979). In general, for a given device fabrication process, the dependence of the interface trap density on substrate orientation is $\langle 100 \rangle < \langle 110 \rangle < \langle 111 \rangle$. Also, a postmetallization or "final" anneal in hydrogen, or in a hydrogen-containing ambient, at temperatures around 400 °C is quite effective in minimizing the density of interface traps. Consequently, $\langle 100 \rangle$ silicon and postmetallization anneal in hydrogen are commonly used in modern VLSI device fabrication.

2.3.6.2 Fixed Oxide Charge

Fixed oxide charges are positive charges located in the oxide layer very close to the $Si–SiO_2$ interface. In fact, for modeling purposes, the fixed oxide charges are usually assumed to be located at the $Si–SiO_2$ interface. They are primarily due to excess silicon species introduced during oxidation and during postoxidation heat treatment (Deal *et al.*, 1967). The dependence of the density of fixed oxide charges on substrate orientation is the same as that of interface traps, namely $\langle 100 \rangle < \langle 110 \rangle < \langle 111 \rangle$.

The presence of fixed oxide charges at the oxide–silicon interface affects the potential in the silicon, which will be discussed in the next subsection. In addition, the fixed oxide charges act as charged scattering centers and thus reduce the mobility of the carriers in a surface inversion channel (Sah *et al.*, 1972).

2.3.6.3 Mobile Ionic Charge

Mobile ionic charges in SiO_2 are usually due to sodium or potassium contamination introduced during device fabrication. Unlike fixed oxide charges, which are not mobile, Na^+ and K^+ ions are quite mobile in SiO_2 and can be moved from one end of the oxide layer to the other when an electric field is applied across the oxide layer, particularly at somewhat elevated temperatures (>200 °C) (Hillen and Verwey, 1986). As these positively charged ions drift close to the $Si–SiO_2$ interface, they repel holes from, and attract electrons to, the silicon surface, often causing unwanted surface electron current to flow among n^+ diffusion regions in a p-type substrate or well. Also, when these positively charged ions come close to the silicon surface, they can act as charged scattering centers for the carriers in the surface inversion channel, thus reducing their mobility.

In VLSI fabrication processes, ***mobile-ion contamination problems must be avoided.*** This is accomplished by a combination of proper passivation, usually using phosphosilicate glass, and "clean" fabrication technology (Hillen and Verwey, 1986).

2.3.6.4 Oxide Trapped Charge

If electron–hole pairs are generated in an oxide layer, e.g., by ionizing radiation, some of these electrons and holes can be subsequently trapped in the oxide. Also, if electrons or holes are injected into an oxide layer, by tunneling or by hot-carrier injection, some of them can be trapped in the oxide.

Electron and hole traps in SiO_2 can easily be introduced by bombardment with high-energy photons or particles (Bourgoin, 1989). Since bombardment by high-energy particles and photons is involved in many steps in the fabrication of modern VLSI devices (during ion implantation, plasma or reactive-ion etching, sputtering deposition, electron-beam evaporation of metal, electron-beam and x-ray lithography, etc.), electron and hole traps are often introduced in the oxide during device fabrication. Fortunately, most of these traps can be eliminated with subsequent anneals at temperatures above 550 °C (Ning, 1978).

Also, depending on the oxidation condition, electron traps can be introduced during the oxide growth process itself (EMIS, 1988). For example, oxide growth in moisture-containing ambient is known to introduce electron traps (Nicollian *et al.*, 1971).

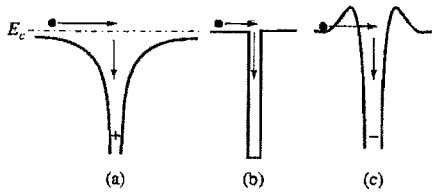


Figure 2.49. Schematics illustrating the potential wells of electron traps in silicon dioxide: (a) Coulomb-attractive trap, (b) neutral trap, and (c) Coulomb-repulsive trap.

- *Capture cross section.* Traps are usually characterized by their *capture cross sections.* Electron traps with cross sections in the range of 10^{-14}–10^{-12} cm^2 are usually *Coulomb-attractive* traps, i.e., the trap centers are positively charged prior to electron capture (Ning *et al.*, 1975; Lax, 1960). Electron traps with cross sections in the 10^{-18}–10^{-14}-cm^2 range are usually due to neutral traps (Lax, 1960), and those with cross sections smaller than 10^{-18} cm^2 are usually associated with *Coulomb-repulsive* traps, i.e., the trap centers are already negatively charged prior to electron capture (Balland and Barbottin, 1989). The potential wells representing these electron traps are illustrated in Fig. 2.49. Since the Coulomb-attractive and neutral centers have the largest capture cross sections, they are also the most important to include when considering the effects of electron traps on device characteristics. Hole traps have not been studied in as much detail as electron traps. This may be due to the fact that holes are very readily trapped when they are injected into an oxide layer (Goodman, 1966). This is consistent with the measured hole capture cross section of about 3×10^{-13} cm^2, which is as large as the largest electron traps in SiO_2 (Ning, 1976a).
- *Temperature dependence.* Consider the capture of a mobile electron into an electron trap. The trapping process has two competing components, namely, capturing the electron into some initial high-energy state of the trap center, and reemitting that same electron from the initial captured state by thermal excitation. If an electron in an initial captured state has a higher probability of cascading down towards the ground state of the trap center than of being reemitted, the electron becomes trapped. On the other hand, if the probability of reemission by thermal excitation from the initial captured state is high enough, trapping will not occur (Lax, 1960). The capture cross section, therefore, decreases with increasing temperature, since the probability for thermal reemission increases with temperature (Lax, 1960).
- *Field dependence.* If an electric field is applied across an oxide layer, it has the effect of increasing the energy of the carriers moving in the oxide layer. As these carriers gain energy from the oxide field, the probability of their being captured in some initial trap state is lowered, since the carriers now must lose more energy in the initial capture process. At the same time, an oxide field has the effect of lowering the energy barriers for the carriers trapped in a potential well, thus increasing the probability for reemitting them from their initial captured states (Lax, 1960). As a result, the capture cross section decreases with increasing oxide field (Ning, 1976b, 1978). ***The commonly used method of injecting carriers into SiO_2 by tunneling at high oxide fields tends to underestimate the amount of traps*** in SiO_2, since the capture cross sections at such high oxide fields are much smaller than those at normal device operation.

2.3.7 Effect of Interface Traps and Oxide Charge on Device Characteristics

The presence of oxide charges and interface traps has three major effects on the characteristics of devices. First, the charge in the oxide, or in the interface traps, interacts with the charge in the silicon near the surface and thus changes the silicon charge distribution and the surface potential. Second, as the density of interface trapped charge changes with changes in the surface potential, it gives rise to an additional capacitance component in parallel with the silicon capacitance C_{si} discussed in Section 2.3.3. Third, the interface traps can act as generation–recombination centers, or assist in the band-to-band tunneling process, and thus contribute to the leakage current in a gated-diode structure. These effects are discussed more quantitatively below.

2.3.7.1 Effect of Oxide Charge on Surface Potential

As discussed in Section 2.3.2, the charge distribution in silicon is a function of the surface potential. Thus, the effect of oxide charge on the charge distribution in silicon can be described in terms of its effect on surface potential. In the case of an MOS structure, the effect of oxide charge is usually described in terms of the change in gate voltage, which is a readily measurable parameter, necessary to counter the effect of the oxide charge or to restore the surface potential to that of zero oxide charge.

For simplicity of illustration, let us consider an MOS structure biased at flat-band condition. Let us assume that a sheet of oxide charge Q per unit area is placed at a distance x from the gate electrode, and a gate voltage δV_g has been applied to restore the MOS structure to its original, i.e., flat-band, condition. With the surface potential restored to its original value, the sheet of oxide charge has induced no change in the charge distribution in the silicon, which is a function of the surface potential, but a charge of magnitude $-Q$ per unit area on the gate electrode. This is illustrated in Fig. 2.50. Gauss's law in Eq. (2.43) implies that the electric field in the oxide between 0 and x due to the sheet of oxide charge and its image charge on the gate electrode is $-Q / \varepsilon_{ox}$ (see Exercise 2.5). This is also illustrated in Fig. 2.50. The potential difference supporting this electric field is $-xQ / \varepsilon_{ox}$, which is provided by the applied gate voltage. Therefore,

$$\delta V_g = -\frac{xQ}{\varepsilon_{ox}}. \tag{2.219}$$

The gate voltage necessary to offset the effect of an arbitrary oxide charge distribution can be obtained by superposition of individual elements of the charge distribution and applying Eq. (2.219) to each element. For an oxide charge distribution of $\rho_{net}(x, \psi_s) = \rho_{net}(x) + Q_{it}(\psi_s)\delta(x - t_{ox})$, which consists of an arbitrary distribution $\rho_{net}(x)$ that is independent of the surface potential and a delta-function distribution of the interface trap charge $Q_{it}(\psi_s)$ located at $x = t_{ox}$, the gate voltage necessary to offset it is

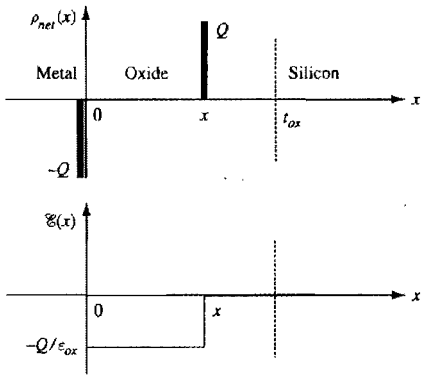


Figure 2.50. Schematic illustrating the effect of a sheet charge of areal density Q within the oxide layer of an MOS capacitor biased at flat-band condition.

$$\Delta V_g = \Delta V_g(\psi_s) = -\frac{1}{\varepsilon_{ox}}\int_0^{t_{ox}} x\rho_{net}(x,\psi_s)dx$$
$$= -\frac{1}{\varepsilon_{ox}}\left(\int_0^{t_{ox}} x\rho_{net}(x)dx + Q_{it}(\psi_s)t_{ox}\right). \tag{2.220}$$

According to the charge nomenclature discussed in Subsection 2.3.6, $\rho_{net}(x)$ includes the mobile charge, the oxide trapped charge, and the fixed oxide charge. It is a common practice to define an *equivalent oxide charge per unit area*, Q_{ox}, by

$$Q_{ox} = Q_{ox}(\psi_s) \equiv \int_0^{t_{ox}} \frac{x}{t_{ox}}\rho_{net}(x,\psi_s)dx$$
$$= \int_0^{t_{ox}} \frac{x}{t_{ox}}\rho_{net}(x)dx + Q_{it}(\psi_s). \tag{2.221}$$

Equation (2.220) can then be rewritten in the simple form of

$$\Delta V_g(\psi_s) = -\frac{Q_{ox}(\psi_s)}{C_{ox}}, \tag{2.222}$$

where $C_{ox} = \varepsilon_{ox}/t_{ox}$ is the oxide capacitance per unit area introduced in Eq. (2.195). Equation (2.222) states that the effect of an arbitrary oxide charge distribution is equivalent to an oxide sheet charge of areal density $Q_{ox}(\psi_s)$ located at the oxide–silicon interface.

2.3.7.2 Interface-Trap Capacitance

In Section 2.3.3, the silicon part of the capacitance is defined without including any interface trapped charge. As the interface traps are filled and emptied in response to changes in the surface potential, they give rise to an *interface-trap capacitance per unit area*, C_{it}, defined by

$$C_{it}(\psi_s) \equiv -\frac{dQ_{it}(\psi_s)}{d\psi_s}. \tag{2.223}$$

Just as in Eq. (2.197), the − sign in Eq. (2.223) is inserted to ensure that the capacitance is always a positive quantity. In Eq. (2.223), we have indicated explicitly that the interface-trap capacitance is a function of surface potential.

To include the effect of Q_{ox} in the operation of an MOS capacitor, Eq. (2.195) should be modified by adding to its right-hand side a ΔV_g term due to Q_{ox}. That is, Eq. (2.195) becomes

$$V_g - V_{fb} = \Delta V_g(\psi_s) - \frac{Q_s(\psi_s)}{C_{ox}} + \psi_s$$
$$= -\frac{Q_s(\psi_s) + Q_{ox}(\psi_s)}{C_{ox}} + \psi_s. \tag{2.224}$$

The total charge on the gate electrode is now $Q_s(\psi_s) + Q_{ox}(\psi_s)$. Equation (2.196) then becomes

$$C_g = -\frac{d[Q_s(\psi_s) + Q_{ox}(\psi_s)]}{dV_g}, \tag{2.225}$$

and Eq. (2.198) becomes

$$\frac{1}{C_g} = \frac{1}{C_{ox}} + \frac{1}{C_{si} + C_{it}}. \tag{2.226}$$

That is, the interface-trap capacitance is in parallel with the silicon capacitance.

As discussed in the previous subsection, the probability of a surface state being filled with an electron is governed by its energy level relative to the Fermi level. ***Only those interface traps that can be filled and emptied at a rate faster than the capacitance-measurement signal can contribute to C_{it}.*** Traps too slow to follow the capacitance-measurement signal will not contribute to C_{it}.

2.3.7.3 *C–V* Curves as a Monitoring and Diagnostic Tool for Oxide and Interface Quality

As discussed in Section 2.3.6, the amount of oxide charge and surface states is a function of the silicon substrate orientation and a strong function of the device fabrication process. For modern MOS devices, by the time a device fabrication process is ready for manufacturing, the amount of oxide charge and surface states is usually quite low, with Q_{ox}/q typically about 10^{11} cm^{-2} or less. For an MOS device having an oxide thickness of 10 nm, the corresponding gate voltage shift according to Eq. (2.222) is only 46 mV. At such low surface-state densities, the MOS *C–V* characteristics are quite ideal in that the measured *C–V* curves match well with the calculated ones (see Section 2.3.3 and Fig. 2.38).

However, many experimental fabrication processes, particularly those involving high-energy plasma, reactive-ion etching, and electron or ion beams, can generate significant oxide charge and surface states in MOS devices. Unless these *oxide damages* can be removed by post-process thermal annealing, the amount of residual oxide charge and surface states can be appreciable. High-field stress of silicon MOS devices can also

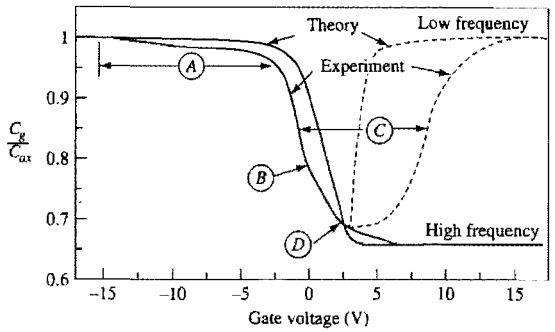


Figure 2.51. Comparison of experimental and theoretical high-frequency and low-frequency C–V curves, showing typical distortion caused by interface traps. The MOS capacitor has $N_a = 10^{16}$ cm^{-3} and $t_{ox} = 200$ nm. The symbols are explained in the text. (After Deal *et al.*, 1969.)

generate oxide charge and surface states. (High-field effects will be discussed in Section 2.5.) The presence of oxide charge and surface states can cause the measured C–V curve to appear distorted compared to the ideal C–V curve. There are two contributions to this distortion. First, Q_{it} is a part of Q_{ox} which shifts the C–V curve along the gate voltage axis according to Eq. (2.224). The shift is distorted because Q_{it} is a function of surface potential, which in turn is a function of gate voltage. Second, the additional capacitance due to the interface traps also distorts the measured C–V curve because C_{it} is a function of surface potential, which in turn is a function of gate voltage.

If the amount of oxide charge and surface states is large, the distortions in the C–V curve can be quite prominent, as illustrated in Fig. 2.51 (Deal *et al.*, 1969). The physical mechanisms responsible for the various distorted regions can be understood as follows. The distortion labeled A is where the MOS capacitor is normally in accumulation. In this gate voltage region, the valence-band edge at the silicon–oxide interface approaches or crosses the Fermi level (see Fig. 2.31). As a result, the interface states near the valence band become ionized and positively charged. (The interface states near the valence band are called donor states. They are neutral when they lie below the Fermi level and become positively charged by donating electrons when they lie above the Fermi level.) As the donor interface states become ionized, they contribute to a build up of positive interface trap charge which shifts the gate voltage in the negative direction according to Eq. (2.224). The distortion near the label B is related to interface states near the midgap, since it occurs at a gate voltage range where the MOS capacitor is between flatband and weak-inversion conditions (see Figs 2.31 and 2.38). The distortion labeled D is where the MOS capacitor is near weak inversion. To the right side of D, the capacitor is in inversion where the conduction-band edge at the silicon–oxide interface approaches or crosses the Fermi level. In this gate voltage range, the interface states near the conduction band become ionized and negatively charged. (The interface states near the conduction band are called acceptor states. They are neutral when they lie above the Fermi level and become negatively charged by accepting electrons when they lie below the Fermi level.) As the acceptor interface states become ionized, they contribute to a build up of negative interface trap charge which shifts

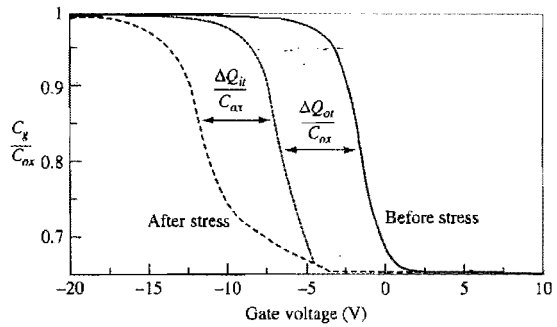


Figure 2.52. Typical high-frequency C–V plot of an MOS capacitor showing the distortion due to interface traps. The MOS capacitor has $N_a = 10^{16}$ cm^{-3} and $t_{ox} = 200$ nm. The oxide trapped charge and interface trapped charge are caused by subjecting the capacitor to a negative bias stress of 2MV/cm at 400 °C for 2 minutes. (After Deal *et al.*, 1967.)

the gate voltage in the positive direction according to Eq. (2.224). This causes the low-frequency C–V curve to shift to the right. The broadening of the C–V curve at its midpoint is labeled C. It is a result of the interface states near the conduction band (Deal *et al.*, 1969).

Figure 2.52 illustrates the distortion of a typical high-frequency C–V curve of an MOS capacitor after trapped charge has been created inside the oxide layer and at the oxide–silicon interface (Deal *et al.*, 1967). (The creation of bulk oxide and interface traps by high electric fields will be discussed in Section 2.5.) The oxide trapped charge causes a parallel shift of the C–V curve (dotted line) to the left. The interface-trap capacitance causes the curve to be distorted and shifted to the left by an additional amount.

The C–V distortions depicted in Figs 2.51 and 2.52 are for 200 nm thick oxides having significant oxide charge and surface states. It can be inferred from Eqs. (2.221) and (2.222) that the magnitude of the gate voltage shift caused by a certain areal density of interface states, Q_{it}, is proportional to t_{ox}. The voltage shift caused by a certain uniform volume density of oxide trapped charge, ρ_{net}, is proportional to t_{ox}^2. Therefore, for the same Q_{it} and ρ_{net}, the C–V curves of thinner oxide devices should appear less distorted.

There is a vast amount of published literature on the subject of interface states and the measurement of interface states. Interested readers are referred to the literature for a discussion on the general characteristics of interface states in MOS capacitors (Deal *et al.*, 1969) and on the various techniques for measuring interface states (Schroder, 1990).

2.3.7.4 Surface Generation–Recombination Centers

As discussed in the previous subsection, interface states can serve as generation–recombination centers. In the case of a gated-diode structure, the surface generation–recombination current adds to the diode leakage current. The magnitude of the surface leakage current depends on whether or not the surface states are *exposed*, i.e., whether or not the silicon surface is depleted (Grove and Fitzgerald, 1966). If the surface is inverted, the surface states are all filled with minority carriers and do not function efficiently as generation centers. Similarly, if the surface is in accumulation, the surface states are all

filled with majority carriers and do not function efficiently as generation centers either. ***Only when the silicon surface is depleted will the surface states function efficiently as generation centers***. Thus, surface leakage current can be suppressed by biasing the gate to keep the silicon surface either in inversion or in accumulation. The reader is referred to Appendix 5 for a detailed discussion of the physics involved in generation–recombination processes.

As recombination centers, surface states can degrade the minority-carrier lifetime of devices. Consequently, devices where long minority-carrier lifetimes are required are usually designed to confine the minority carriers in them away from the silicon surface. In addition, the device fabrication processes are usually optimized to minimize the density of surface states.

2.3.7.5 Surface-State or Trap-Asssisted Band-to-Band Tunneling

As will be discussed in Subsection 2.5.2, band-to-band tunneling occurs when the electric field across a p–n junction is sufficiently large. For a gated diode, or for a p–n diode with silicon surface components, the presence of surface states or interface traps in the high-field region can enhance the band-to-band tunneling current very significantly. Thus, gate-induced drain leakage currents in MOSFETs, which will be discussed in Section 2.5.5, and emitter–base diode tunneling currents in bipolar transistors, which will be discussed in Section 6.3.4, depend strongly on the density of interface states at the oxide–silicon interface of these devices.

2.4 Metal–Silicon Contacts

The metal–semiconductor contact is a critically important element in all semiconductor devices and technology. As a contact to a silicon device terminal, a metal–silicon contact should be non-rectifying and have a small contact resistance in order to minimize the voltage drop across the contact. Such contacts are usually referred to as *ohmic contacts*. In general, a metal–semiconductor contact has rectifying current–voltage characteristics similar to those of a p–n diode (see Section 2.2). Rectifying metal–semiconductor devices are called *Schottky diodes* or *Schottky barrier diodes*. Here we discuss the basic physics and operation of a metal–silicon contact, focusing on its current–voltage characteristics as a Schottky diode and as an ohmic contact. A brief discussion of Schottky diodes as active devices is also given.

2.4.1 Static Characteristics of a Schottky Barrier Diode

The static characteristics of a Schottky barrier diode can be inferred from those of an MOS capacitor (see Section 2.3) by letting the oxide layer thickness go to zero. Just as the surface potential of the semiconductor in an MOS capacitor is affected by the interface trapped charge, the surface potential of the semiconductor in a Schottky barrier diode is affected by the electron occupation of the surface states on the semiconductor surface. Therefore, the characteristics of a Schottky barrier diode depend on the properties of the metal and the properties of the semiconductor and its surface states. Here we first discuss

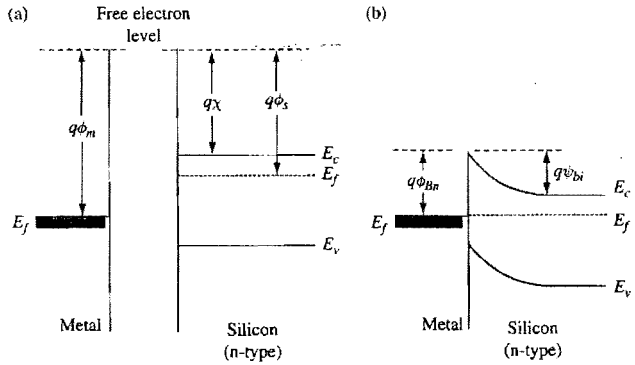


Figure 2.53. Energy-band diagrams of a metal–silicon system where the silicon surface is assumed to be absent of any surface states. (a) When the metal and the silicon are far apart. (b) When the metal is in contact with the silicon, with no externally applied voltage.

the static characteristics of a Schottky barrier diode ignoring all surface states, and then discuss how the surface states can modify the diode characteristics.

2.4.1.1 Schottky Barrier Diodes Without Surface States

When surface states are ignored, the energy-band diagram of an MOS capacitor shown in Fig. 2.29 can be readily adapted to give the energy-band diagram of a metal–silicon system. It was pointed out in Section 2.3.1.1 that when two different materials are brought into contact, they must share the same free electron level at the interface. Also, it was shown in Sections 2.1.1.3 and 2.1.4.5 that at thermal equilibrium or when there is no net electron or hole current through a system, the Fermi level of the system is spatially constant. These two factors together lead to the energy-band diagrams in Fig. 2.53 for a metal–n-silicon contact at thermal equilibrium. From consideration of free electron level at the interface, the Schottky barrier height for electrons, $q\phi_{Bn}$, is

$$q\phi_{Bn} = q(\phi_m - \chi), \tag{2.227}$$

where $q\phi_m$ is the metal work function and $q\chi$ is the electron affinity of silicon. From consideration of Fermi level being spatially constant, the built-in potential ψ_{bi} is

$$\begin{aligned} q\psi_{bi} &= q\phi_{bn} - (E_c - E_f)_{bulk} \\ &= q\left(\phi_{bn} - \frac{E_g}{2q} + \psi_B\right), \end{aligned} \tag{2.228}$$

where $\psi_B = |\psi_f - \psi_i|$ [see Eq. (2.48)].

The built-in potential implies an amount of depletion-layer charge Q_d in the silicon given by Eq. (2.189). To maintain overall charge neutrality of the metal–silicon system, this depletion charge induces a sheet of electronic charge of the same density per unit area in the metal at the metal–silicon interface.

The physical picture for a Schottky diode illustrated in Fig. 2.53 is based on energy-band diagrams for *bulk* silicon and metal. It does not take into consideration any surface properties of the semiconductor. Experimentally measured barrier heights are not consistent with Eq. (2.227). For some semiconductors, the measured barrier heights show little dependence on metal work function. For others, the dependence on work function is weaker than suggested by Eq. (2.227). That the measured barrier height has a weaker dependence on metal work function than suggested by Eq. (2.227) is attributable to the presence of surface states. This is discussed in the next subsection.

2.4.1.2 Schottky Barrier Diodes with Surface States

Many ideas and models for improving the understanding of real Schottky diodes have been proposed. In terms of explaining the relationship between measured barrier heights and metal work functions, the models all include the effects of surface states. Here we discuss qualitatively how surface states can influence the barrier height. A good reference where many ideas and models are discussed at length can be found in Henisch (1984).

The inclusion of surface states makes the physical picture of a metal–semiconductor contact more complex than the bulk model in Fig. 2.53. This is illustrated in Fig. 2.54. Let us first consider the situation where the metal and the silicon are physically and electrically two separate systems, i.e., Fig. 2.54(a). As electrically separate systems, there can be no charge exchange between the metal and the silicon. Occupation of some of the surface states by electrons induces a positive depletion-layer charge of Q_d per unit area in the silicon near the surface, causing the energy bands to bend upward near the surface. The amount of band bending is represented by the surface potential ψ_s. The upward band bending means $\psi_s < 0$. The relationship between the depletion charge density and surface potential is given by Eq. (2.189), namely

$$-q\psi_s(\text{free surface}) = \frac{[Q_d(\text{free surface})]^2}{2\varepsilon_{si}N_d}, \tag{2.229}$$

where N_d is the doping concentration of the n-type silicon. In Eq. (2.229), we have indicated that the surface potential and the corresponding depletion charge are for a free semiconductor surface.

Next, let us consider the metal and the silicon being electrically connected but physically separated, with a physical gap between the metal and silicon surfaces. (In the literature, for purposes of establishing a model for a Schottky diode, this physical gap is often replaced by an oxide layer. In this case, the discussion follows that for an MOS capacitor in Section 2.3.1.1.) The energy bands for this electrically connected system are as illustrated in Fig. 2.54(b). As discussed in Sections 2.1.1.3 and 2.1.4.5, a charge exchange between the metal and the silicon occurs until the Fermi level is spatially constant across the entire system. As discussed in Section 2.3.1.1, the free electron levels in the interfaces are continuous, suggesting the presence of an electric field in the gap between the surfaces of the metal and the silicon. This electric field is supported by a *net* charge Q_M on the metal surface, as required by Gauss's law [Eq. (2.43)]. Q_M is a negative charge for the electric field direction indicated in Fig. 2.54(b). If the amount of net charge

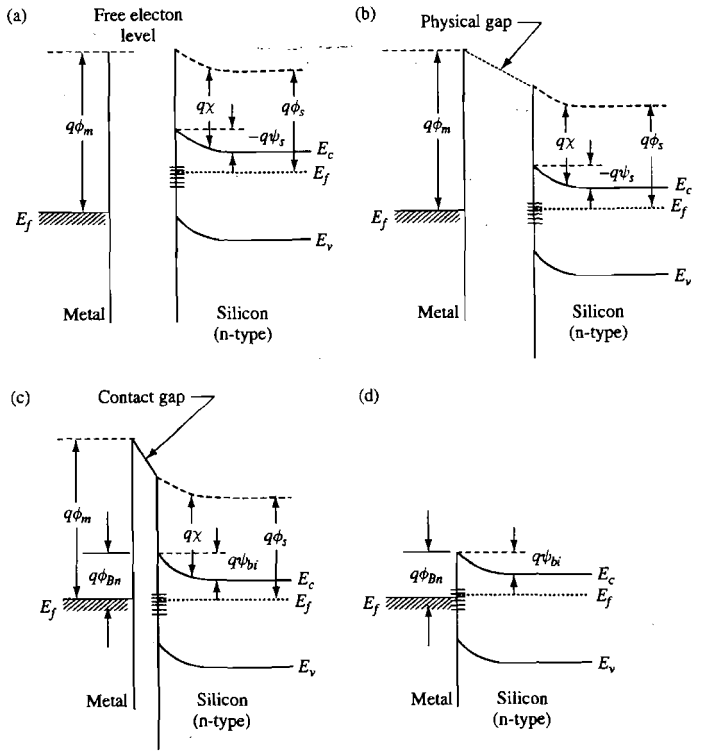


Figure 2.54. Energy-band diagram of a metal–silicon system where the silicon is assumed to have a large density of surface states. (a) When the metal and the silicon are not electrically connected as one system. Some of the interface states are filled with electrons, causing the bands to bend upward. (b) When the metal and the silicon are electrically connected to form one system, but the metal is physically separate from the silicon surface by a gap space. (c) When the metal is in contact with the silicon to form a Schottky diode. A contact gap of atomic dimension is shown. The contact-gap region is discussed in the text. (d) A simplified diagram where the contact gap is omitted.

per unit area in the surface states is Q_{it}, then overall charge neutrality of the metal–silicon system requires that $Q_M = -(Q_{it} + Q_d)$. As the gap between the metal surface and the silicon surface is reduced, the electric field in the gap increases, and hence the magnitude of Q_M increases, requiring the magnitude of $(Q_{it} + Q_d)$ to increase. In other words, a change in Q_M can cause a change in Q_{it} and/or Q_d.

The Bardeen model (Bardeen, 1947) for the roles played by surface states provides a physical picture for explaining how the charges Q_M, Q_{it} and Q_d may change together. According to the Bardeen model, the charges involved in the charge-transfer process in the formation of a metal–semiconductor contact come from four double layers. (A double layer consists of a layer of positive charge and a layer of negative charge of the same magnitude.) These layers are: (i) a double layer of atomic dimensions at the metal

surface, (ii) a double layer of atomic dimensions at the semiconductor surface, (iii) a double layer formed from the surface charges on the metal and semiconductor, both of atomic dimensions, and (iv) a double layer formed from a surface charge of atomic dimensions on the semiconductor surface and a depletion charge layer in the semiconductor. According to this model, Q_M has contributions from double layers (i) and (iii); Q_{it} has contributions from double layers (ii), (iii) and (iv); Q_d has contribution from double layer (iv).

Bardeen showed that the strength of double layer (iii) is small for semiconductors where the density of surface states is small (less than about 10^{13} cm^{-2}), and a change in Q_M is balanced by charge transfer primarily among double layers (ii) and (iv). In this case, as the physical gap in Fig. 2.54(b) is reduced, the change in Q_M is balanced primarily by a change in Q_d, with little change in Q_{it}. For semiconductors with sufficiently high (greater than 10^{13} cm^{-2}) density of surface states, double layer (iii) can be the primary source for any change in Q_M. In this case, as the physical gap in Fig. 2.54(b) is reduced, the change in Q_M is balanced primarily by a change in Q_{it}, with little change in Q_d. When there is little change in Q_d, there is little change in the surface potential which in turn implies little change in the energy-band edges at the semiconductor surface relative to the Fermi level. The Fermi level at the semiconductor surface is said to be more or less pinned by the high density of surface states.

When the metal makes contact with the silicon to form a Schottky diode, the energy-band diagram is as illustrated in Fig. 2.54(c). Note that a contact gap is shown to represent the region containing the double layers (i), (ii) and (iii) in the Bardeen model. The width of the contact gap is of atomic dimensions, at least for good contacts where there is no unintended interfacial material. The contact gap is assumed to be sufficiently thin to play no role in the transport of electrons between the metal and the silicon. Notice that, just as in Fig. 2.54(b), the potential difference across this infinitesimal contact gap is equal to $(\phi_m - \chi + \psi_s - E_g/2q + \psi_B)$. In Fig. 2.54(b), before the Schottky diode is formed, the surface potential ψ_s depends on the physical gap space. If there is weak Fermi-level pinning, ψ_s changes with the physical gap space. If there is strong Fermi-level pinning, ψ_s is relatively insensitive to the physical gap space. In Fig. 2.54(c), we have $\psi_s = -\psi_{bi}$, where ψ_{bi} is the built-in potential of the Schottky barrier diode. The electron energy barrier is $q\phi_{Bn}$. (The contact gap is transparent to electron transport between the metal and the silicon.)

Since the contact gap in a good metal-semiconductor contact is assumed to be transparent to electron transport between the metal and the semiconductor, we can omit it from the energy-band diagram completely for purposes of modeling device characteristics of a Schottky diode. (As discussed in the paragraph below, we should do this only with care.) The *simplified* energy-band diagram for a Schottky diode is as illustrated in Fig. 2.54(d).

It should be noted that Fig. 2.54(c) suggests that $\phi_{Bn} \neq \phi_m - \chi$ in cases of high surface-state density. This should be contrasted with the case of no surface states shown in Fig. 2.53, where we have $\phi_{Bn} = \phi_m - \chi$. If we had drawn the contact gap as a layer of zero thickness, the free electron level would have appeared discontinuous at the metal–silicon interface. Therefore, while in common practice Fig. 2.54(d) is shown and used by itself

for description of a Schottky diode, the correct and complete physical picture must include a contact gap of finite thickness in order to maintain continuity of free electron level across the metal–semiconductor system. A more detailed discussion on Schottky barriers and surface states with a mathematical model for the barrier height taking into consideration the work function difference, the density of surface states and the contact-gap thickness, can be found in Cowley and Sze (1965).

The energy-band diagrams in Fig. 2.54 are sketched based on the assumption of well defined metal and silicon surfaces and a well defined contact-gap region. In developing the model for Fermi-level pinning by surface states, Bardeen (1947) pointed out explicitly that if the contact between a metal and a semiconductor is very intimate, it may not be possible to distinguish between the double layers (i), (ii) and (iii). For an intimate contact, the metal will tend to broaden the surface levels. Furthermore, Heine (1965) showed that the wavefunction of an electron in a surface state does spread over some finite distance and that the concepts of band bending in a metal–semiconductor contact must not be taken too seriously over distances of the spread of localized electron wavefunctions. Therefore, for an intimate metal–semiconductor contact, the interface boundaries that define the contact-gap region are not as well defined as implied in Fig. 2.54(c). Fortunately, for purposes of describing and establishing the electrical characteristics of a Schottky diode, the details of the contact-gap region are not important because the region is transparent to electron transport.

For a metal–silicon contact represented by Figs 2.54(c) or 2.54(d), the built-in potential is given by

$$q\psi_{bi} = -q\psi_s(\text{contacted surface}) = \frac{[Q_d(\text{contacted surface})]^2}{2\varepsilon_{si}N_d}. \tag{2.230}$$

Once the built-in potential is known, the electron energy barrier can be determined from Eq. (2.228).

2.4.1.3 Measured Barrier Heights

That the Fermi level at the interface of a metal–semiconductor contact tends to be pinned by surface states has been well verified experimentally. The degree of pinning, however, varies with the semiconductor and often depends on the details of the process used for forming the metal–semiconductor contact. In practice, the measured barrier heights vary widely, even when the Schottky diodes are prepared in presumably identical processes. This is likely to be caused by some oxide layer or sublayer at the metal–semiconductor interface which alters the chemical bonding and structure of the interface. For metal contacts to many "clean" group IV and III-V semiconductor surfaces, the Fermi level pinning by surface states appears to be total, with the measured electron barrier heights practically independent of the metal used (Mead and Spitzer, 1964). For silicon and many other semiconductors, the measured barrier heights can be modeled by assuming partial pinning of the Fermi level by surface states (Cowley and Sze, 1965). In the case of silicon, annealing a metal–silicon system to form a metal-silicide–silicon contact can lead to barrier heights that are different from but more reproducible than the corresponding metal–silicon contact (Andrews and Phillips, 1975; Andrews, 1974).

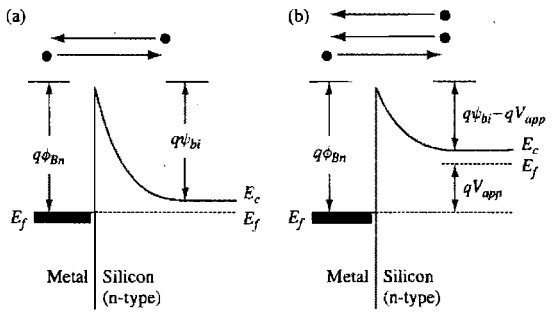


Figure 2.55. Schematic energy-band diagrams illustrating the flow of electrons in an n-type Schottky diode. (a) At thermal equilibrium, there is an equal and opposite flow of electrons. (b) At forward bias, there is a net flow of electrons from the silicon into the metal. For simplicity of illustration, barrier-lowering effect is not shown.

Tung (1992) suggested that there can be lateral inhomogeneity in the distribution of surface states or surface charge as well. Thus, a metal–semiconductor interface can be modeled as consisting of nanometer-sized local patches, with each patch having its own local electron energy barrier (Im *et al.*, 2001). The measured barrier height represents the averaged barrier height of the entire contact. Since there can be contact-to-contact variation in the lateral inhomogeneity, there can be inhomogeneity-induced variation in the measured barrier heights as well.

As can be inferred readily from Figs 2.53(b) and 2.54(d), the hole energy barrier $q\phi_{Bp}$ is related to the electron energy barrier $q\phi_{Bn}$ through

$$q\phi_{Bn} + q\phi_{Bp} = E_g, \tag{2.231}$$

where E_g is the energy gap of the semiconductor. We will focus our discussion on metal contacts to n-type silicon where the barrier height is $q\phi_{Bn}$. Metal contacts to p-type silicon where the barrier height is $q\phi_{Bp}$ will not be discussed explicitly.

2.4.1.4 Effect of Electric Field on Barrier Height

In Appendix 7, it is shown that the image-force effect causes the energy barrier for electron transport across a metal–silicon interface to be lowered by

$$q\Delta\phi \approx \sqrt{\frac{q^3\mathscr{E}_m}{4\pi\varepsilon_{si}}}, \tag{2.232}$$

where $\mathscr{E}_m$ is the maximum electric field in the silicon. The actual energy barrier for electron transport in a Schottky barrier diode is therefore $(q\phi_{Bn} - q\Delta\phi)$. The total band bending in the silicon is $q(\psi_{bi} - V_{app})$, where V_{app} is the forward-bias voltage across the Schottky diode (see Fig. 2.55), and Eq. (2.184) gives $\mathscr{E}_m = \text{sqrt}[2qN_d(\psi_{bi} - V_{app})/\varepsilon_{si}]$ for n-type silicon with a uniform doping concentration of N_d. That is, the effective energy barrier of a Schottky barrier is smaller than that suggested in Figs 2.53 and 2.54. A forward

bias ($V_{app} > 0$) across a diode reduces the electric field and hence increases the effective energy barrier, while a reverse bias ($V_{app} < 0$) increases the electric field and hence reduces the effective energy barrier.

2.4.2 Current Transport in a Schottky Barrier Diode

Consider an n-type silicon Schottky barrier diode. The energy-band diagrams illustrating the flow of electrons across the interface are shown schematically in Fig. 2.55. In modeling the transport of an electron across the interface, we need to consider the kinetic energy of the electron relative to the energy barrier for current flow across the interface, as depicted in the energy-band diagrams. For example, for an electron in the metal having an energy $E = E_f$, it sees an energy barrier of $q\phi_{Bn}$ (barrier-lowering effect is ignored for simplicity of discussion). For an electron having an energy $E = E_f + \Delta E$ it sees an energy barrier of $q\phi_{Bn} - \Delta E$. Similarly, for an electron in the quasineutral silicon region having an energy of $E = E_c$, it sees an energy barrier of $q(\psi_{bi} - V_{app})$. For an electron having an energy ΔE above the conduction-band edge, its barrier for transport across the interface is $q(\psi_{bi} - V_{app}) - \Delta E$. These energy barriers for current flow should not be confused with the energy barrier of the Schottky diode itself, which is $q\phi_{Bn}$.

At thermal equilibrium, there is no net electron flow in either direction in the diode, as indicated in Fig. 2.55(a). If a forward voltage V_{app} is applied to the diode, there will be a net electron flow from the n-silicon to the metal, but there are no holes (minority carriers) flowing into the n-silicon, as indicated in Fig. 2.55(b). Similarly, for a forward-biased p-type silicon Schottky barrier diode, there is a net flow of holes from the p-silicon into the metal, which is equivalent to a net flow of electrons from the metal into the valence band of the p-silicon. There are no excess electrons (minority carriers) injected from the metal into the conduction band of the p-silicon. That is, the current transport in a Schottky barrier diode is mainly due to majority carriers. This should be contrasted with a p–n diode where current transport is mainly due to minority carriers (electrons injected into the conduction band of the p-side and holes injected into the valence band of the n-side). The switching speed of a p–n diode is limited by the time it takes to discharge the minority carriers stored in the diode during forward bias (see Section 2.2.5.3). Therefore, without minority-carrier storage, a Schottky barrier diode is inherently faster than a p–n diode.

2.4.3 Current–Voltage Characteristics of a Schottky Barrier Diode

The processes by which electrons are transported from one side to the other in a Schottky barrier diode are illustrated in Fig. 2.56. Thermionic emission refers to the electrons having sufficient energy to surmount the effective (image-force effect included) energy barrier. Field emission refers to the tunneling of electrons from around the conduction-band edge. Thermionic-field emission describes the tunneling of electrons having energy above the conduction band but not enough energy to surmount the barrier. For a diode designed to function as an active device or a circuit component, the doping concentration is usually sufficiently light, and therefore the depletion layer thickness sufficiently large, so that thermionic emission is the dominant process for electron transport. Field emission and

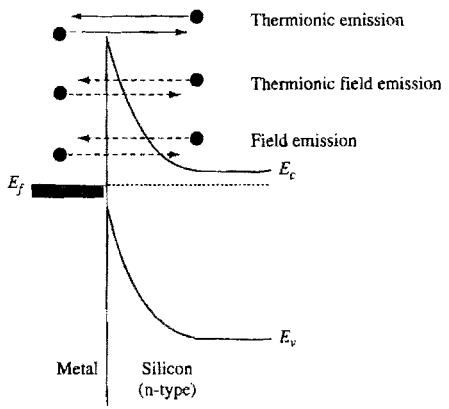


Figure 2.56. Schematic energy-band diagram of a Schottky barrier diode illustrating the principal transport processes.

thermionic-field emission are important transport processes in metal–semiconductor contacts where the doping levels are high (e.g., in ohmic contacts, which will be discussed later).

2.4.3.1 Thermionic Emission

In thermionic emission, the simplest theory is to treat the electrons as an ideal gas that follows Boltzmann statistics in energy distribution. Electron collision within the semiconductor depletion region is ignored, and only those electrons traveling in the direction of emission and having sufficient energy to surmount the barrier are emitted. In the case of a multi-valley semiconductor having anisotropic effective electron masses like silicon, we should consider the emission current from each valley and then sum the currents to obtain the total current.

The conduction band of silicon has six identical valleys located on the k_x-, k_y-, and k_z- axes. Each valley is an ellipsoid, with a longitudinal mass of $m_l = 0.92\,m_0$ and a transverse mass of $m_t = 0.19m_0$, where m_0 is the free-electron mass. The thermionic electron emission current density from an arbitrarily oriented silicon surface has been derived by Crowell (1965, 1969). The derivation is simplest for <100> silicon. Since <100> is the most commonly used silicon orientation, we shall only consider this orientation here. The reader interested in other orientations is referred to the paper by Crowell (1965).

Since electron collision within the depletion region is ignored, we can consider the thermionic emission current to be due to electrons originating from the quasineutral silicon region having sufficient energy to surmount the energy barrier for emission. As discussed in Section 2.1.1.2, the number of electronic states per unit volume having momenta between p_x and $p_x + dp_x$, between p_y and $p_y + dp_y$, and between p_z and $p_z + dp_z$ in one of the conduction-band valleys is

$$N(p_x, p_y, p_z)dp_x dp_y dp_z = \frac{2}{h^3} dp_x dp_y dp_z. \tag{2.233}$$

Note that the factor g denoting the number of equivalent minima in the conduction band is unity in Eq. (2.233) because only one valley is being considered. The kinetic energy of an electron in the quasineutral region is $E - E_c$, and the relationship between kinetic energy and momenta is given by Eq. (2.2). In terms of Boltzmann statistics, the probability that an electronic state at energy E is occupied by an electron is $\exp[-(E - E_f)/kT]$ [see Eq. (2.5)]. Therefore, Eq. (2.233) gives the number of electrons per unit volume having momenta between p_x and $p_x + dp_x$, between p_y and $p_y + dp_y$, and between p_z and $p_z + dp_z$ in one of the conduction-band valleys as

$$\begin{aligned} dn(p_x, p_y, p_z) &= \frac{2}{h^3} e^{-(E-E_f)/kT} dp_x dp_y dp_z \\ &= \frac{2}{h^3} e^{-(E_c-E_f)/kT} e^{-(p_x^2/2m_x kT + p_y^2/2m_y kT + p_z^2/2m_z kT)} dp_x dp_y dp_z. \end{aligned} \tag{2.234}$$

The number of electrons per unit volume having momenta between p_x and $p_x + dp_x$ is given by integrating Eq. (2.234) over all values of p_y and p_z. That is,

$$\begin{aligned} dn(p_x) &= \frac{2}{h^3} e^{-(E_c-E_f)/kT} e^{-p_x^2/2m_x kT} dp_x \int_{-\infty}^{\infty} e^{-p_y^2/2m_y kT} dp_y \int_{-\infty}^{\infty} e^{-p_z^2/2m_z kT} dp_z \\ &= \frac{4\sqrt{m_y m_z}}{h^3} \pi kT e^{-(E_c-E_f)/kT} e^{-p_x^2/2m_x kT} dp_x. \end{aligned} \tag{2.235}$$

- *Ignoring Barrier Lowering Effect.* In this case, when a voltage V_{app} is applied to a Schottky diode, the minimum energy an electron traveling perpendicularly to the emission surface must have in order to surmount the emission barrier is $q(\psi_{bi} - V_{app})$. For <100> silicon, the emission surface is perpendicular to the k_x-axis. The current density due to thermionic emission of electrons from a single conduction-band valley into the metal is

$$\begin{aligned} J_{1,s\to m}(V_{app}) &= q \int_{p_x = p_{x0}}^{\infty} v_x dn(p_x) \\ &= \frac{4\sqrt{m_y m_z}}{h^3} q\pi kT e^{-(E_c - E_f)/kT} \int_{p_{x0}}^{\infty} \frac{p_x}{m_x} e^{-p_x^2/2m_x kT} dp_x \\ &= \frac{4\pi q\sqrt{m_y m_z}}{h^3} k^2 T^2 e^{-(E_c-E_f)/kT} e^{-q(\psi_{bi} - V_{app})/kT}, \end{aligned} \tag{2.236}$$

where $v_x = p_x / m_x$ is the velocity of an electron traveling in the x-direction, and the lower integration limit p_{x0} is given by $p_{x0}^2/2m_x = q(\psi_{bi} - V_{app})$. The subscript 1 indicates that the current density is from only one conduction-band valley. Since $q\phi_{Bn} = q\psi_{bi} + E_c - E_f$, Eq. (2.236) can be rewritten as

$$\begin{aligned} J_{1,s\to m}(V_{app}) &= \frac{4\pi q\sqrt{m_y m_z}}{h^3} k^2 T^2 e^{-q\phi_{Bn}/kT} e^{qV_{app}/kT} \\ &= A \frac{\sqrt{m_y m_z}}{m_0} T^2 e^{-q\phi_{Bn}/kT} e^{qV_{app}/kT}, \end{aligned} \tag{2.237}$$

where the quantity

$$A \equiv \frac{4\pi q m_0 k^2}{h^3} \tag{2.238}$$

is the Richardson constant for free electrons. $A = 120$ A/cm^2/K^2.

To obtain the total thermionic electron emission current from the silicon, we need to sum the currents from all six valleys. For <100> silicon, the two valleys on the k_x-axis have $m_y = m_z = m_t$, and the four valleys on the k_y- and k_z-axes have $m_y = m_l$ and $m_z = m_t$. The total thermionic electron emission current density from silicon into metal is

$$\begin{aligned} J_{n-Si<100>,s\to m}(V_{app}) &= \sum_1^6 J_{i,n-Si<100>,s\to m}(V_{app}) \\ &= A\left(\frac{2m_t}{m_0} + \frac{4\sqrt{m_t m_l}}{m_0}\right) T^2 e^{-q\phi_{Bn}/kT} e^{qV_{app}/kT} \\ &= A^*_{n-Si<100>} T^2 e^{-q\phi_{Bn}/kT} e^{qV_{app}/kT}, \end{aligned} \tag{2.239}$$

where

$$A^*_{n-Si<100>} = A\left(\frac{2m_t}{m_0} + \frac{4\sqrt{m_t m_l}}{m_0}\right) = 2.05A \tag{2.240}$$

is the Richardson constant for n-type <100> silicon.

For silicon, the orientation dependence is relatively weak. For n-type <111> silicon, the Richardson's constant is $2.15A$ (Crowell, 1965). In theory, the measured Richardson's constant contains information about the effective mass tensor of the semiconductor. However, the simple thermionic emission model gives only a qualitative description of experimentally measured currents in a typical Schottky barrier diode. The measured Richardson's constant should not be used to infer information about the effective mass tensor (Crowell, 1969). In practice, the Richardson's constant is often treated as an adjustable parameter for fitting experimental data (Henisch, 1984).

At zero applied bias, the electron emission current from the metal into the silicon is equal in magnitude but opposite in direction to the electron emission current from the silicon into the metal. That is,

$$\begin{aligned} J_{n-Si<100>,m\to s}(V_{app} = 0) &= -J_{n-Si<100>,s\to m}(V_{app} = 0), \\ &= -A^*_{n-Si<100>} T^2 e^{-q\phi_{Bn}/kT}. \end{aligned} \tag{2.241}$$

When the barrier lower effect is ignored, the energy barrier for electron emission from metal into silicon is independent of V_{app}. Therefore, we expect the electron emission current from metal into silicon to be independent of V_{app} when barrier lowering effect is ignored. The total thermionic emission current density for an n-type <100> silicon Schottky barrier diode, when barrier lower effect is ignored, is therefore

$$\begin{aligned} J_{thermionic,n-Si<100>}(V_{app}) &= J_{n-Si<100>,s\to m}(V_{app}) + J_{n-Si<100>,m\to s}(V_{app}) \\ &= A^*_{n-Si<100>} T^2 e^{-q\phi_{Bn}/kT}\left(e^{qV_{app}/kT} - 1\right). \end{aligned} \tag{2.242}$$

For a forward-biased diode ($V_{app} > 0$), the current is dominated by the emission from the semiconductor into the metal. For a reverse-biased diode ($V_{app} < 0$), the current is dominated by the emission from the metal into the semiconductor. Equation (2.242) shows that, when barrier lowering effect is ignored, a Schottky barrier diode has I–V characteristics similar to those of a p–n diode [cf. Eq. (2.102)], with an $\exp(qV_{app}/kT)$ dependence on V_{app} in forward bias, and a saturation current that is independent of V_{app} in reverse bias.

- *Including Barrier Lowering Effect.* There is a subtle difference between a Schottky diode and a p–n diode when barrier lowering effect is included. When barrier lowering effect is included, the Schottky barrier $q\phi_{Bn}$ in Fig. 2.55 and in Eqs. (2.237) to (2.242) should be replaced by an effective Schottky barrier $q(\phi_{Bn} - \Delta\phi)$. The barrier-lowering term $q\,\Delta\phi$ depends on the applied voltage through the electric field $\mathscr{E}_m$ [see Eq. (2.232)]. As discussed in Section 2.4.1.4, a forward bias ($V_{app} > 0$) reduces $q\Delta\phi$ and hence increases the effective Schottky barrier, while a reverse bias ($V_{app} < 0$) increases $q\Delta\phi$ and hence reduces the effective Schottky barrier. Thus, replacing $q\phi_{Bn}$ by $q(\phi_{Bn} - \Delta\phi)$ in Eq. (2.242) suggests that the forward-bias current of a Schottky diode increases with V_{app} at a rate somewhat slower than $\exp(qV_{app}/kT)$. This should be compared with the forward-bias current of a p–n diode which is proportional to $\exp(qV_{app}/kT)$ [see Eq. (2.123)]. See also Fig. 2.23.

In the literature, more complex theories have been proposed for describing the transport process in Schottky diodes. There is a diffusion emission theory which includes the effect of electron collisions within the semiconductor depletion region. There is also a theory which combines the physics involved in the simple thermionic emission process and the diffusion emission process (Crowell and Sze, 1966a). All theories result in an equation similar to Eq. (2.242), with the difference only in the pre-exponential factor. The interested reader is referred to the literature for the details (Sze, 1981; Henisch, 1984). From a device point of view, the important I–V characteristics of a Schottky barrier diode are contained in the exponential factors in Eq. (2.242), namely in the $\exp(-q\phi_{Bn}/kT)$ dependence on $q\phi_{Bn}$ and the $[\exp(qV_{app}/kT) - 1]$ dependence on qV_{app}.

2.4.3.2 Field Emission and Thermionic-Field Emission

If the semiconductor is heavily doped, the depletion region thickness will be thin, and the electron transport can become dominated by a combination of field emission and thermionic-field emission. In this case, large currents can flow even at low applied biases. In general, when field emission and thermionic-field emission dominate the electron transport, a metal–semiconductor contact is no longer useful as a rectifying diode. As a result, we will not consider the general theory of field emission and thermionic-field emission any further. The interested reader is referred to the literature for details (Padovani and Stratton, 1966; Crowell and Rideout, 1969).

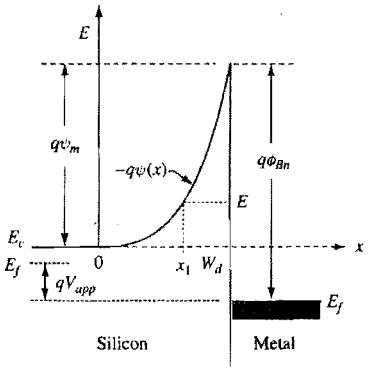


Figure 2.57. Schematic showing the energy bands appropriate for considering field emission in a metal–silicon contact. As illustrated, the Schottky diode is forward biased, as indicated by the Fermi level in the silicon being higher than that in the metal.

2.4.3.3 Schottky Barrier Diode as an Active Device

A rectifying Schottky barrier diode has *I–V* characteristics similar to those of a p–n diode, but a Schottky barrier diode is a much faster device than a p–n diode because it is a majority-carrier device. As a result, Schottky barrier diodes are often used as microwave diodes and as gates of microwave transistors where speed is important (see e.g., Irvin and Vanderwal, 1969). Also, Schottky barrier diodes are often added to bipolar circuits as voltage clamps to improve circuit speed.

2.4.4 Ohmic Contacts

Ohmic contacts are usually made with metal or metal silicide in contact with heavily doped semiconductor. The electron transport process in this case is dominated by field emission. Let us first consider the tunneling of a conduction-band electron from the quasineutral semiconductor region into the metal. The band bending near the metal–semiconductor contact is illustrated in Fig. 2.57. The total band bending is $q\psi_m = q(\psi_{bi} - V_{app})$ when a forward bias of V_{app} is applied. For a given V_{app}, let us assume the conduction-band starts to bend upward at $x = 0$, and the interface is located at $x = W_d$, where W_d is the depletion-layer thickness. Since we are considering an electron in the conduction band, it is convenient to use the conduction-band edge of the quasineutral silicon region $E_c(x < 0)$ as the energy reference, as indicated in Fig. 2.57. $\psi(x)$ is the electrostatic potential at location x relative to $E_c(x < 0)$, i.c., $-q\psi(0) = E_c(x < 0)$, and $-q\psi(x)$ is the potential energy of an electron at location x. The Poisson equation [Eq. (2.44)] can be integrated twice to give

$$\psi(x) = -\frac{qN_d x^2}{2\varepsilon_{si}} - \frac{E_c(x<0)}{q}, \tag{2.243}$$

where N_d is semiconductor doping concentration. From Eq. (2.188), we have

$$W_d = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} - V_{app})}{qN_d}}. \tag{2.244}$$

In the WKB approximation for tunneling through an energy barrier, the transmission coefficient through the energy barrier represented by $-q\psi(x)$ for an electron with energy E is

$$\begin{aligned} T(E) &= \exp\left[\frac{-4\pi}{h}\int_{x_1}^{W_d}\sqrt{2m^*}\sqrt{-q\psi(x) - E}dx\right] \\ &= \exp\left[\frac{-4\pi}{h}\int_{x_1}^{W_d}\sqrt{2m^*}\sqrt{\frac{q^2N_dx^2}{2\varepsilon_{si}} + E_c(x<0) - E}\,dx\right], \end{aligned} \tag{2.245}$$

where the lower integration limit x_1 is given by $-q\psi(x_1) = E$. In considering an ohmic contact, we are interested in the current due to electrons tunneling from the quasineutral region of the silicon through the potential barrier into the metal at small applied voltages. These electrons have only thermal energy ($kT \approx 26$ meV at room temperature) which is small compared to the maximum tunneling barrier height $q(\psi_{bi} - V_{app})$, which is approximately equal to $q\psi_{bi}$ at small V_{app}. Therefore, we can assume the tunneling electrons to have an energy $E \approx E_c(x < 0)$. For these electrons, the tunneling process starts at $x_1 = 0$ and the corresponding transmission coefficient is

$$T[E = E_c(x<0)] = \exp\left(\frac{-q(\psi_{bi} - V_{app})}{E_{00}}\right), \tag{2.246}$$

where

$$E_{00} \equiv \frac{q\hbar}{4\pi}\sqrt{\frac{N_d}{m^*\varepsilon_{si}}}. \tag{2.247}$$

For a heavily doped quasineutral silicon region, its Fermi level and conduction-band edge are about equal. That is, referring to Fig. 2.55(a), we have ψ_{bi}(heavily doped) $\approx \phi_{Bn}$. Therefore, for a heavily doped Schottky barrier, we have

$$T[(E = E_c(x<0)] = \exp\left(\frac{-q(\phi_{Bn} - V_{app})}{E_{00}}\right). \tag{2.248}$$

That is, for an ohmic contact, the current density varies as

$$J_{ohmic} \propto \exp\left(\frac{-q(\phi_{Bn} - V_{app})}{E_{00}}\right). \tag{2.249}$$

The specific contact resistance, or contact resistivity, ρ_c, is an important figure of merit for ohmic contacts:

$$\rho_c \equiv \left(\frac{\partial J}{\partial V_{app}}\right)^{-1}_{V_{app}=0}. \tag{2.250}$$

Using Eq. (2.249) for ohmic contacts, we have

$$\rho_c \propto \frac{E_{00}}{q}\exp(q\phi_{Bn}/E_{00}). \tag{2.251}$$

The behavior of ρ_c is dominated by the exponential factor. That is, to ensure a low contact resistance, a low-barrier metal should be used and the silicon should be as heavily doped as possible (to maximize E_{00}). For $N_d = 10^{20}$ cm^{-3} and $T = 300$ K, ρ_c is about 8×10^{-6} Ω-cm^2 for $q\phi_{Bn} = 0.6$ eV, and about 8×10^{-8} Ω-cm^2 for $q\phi_{Bn} = 0.4$ eV (Yu, 1970; Chang, *et al.*, 1971). Experimental determination of the resistance of a contact can be very involved for low-resistance contacts. The reader is referred to the literature for a discussion and comparison of the many contact-resistance measurement methods (Schroder, 1990).

2.5 High-Field Effects

In the presence of an electric field, carriers gain energy from the field as they drift along. These carriers in turn lose energy by emitting phonons. As the field increases, the average energy of the carriers increases. At sufficiently high fields, a number of physical phenomena which have important implications on the design and operation of VLSI devices can occur. In the case of high fields in silicon, these phenomena include impact ionization, or generation of electron–hole pairs, junction breakdown, band-to-band tunneling, and injection of hot carriers from locations near the silicon–oxide interface into the silicon dioxide region. In the case of high fields in silicon dioxide, the important phenomena include tunneling through the oxide layer and dielectric breakdown. The basic physics of these phenomena as they relate to VLSI devices is discussed in this section.

2.5.1 Impact Ionization and Avalanche Breakdown

Consider the depletion region of a p–n diode. At sufficiently high fields, an electron in the conduction band can gain enough energy to "lift" an electron from the valence band into the conduction band, thus generating one free electron in the conduction band and one free hole in the valence band. This process is known as *impact ionization*. Similarly, a hole in the valence band can gain enough energy to cause impact ionization. If the field is high enough, these secondary electrons and holes can themselves cause further impact ionization, thus beginning a process of carrier multiplication in the high-field region. The p–n diode breaks down when the multiplication process runs away or becomes an avalanche. The equations relating the rate of impact ionization to the condition for avalanche breakdown are derived below.

Consider a reverse-biased p–n diode with an electric field in its depletion region high enough to cause impact ionization. Let $x = 0$ and $x = W$ be the locations of the two boundaries of the depletion region. Suppose there is a hole current I_{p0} entering the depletion region at $x = 0$, as illustrated in Fig. 2.58. This hole current will generate electron–hole pairs. The secondary electrons and holes in turn cause further impact ionization as they traverse the depletion region. Thus, the hole current will increase

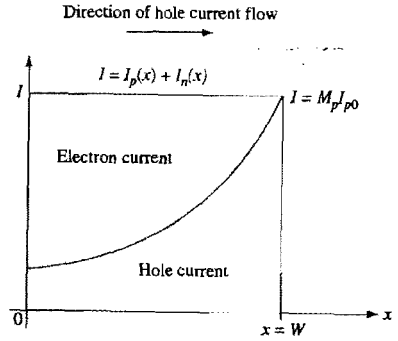


Figure 2.58. Schematic illustration of the steady-state current caused by hole-initiated impact ionization within the depletion region for a p–n diode.

with distance, reaching a value of $M_p I_{p0}$ at $x = W$, where M_p is the multiplication factor for holes. At steady state, the total current I is constant and independent of distance, i.e., $I = M_p I_{p0}$. Within the depletion region, the total current is the sum of the hole and electron currents (Moll, 1964), i.e.,

$$I = I_p(x) + I_n(x). \tag{2.252}$$

These current components are illustrated in Fig. 2.58. The field is such that holes move towards the right ($x = W$), and electrons move towards the left ($x = 0$).

Consider a differential distance between x and $x + dx$. There are $I_p(x)/q$ holes and $I_n(x)/q$ electrons crossing this differential distance per unit time. In crossing this differential distance, the holes cause $\alpha_p(x)I_p(x)dx/q$ electron–hole pairs to be generated, where α_p is the *hole-initiated rate of electron–hole pair generation per unit distance*. Similarly, the number of electron–hole pairs generated by the electrons is $\alpha_n(x)I_n(x)dx/q$, where α_n is the *electron-initiated rate of electron–hole pair generation per unit distance*. Thus the increase in the hole current as the electrons and holes cross the differential distance dx is

$$dI_p = \alpha_p I_p dx + \alpha_n I_n dx. \tag{2.253}$$

Equations (2.252) and (2.253) give

$$\frac{dI_p}{dx} = (\alpha_p - \alpha_n)I_p + \alpha_n I, \tag{2.254}$$

which, subject to the boundary condition $I_p(0) = I/M_p$, has a solution (Sze, 1981)

$$I_p(x) = I\left[\frac{1}{M_p} + \int_0^x \alpha_n \exp\left(-\int_0^{x'}(\alpha_p - \alpha_n)dx''\right)dx'\right]\exp\left(\int_0^x (\alpha_p - \alpha_n)dx'\right). \tag{2.255}$$

Since we are considering hole-initiated impact ionization, and there is no electron current entering the depletion region at $x = W$, the hole current at $x = W$ is simply equal to I. Therefore, Eq. (2.255) gives

$$\frac{1}{M_p} = \exp\left(-\int_0^W (\alpha_p - \alpha_n)dx\right) - \int_0^W \alpha_n \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx. \quad (2.256)$$

Similarly, for impact ionization initiated by electrons, the electron multiplication factor M_n is given by

$$\frac{1}{M_n} = \exp\left(-\int_0^W (\alpha_n - \alpha_p)dx\right) - \int_0^W \alpha_p \exp\left(-\int_x^W (\alpha_n - \alpha_p)dx'\right)dx. \quad (2.257)$$

Avalanche breakdown occurs when carrier multiplication by impact ionization runs away, i.e., when the multiplication factors become infinite. It is shown in Appendix 8 that ***the condition for avalanche breakdown is the same whether the breakdown process is initiated by electrons or by holes***. That is, when a p–n junction breaks down, it does not matter if the avalanche breakdown process is initiated by an electron or by a hole.

It should be noted that the avalanche multiplication of electrons and holes is a positive feedback process where both the electrons and holes generated by impact ionization take part in generating additional electrons and holes. As a result, it is possible to have avalanche breakdown, i.e., M_n and M_p being infinite, for finite values of W, α_n and α_p at some large reverse bias voltage. If the feedback process by either the secondary electrons or the secondary holes were absent, avalanche breakdown would not occur for finite values of α_n and α_p. To demonstrate this point, let us consider the case of impact ionization initiated by holes, where the multiplication factor M_p is given by Eq. (2.256). If there were no positive feedback by the secondary electrons, we would have $\alpha_n = 0$. In this case, Eq. (2.256) shows that M_p is finite (no breakdown) for any finite values of α_p.

If the high-field region where impact ionization occurs is sufficiently wide, the positive feedback process will eventually lead to avalanche breakdown. It is left to the reader to show that, for the special case of both α_n and α_p being constant, independent of distance or electric field, avalanche breakdown occurs when the width of the high-field region approaches a value of $[\ln(\alpha_p/\alpha_n)]/(\alpha_p - \alpha_n)$ (see Exercise 2.20).

In theory, Eqs. (2.256) and (2.257) can be used to calculate the multiplication factors, and hence the breakdown voltage. In practice, however, the ionization rates, as well as the junction doping profiles, are simply not known accurately enough for calculation of breakdown voltages to be made with sufficient accuracy for VLSI device design purposes. Breakdown voltages in modern VLSI devices are usually determined experimentally.

2.5.1.1 Empirical Impact Ionization Rates

The measured ionization rates are often expressed in the ***empirical*** form of

$$\alpha = A \exp(-b/\mathscr{E}) \quad (2.258)$$

where A and b are constants, and $\mathscr{E}$ is the electric field (Chynoweth, 1957). There is quite a bit of spread in the measured impact ionization rates reported in the literature. However, the most recent measurements give similar results (van Overstraeten and de Man, 1970; Grant, 1973). These results are shown in Table 2.2 and plotted in Fig. 2.59.

Two points are clear from Fig. 2.59. First, α_n is much larger than α_p, particularly at low electric fields. This is due to the effective mass of holes being much larger than that of

Table 2.2 Impact-Ionization Rates in Silicon

		α_n (cm^{-1})		α_p (cm^{-1})	
Data	Field Range (V/cm)	A_n (cm^{-1})	b_n (V/cm)	A_p (cm^{-1})	b_p (V/cm)
van Overstraeten and de Man	$1.75 \times 10^5 < \mathscr{E} < 4.0 \times 10^5$	7.03×10^5	1.231×10^6	1.582×10^6	2.036×10^6
	$4.0 \times 10^5 < \mathscr{E} < 6.0 \times 10^5$	7.03×10^5	1.231×10^6	6.71×10^5	1.693×10^6
Grant	$2.0 \times 10^5 < \mathscr{E} < 2.4 \times 10^5$	2.6×10^6	1.43×10^6	2.0×10^6	1.97×10^6
	$2.4 \times 10^5 < \mathscr{E} < 5.3 \times 10^5$	6.2×10^5	1.08×10^6	2.0×10^6	1.97×10^6
	$5.3 \times 10^5 < \mathscr{E}$	5.0×10^5	0.99×10^6	5.6×10^5	1.32×10^6

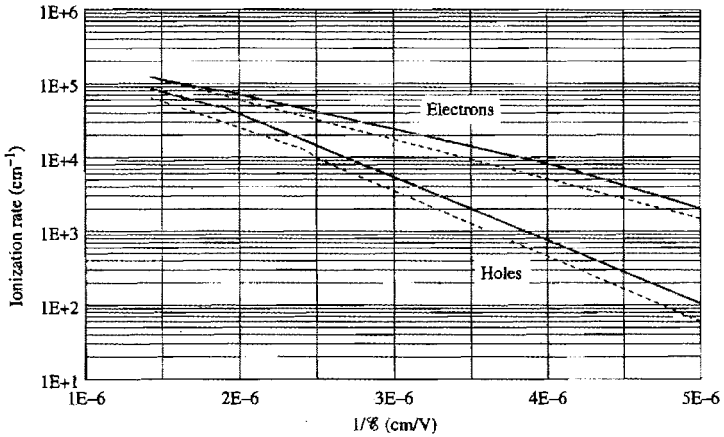


Figure 2.59. Impact-ionization rates in silicon. The solid curves are data of Grant (1973), and the dash curves are data of van Overstraeten and de Man (1970).

electrons. Second, the impact ionization rates increase very rapidly with electric field. For the depletion region of a p–n diode where the electric field is not constant, it is the small region surrounding the maximum-field point that contributes the most to the impact-ionization currents. Thus, to minimize impact ionization in a p–n diode, the maximum electric field should be minimized. As mentioned in Section 2.2.2.4, doping-profile grading, or using lightly doped regions or i-layers, can effectively reduce the peak electric field in a p–n junction.

Impact ionization rates decrease as temperature increases (Grant, 1973). This is due to the increased lattice scattering at higher temperatures. The data in Table 2.2 and Fig. 2.59 are for room temperature.

2.5.2 Band-to-Band Tunneling

When the electric field across a reverse-biased p–n junction approaches 10^6 V/cm, significant current flow can occur due to tunneling of electrons from the valence band

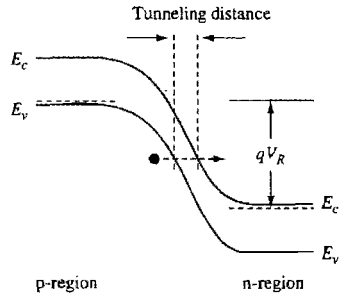


Figure 2.60. Schematic illustrating band-to-band tunneling in a p–n junction.

of the p-region into the conduction band of the n-region. This phenomenon is illustrated schematically in Fig. 2.60. In silicon this tunneling process usually involves the emission or absorption of phonons (Kane, 1961; Chynoweth *et al.*, 1960), and the tunneling current density is given by (Fair and Wivell, 1976)

$$J_{b-b} = \frac{\sqrt{2m^*}q^3\mathscr{E}V_R}{4\pi^3\hbar^2E_g^{1/2}}\exp\left(-\frac{4\sqrt{2m^*}E_g^{3/2}}{3q\mathscr{E}\hbar}\right) \tag{2.259}$$

where $\mathscr{E}$ is the electric field, E_g is the energy bandgap, and V_R is the reverse bias voltage across the junction. An upper-bound estimate of the peak electric field can be made by assuming a one-sided junction. In this case, the analyses in Section 2.2.2 give the upper-bound for the electric field as

$$\mathscr{E}_{\max} = \sqrt{\frac{2qN_a(V_R + \psi_{bi})}{\varepsilon_{si}}}, \tag{2.260}$$

where N_a is the doping concentration of the lightly doped side (assumed p-type) of the diode and ψ_{bi} is the built-in potential of the diode. With these approximations, the band-to-band tunneling current density is about 1 A/cm^2 for $N_a = 5 \times 10^{18}$ cm^{-3} and $V_R = 1$ V (Taur *et al.*, 1995a). More recently, Solomon *et al.* (2004) showed that band-to-band tunneling current can be modeled using the concept of an effective tunneling distance. In this model, the tunneling current is assumed to be proportional to $\exp(-w_T/\lambda_T)$, where w_T is the tunneling distance, illustrated schematically in Fig. 2.60, and λ_T is an effective tunneling decay length. The reader is referred to Solomon's paper for the details.

As will be discussed in Chapters 4 and 7, in scaling down the dimensions of a transistor, the doping concentrations increase and the junction doping profiles become more abrupt, and hence band-to-band tunneling effect increases. Once the leakage current due to band-to-band tunneling is appreciable, it increases very rapidly with electric field or reduction of the tunneling distance. For modern VLSI devices, band-to-band tunneling is becoming one of the most important leakage-current components, particularly for applications such as DRAM and battery-operated systems where leakage currents must be kept extremely low.

2.5.3 Tunneling into and through Silicon Dioxide

Consider an MOS capacitor discussed in Section 2.3. For simplicity, the gate electrode is assumed to be heavily doped n-type polysilicon. When biased at the flatband condition, the energy-band diagram is as shown in Fig. 2.61(a), where $q\phi_{ox}$ denotes the Si–SiO_2 interface energy barrier for electrons which, as indicated in Fig. 2.29, is about 3.1 eV. When a large positive bias is applied to the gate electrode, electrons in the strongly inverted surface can tunnel through the oxide layer and hence give rise to a gate current. Similarly, if a large negative voltage is applied to the gate electrode, electrons from the n^+ polysilicon can tunnel through the oxide layer, and again give rise to a gate current.

2.5.3.1 Fowler–Nordheim Tunneling

Fowler–Nordheim tunneling occurs when electrons tunnel into the conduction band of the oxide layer and then drift through the oxide layer. Figure 2.61(b) illustrates Fowler–Nordheim tunneling of electrons from the silicon surface inversion layer. The

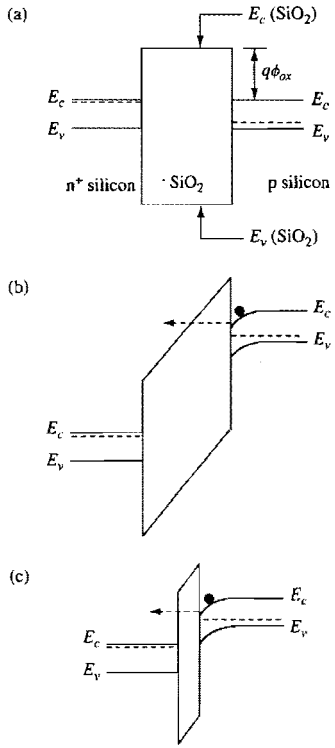


Figure 2.61. Tunneling effects in an MOS capacitor structure: (a) energy-band diagram of an n-type polysilicon-gate MOS structure at flat band; (b) Fowler–Nordheim tunneling; (c) direct tunneling.

complete theory of Fowler–Nordheim tunneling is rather complicated (Good Jr. and Müller, 1956). For the simple case, where the effects of finite temperature and image-force barrier lowering (which is discussed in Appendix 7) are ignored, the tunneling current density is given by (Lenzlinger and Snow, 1969)

$$J_{FN} = \frac{q^2 \mathscr{E}_{ox}^2}{16\pi^2 \hbar \phi_{ox}} \exp\left(-\frac{4(2qm^*)^{1/2} \phi_{ox}^{3/2}}{3\hbar \mathscr{E}_{ox}} \right), \tag{2.261}$$

where $\mathscr{E}_{ox}$ is the electric field in the oxide. Equation (2.261) shows that Fowler–Nordheim tunneling current is characterized by a straight line in a plot of $\log\left(J/\mathscr{E}_{ox}^2\right)$ versus $1/\mathscr{E}_{ox}$.

As discussed later in Section 2.5.3.3, electrons tunneling into an oxide layer can be trapped in an oxide layer. If the tunneling current is measured at a constant voltage, then the trapped electrons in turn can cause the observed tunneling current to decrease with time. Depending on the thickness of the oxide layer and its formation process, this decrease in tunneling current can go on for some time before it reaches a more-or-less steady state. The tunneling currents reported in the classic paper by Lenzlinger and Snow were taken after the samples were first subjected to a current density of about 10^{-10} A/cm^2 for two hours, during which time the tunneling currents decreased by about one order of magnitude from their initial values (Lenzlinger and Snow, 1969). At an oxide field of 8 MV/cm, the measured steady-state Fowler–Nordheim tunneling current density is about 5×10^{-7} A/cm^2. The initial tunneling current is about ten times larger. For normal device operation, Fowler–Nordheim tunneling current is negligible.

The characteristics of the tunneling currents represented by Eqs. (2.259) and (2.261) are determined primarily by their exponential factors. It should be noted that the exponents of the two equations are basically the same. The Fowler–Nordheim tunneling is through a triangular barrier of height $q\phi_{ox}$, slope $q\mathscr{E}_{ox}$ and tunneling distance $\phi_{ox}/\mathscr{E}_{ox}$. It is left as an exercise (Exercise 2.21) for the reader to show that the band-to-band tunneling exponent in Eq. (2.259) can be derived from the WKB approximation, i.e., Eq. (2.245), for tunneling through a triangular barrier of height E_g, slope $q\mathscr{E}$ and tunneling distance $E_g / q\mathscr{E}$.

2.5.3.2 Direct Tunneling

If the oxide layer is very thin, say 4 nm or less, then, instead of tunneling into the conduction band of the SiO_2 layer, electrons from the inverted silicon surface can tunnel directly through the forbidden energy gap of the SiO_2 layer. This is illustrated in Fig. 2.61(c). The theory of *direct tunneling* is even more complicated than that of Fowler–Nordheim tunneling, and there is no simple dependence of the tunneling current density on voltage or electric field (Chang *et al.*, 1967; Scheugraf *et al.*, 1992). Direct-tunneling current can be very large for thin oxide layers. Figure 2.62 is a plot of the measured and simulated thin-oxide tunneling current versus voltage in polysilicon-gate MOSFETs (Lo *et al.*, 1997). For the gate-voltage range shown in Fig. 2.62, the current is primarily a direct-tunneling current. Direct-tunneling current is important in MOSFETs of very small dimensions, where the gate oxide layers can approach 1 nm in thickness.

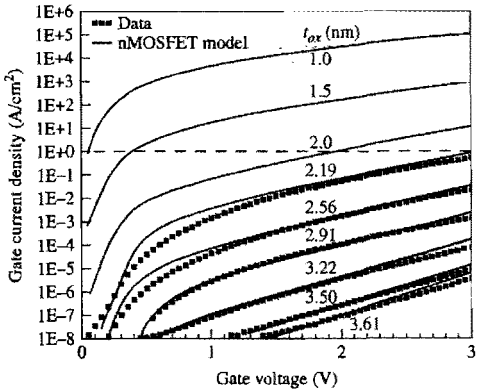


Figure 2.62. Measured (dots) and simulated (solid lines) tunneling currents in thin-oxide polysilicon-gate MOS devices. The dashed line indicates a tunneling-current level of 1 A/cm^2. (After Lo *et al.*, 1997.)

2.5.3.3 Defect Generation Caused By Tunneling Current

The tunneling of electrons into and through a silicon dioxide layer can cause "defects" to be generated within the oxide layer and/or at the oxide–silicon interface. These defects can take the form of electron traps, hole traps, trapped electrons, trapped holes, or interface states (DiStefano and Shatzkes, 1974; Harari, 1978; Chen *et al.*, 1986; DiMaria *et al.*, 1993). These defects govern the time dependent behavior of the tunneling current and play an important role in the wear-out and eventual breakdown of the oxide layer. In this subsection, we briefly discuss how these defects can influence the tunneling process. The reader is referred to the vast literature on the subject for more details (DiMaria and Cartier, 1995, and Suehle, 2002, and the references therein).

- *Tunneling into an electron trap.* As electrons tunnel into an oxide layer, some of the electrons can get trapped. The trapped electrons modify the oxide field such that the field near the cathode (the electrode that acts as an electron source) is decreased, while the field near the anode (the electrode that acts as an electron sink) is increased. This is illustrated in Fig. 2.63. The reduced field near the cathode, in turn, causes the tunneling current to decrease. In a constant-voltage tunneling current measurement, electron trapping is what causes the current to decrease with time. In a ramped-voltage (voltage increasing with time at a constant rate) tunneling current measurement, electron trapping often leads to a hysteresis in the voltage–current plot.
- *Hole generation, injection, and trapping.* As an electron travels in the conduction band of an oxide layer, it gains energy from the oxide field. If the voltage drop across the oxide layer is larger than the bandgap energy of silicon dioxide, which, as indicated in Fig. 2.28, is about 9 eV, the electron can gain sufficient energy to cause impact ionization in the oxide. The holes generated by impact ionization can be trapped in the oxide. Holes can be injected indirectly into the oxide layer during electron tunneling as well. A tunneling electron arriving at the anode can cause impact

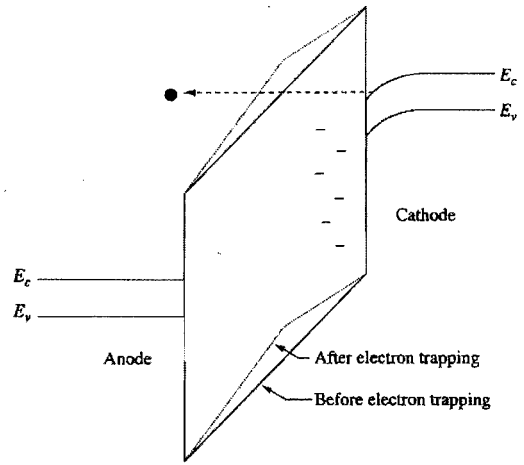


Figure 2.63. Schematic illustrating the trapping of tunneling electrons. As electrons are trapped, the oxide field near the cathode (electron source) is decreased, while the oxide field near the anode (electron sink) is increased.

ionization in the anode near the oxide–anode interface. Depending on the energy of the tunneling electron, the hole thus generated can be from deep down in the valence band of the anode, and thus can be "hot." A hot hole in the anode near the anode–oxide interface can be injected into the oxide layer. This process is illustrated in Fig. 2.64. The injected hole can be trapped in the oxide layer as it travels towards the cathode. The trapped holes in the oxide layer cause the oxide field near the cathode to increase, which in turn causes the tunneling current to increase. This is illustrated in Fig. 2.65. Thus the trapping of holes provides a positive feedback to the electron tunneling process. In a constant-voltage tunneling current measurement, hole trapping is the primary reason the current increases with time.

- *Trap and interface-state generation.* Traps can also be generated in the silicon dioxide layer and at the oxide–silicon interface by the electron current (Harari, 1978; DiMaria, 1987; Hsu and Ning, 1991). In addition to increasing electron and hole trapping, these traps can enhance the tunneling current by assisting in the tunneling process, as discussed in the two subsections below.

2.5.3.4 Bulk-Trap-Assisted Tunneling

Instead of tunneling directly through an oxide layer, an electron can first tunnel from the cathode into an electron trap in the oxide and then tunnel from the trap to the anode. Thus, traps in the oxide layer can act as stepping stones for the tunneling electrons. This is illustrated in Fig. 2.66. The enhanced tunneling current in turn can increase the generation of traps in the oxide. Thus, trap-assisted tunneling plays an important role in the degradation of an oxide under voltage stress because of the positive feedback between trap-generation and trap-assisted tunneling (DiMaria and Cartier, 1995).

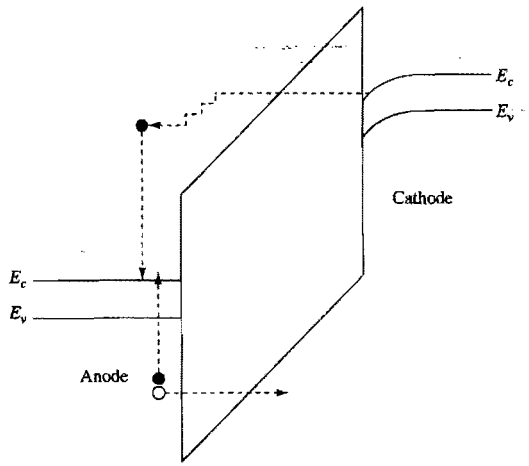


Figure 2.64. Schematic illustrating the generation of an electron–hole pair in the anode by a tunneling electron. The hole thus generated can then be injected (by tunneling in this example) into the oxide layer.

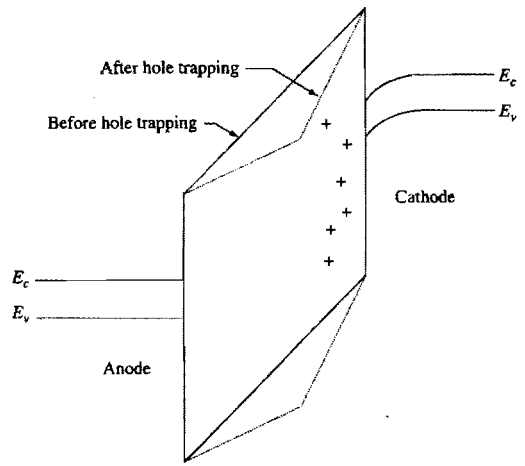


Figure 2.65. Schematic showing the trapping of holes in the oxide layer. The trapped holes enhance the electric field near the cathode, and decrease the electric field near the anode.

2.5.3.5 Interface-Trap-Assisted Tunneling at Low Voltages

Interface traps can also assist in the tunneling process. This is illustrated in Fig. 2.67. Interface states exist at both the substrate silicon–oxide interface and the gate silicon–oxide interface. In general, states above the Fermi level are empty of electrons and states below the Fermi level are filled with electrons. In Fig. 2.67(a), the gate electrode is biased

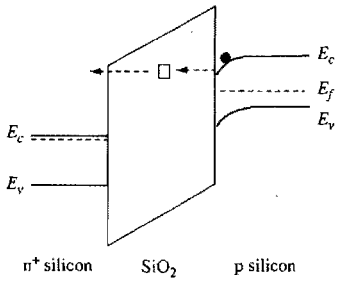


Figure 2.66. Schematic illustrating bulk-trap-assisted tunneling in an MOS capacitor structure.

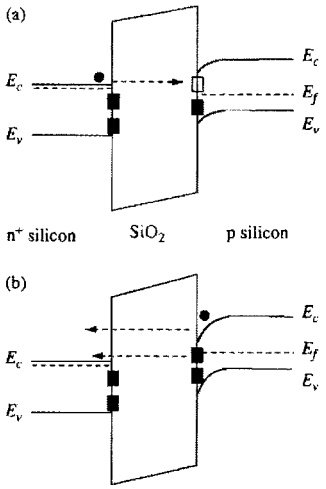


Figure 2.67. Schematics illustrating interface-trap-assisted tunneling of electrons in a silicon-gate MOS structure. (a) The gate electrode has a small negative bias. The Fermi level in the p-silicon lies somewhat below that in the n^+ silicon gate. (b) The gate electrode has a small positive bias. The Fermi level in the p silicon lies somewhat above that of the n^+ silicon gate.

slightly negatively (towards flatband condition and surface accumulation of the p-silicon). As the surface of the p silicon is driven towards flat band and surface accumulation, more and more states at the interface become empty of electrons. An electron in the conduction band of the gate electrode can tunnel into an empty state at the p-silicon surface. Also, not shown in the figure, an electron in an occupied surface state of the gate electrode can also tunnel into an empty surface state of the p silicon. These interface-trap-assisted tunneling processes are in addition to the normal tunneling of electrons from the conduction band of the gate electrode into the conduction band of the p substrate.

In Fig. 2.67(b), the gate is biased somewhat positively (towards surface inversion of the p silicon). As the surface of the p silicon is driven towards inversion, more and more

states at the surface become filled with electrons. Until the p silicon surface is inverted, there are very few electrons available to tunnel from the conduction band of the p silicon to the conduction band of the gate electrode. However, electrons from the filled surface states of the p silicon can tunnel into the conduction band of the gate electrode. Since interface-trap-assisted tunneling is a direct tunneling process, it is important only for thin oxides. Also, as can be inferred from Fig. 2.67, interface-trap-assisted tunneling is effective only when the p silicon surface potential lies between inversion and weak accumulation. That is, interface-trap-assisted tunneling is important only at low voltages. ***Interface-trap-assisted tunneling can enhance the low-voltage tunneling current in thin oxides whether the gate electrode is biased positively or negatively.*** Interface-trap-assisted tunneling in modern CMOS devices is a widely studied subject. The reader is referred to the literature for more details (see e.g. Crupi *et al.*, 2002, and the references therein).

2.5.4 Injection of Hot Carriers from Silicon into Silicon Dioxide

If a region of sufficiently high electric field is located near the Si–SiO_2 interface, some electrons or holes in the region can gain enough energy from the electric field to surmount the interface barrier and enter the SiO_2 layer. In general, injection from Si into SiO_2 is much more likely for hot electrons than for hot holes because (a) electrons can gain energy from the electric field much more readily than holes due to their smaller effective mass, and (b) the Si–SiO_2 interface energy barrier is larger for holes (≈ 4.6 eV) than for electrons (≈ 3.1 eV), as indicated in Fig. 2.28. The process of hot-electron and hot-hole injection from silicon into silicon dioxide is much too complex to model quantitatively. Thus far, quantitative agreement has been shown only for the special case of hot electrons traveling from the silicon substrate perpendicularly towards the Si–SiO_2 interface, and only with Monte-Carlo models that take into account the correct band structures, all the relevant scattering processes, and nonlocal transport properties (Fischetti *et al.*, 1995). Here we discuss a simple model for the injection of hot electrons and hot holes from Si into SiO_2. The same model can be modified readily to describe the injection of hot holes from Si into SiO_2.

2.5.4.1 Energy Barrier for Hot Electron Injection

The energy barrier $q\phi_{ox}$ shown in Fig. 2.61(a) is the difference in energy between the conduction band of SiO_2 and the conduction band of Si. In Appendix 7, it is shown that the image-force effect causes the barrier for injection of a hot electron from Si into SiO_2 to be lowered by an amount equal to

$$q\Delta\phi = \sqrt{\frac{q^3 \mathscr{E}_{ox}}{4\pi\varepsilon_{ox}}}. \tag{2.262}$$

The actual energy barrier for hot electron emission is therefore $(q\phi_{ox} - q\Delta\phi)$.

For $\mathscr{E}_{ox} = 1 \times 10^6$ V/cm, $q\Delta\phi = 0.19$ eV. Thus, for practical oxide fields of 10^6 V/cm or larger, image-force barrier lowering is not negligible compared to the interface energy

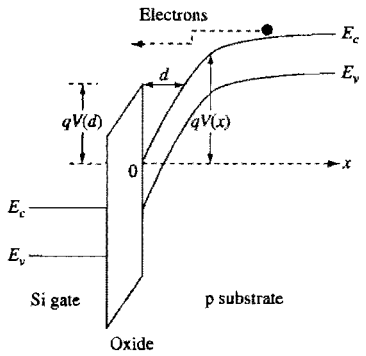


Figure 2.68. Schematic illustrating hot electrons traveling perpendicularly towards the Si–SiO_2 interface and being injected into the SiO_2 layer.

barrier of 3.1 eV. Image-force barrier lowering is included in the more accurate theories of Fowler–Nordheim tunneling (Lenzlinger and Snow, 1969) and direct tunneling (Chang *et al.*, 1967). However, in the literature there are also publications questioning the validity of the concept of image potential at the interface between a semiconductor and an insulator (see e.g. Fischetti *et al.*, 1995, and the references therein).

2.5.4.2 The Lucky Electron Model

The simple one-dimensional injection process is illustrated in Fig. 2.68. The simplest model for describing the injection process is the *lucky electron* model proposed by Shockley (1961). It is an empirical model, but it describes the measured data surprisingly well (Ning *et al.*, 1977a). In this model, the probability that a hot electron at a distance d from the Si–SiO_2 interface will be emitted into the SiO_2 layer is expressed as

$$P(d) = A\exp(-d/\lambda), \tag{2.263}$$

where λ is an effective mean free path for energy loss by hot electrons in silicon, and A is a fitting constant to the experimental data. The relation between the parameter d, the effective hot-electron emission energy barrier, and the electron potential energy, is illustrated in Fig. 2.68. The parameter d can be obtained as follows. Referring to Fig. 2.68, $qV(x)$ is the potential energy of an electron at x. An electron at $x = d$ has just enough potential energy to overcome the effective energy barrier for emission if it can travel from $x = d$ to the interface at $x = 0$ without encountering any energy-losing collision. That is, $qV(d)$ is equal to the effective energy barrier for emission. The injection of hot holes from Si into SiO_2 can be described by a similar *lucky hole* model (Selmi *et al.*, 1993).

It was determined empirically that the effective energy barrier for electron emission can be written as

$$qV(d) = q\phi_{ox} - q\Delta\phi - \alpha\mathscr{E}_{ox}^{2/3}, \tag{2.264}$$

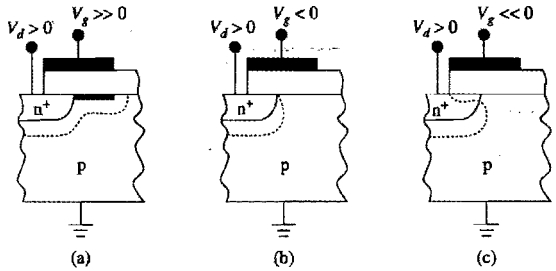


Figure 2.69. Schematics illustrating a gated n^+–p diode when the surface is (a) inverted and (b) accumulated, and (c) when the surface of the n^+ region is depleted or inverted. The dashed lines indicate the boundary of the depletion region.

where the first two terms are the Si–SiO_2 interface energy barrier and the image-force barrier lowering discussed in the previous subsection, and the third term is introduced to account for the fact that hot electrons without enough energy to surmount the image-force-lowered energy barrier can still tunnel into the oxide layer. It was found that setting $\alpha = 1 \times 10^{-5}\text{e}(\text{cm}^2\text{-V})^{1/3}$ and $A = 2.9$ fits a wide range of measured emission probabilities (Ning *et al.*, 1977a).

The temperature dependence of the hot-electron injection process is contained in the temperature dependence of the effective mean free path (Crowell and Sze, 1966b)

$$\lambda(T) = \lambda_0 \tanh(E_R/2kT), \tag{2.265}$$

where $E_R = 63$ meV is the optical-phonon energy, and λ_0 is the low-temperature limit of λ. It was found empirically that $\lambda_0 = 10.8$ nm (Ning *et al.*, 1977a). The mean free path associated with hot-hole injection has a comparable value (Selmi *et al.*, 1993).

2.5.5 High-Field Effects In Gated Diodes

Thus far, the effects of high fields have been considered for p–n diodes and MOS capacitors separately. In a gated diode structure, both the location of the peak-field region and the magnitude of the peak field vary with gate voltage. Let us consider a gated n^+–p diode. As discussed in Section 2.3.5, when the gate is biased to invert the silicon surface, the inverted surface region has about the same potential as the n^+ region, and the gated diode behaves like a large-area n^+–p diode. If the p-region is uniformly doped, the depletion-layer width is about the same below the n^+ silicon region as below the surface inversion region, hence the electric field is rather uniformly distributed and is about the same as in a simple p–n diode. This is illustrated schematically in Fig. 2.69(a).

When the gate is biased somewhat negatively to accumulate the silicon surface, the silicon surface under the gate has about the same potential as the p-type substrate. Owing to the presence of the accumulated holes at the surface, the surface behaves like a p-region more heavily doped than the substrate, causing the depletion layer at the surface

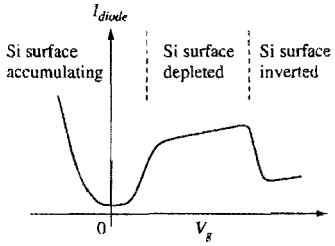


Figure 2.70. Schematic illustrating an n^+–p gated-diode leakage current as a function of gate voltage.

to become narrower than elsewhere. This is illustrated schematically in Fig. 2.69(b). The narrowing of the depletion layer at or near the intersection of the p–n junction and the Si–SiO_2 interface causes *field crowding*, or an increase in the local electric field.

When the negative gate bias is large enough, the n^+ region under the gate can become depleted, and even inverted. This is illustrated in Fig. 2.69(c). In this case, the gate and the n^+ region behave like an MOS capacitor with a heavily doped n-type "substrate." There is more field crowding, and the peak field increases.

As the electric field in and around the gated p–n junction is increased by the gate voltage, all the high-field effects, such as avalanche multiplication and band-to-band tunneling, can increase very dramatically. Thus the leakage current of a reverse-biased gated diode can increase dramatically when the gate voltage begins to cause field crowding in and around the junction region. This is illustrated in Fig. 2.70 which shows the expected gated-diode leakage current as a function of gate voltage. In Fig. 2.70, aside from the increase in current due to field crowding at negative gate voltage, the diode current is simply the sum of the leakage current from the depletion region of a diode and the leakage current from the exposed surface states (Grove and Fitzgerald, 1966). When the Si surface is in accumulation (at $V_g \approx 0$), the leakage current is from the depletion region of the bulk diode alone. When the Si surface is inverted (at large V_g), the leakage current is higher because of the additional leakage current coming from the depletion region of the diode formed by the surface inversion layer and the substrate. When the Si surface is depleted (at intermediate values of V_g), the leakage current is the largest because of the addition leakage current coming from the exposed surface states of the depleted silicon surface. When field crowding occurs in the drain junction of a MOSFET, the increased junction leakage current is called *gate-induced drain leakage*, or GIDL (Chan *et al.*, 1987a; Noble *et al.*, 1989). ***GIDL is an important leakage-current component that must be minimized in modern CMOS devices.***

It should be noted that for the gated n^+–p diode considered here, when the gate is biased to accumulate the silicon surface, the oxide field favors the injection of hot holes from the silicon substrate into the silicon dioxide layer (Verwey, 1972). Similarly, for a gated p^+–n diode, the oxide field favors the injection of hot electrons when the gate is biased to accumulate the silicon surface. Thus injection of majority carriers, instead of minority carriers, from the silicon substrate into the silicon dioxide layer takes place when significant gate-voltage-induced avalanche multiplication occurs in a gated diode.

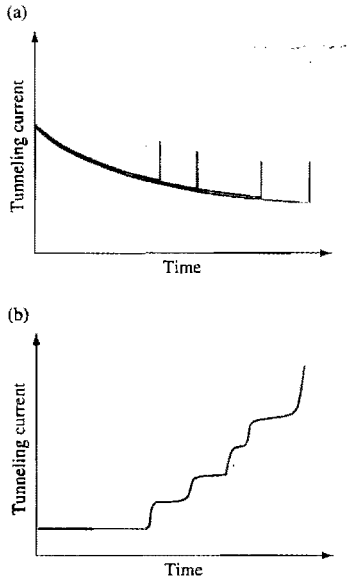


Figure 2.71. Schematics illustrating the typical time dependence of the tunneling current at constant voltage in (a) a thick oxide layer and (b) a thin oxide layer. A sudden jump in tunneling current signals a dielectric breakdown event.

2.5.6 Dielectric Breakdown

As discussed in Section 2.5.3, significant electron tunneling can take place when a large electric field is applied across an oxide layer. Figure 2.71 illustrates schematically the typical time dependence of the tunneling current when a constant voltage is applied across an oxide layer. A sudden jump in the tunneling current indicates that the oxide sample has suffered a *dielectric breakdown* event. For oxides thicker than about 10 nm, the tunneling current typically decreases gradually with time until the oxide breaks down. For these thick oxides, the voltages used to measure tunneling current are usually so large that, unless special care is taken to limit the current, a breakdown event usually leads to the oxide being physically damaged (Shatzkes *et al.*, 1974). There is a distribution in the measured oxide breakdown time (Harari, 1978). This is illustrated in Fig. 2.71(a). For oxides thinner than about 5 nm, the voltages used to measure tunneling current are usually sufficiently small so that not every breakdown event leads to catastrophic breakdown. For most samples, *successive breakdown* events are observed before final or catastrophic breakdown (Suñé *et al.*, 2004). This is illustrated in Fig. 2.71(b). An oxide layer ceases to be a good electrical insulator after it suffers final or catastrophic breakdown.

In thin oxides, the increase in tunneling current from the first breakdown event to final breakdown can occur quite gradually. When the thin gate oxide of a modern MOS transistor in a circuit starts showing signs of breaking down, often the gate tunneling

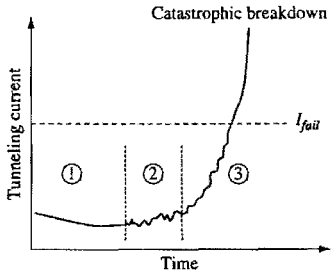


Figure 2.72. Schematic illustrating the evolution of the tunneling current in a thin oxide MOS device at constant applied voltage. I_{fail} indicates the tunneling current at which the device fails to function properly in a circuit. The three stages marked 1, 2, and 3 are discussed in the text.

current can grow to a sufficiently large value to cause the circuit to fail long before the gate oxide layer suffers final breakdown (Kaczer *et al.*, 2000; Linder *et al.*, 2001). Figure 2.72 is a schematic illustrating the time dependence of the tunneling current in a typical thin-oxide MOS device at constant voltage stress. There are roughly three stages in the evolution of the tunneling current. In the initial stage (stage 1 in Fig. 2.72), the current is relatively featureless, typically first decreasing, due to electron trapping, and then rising, due to hole trapping and trap-assisted tunneling, as a function of time. As the current continues to rise, it becomes noisy (stage 2). Finally, the current rises much more rapidly (stage 3) for some time before final breakdown. In Fig. 2.72, I_{fail} denotes the tunneling-current level at which a circuit using the MOS transistor fails to function properly. In the literature, stage 1 is often referred to as the *defect generation* stage or *stress-induced leakage current* stage (DiMaria, 1987; Stathis and DiMaria, 1998); stage 2 as the *soft breakdown* stage (Depas *et al.*, 1996), and stage 3 as the *successive breakdown* or *progressive breakdown* stage (Linder *et al.*, 2002; Okada, 1997; Suñé and Wu, 2002).

2.5.6.1 Breakdown Field

In the literature, the quality of an oxide film is often measured in terms of the electric field at which dielectric breakdown, usually the first breakdown event, occurs. "Good-quality" thick (>100 nm) SiO_2 films typically break down at fields greater than 10 MV/cm, while "good-quality" thin (<10 nm) SiO_2 films usually show larger breakdown fields, often in excess of 15 MV/cm. In bipolar transistors, because there are normally no thin oxide components, the electric fields across the oxide layers are usually so small that dielectric breakdown is not a concern. In CMOS devices, the maximum oxide field varies widely, depending on the application. For devices used in logic and memory circuits, the maximum oxide field is typically in the 3–6 MV/cm range in normal operation, and can reach as high as 5–9 MV/cm in special operations (such as during a device burn-in process). For devices used in electrically programmable nonvolatile memory applications, where normal operation involves tunneling through a thin dielectric layer, the maximum electric field across the thin dielectric layer is in excess of 10 MV/cm. Dielectric breakdown is a real concern in CMOS devices.

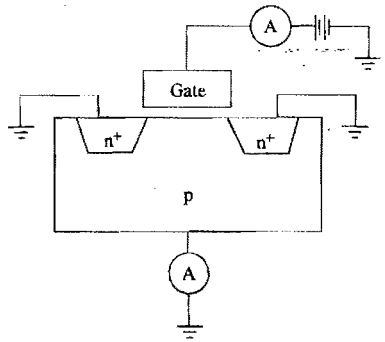


Figure 2.73. Schematic illustrating the bias configuration of an n-channel MOSFET for measuring the charge to breakdown and its hole-charge component.

2.5.6.2 Time to Breakdown and Charge to Breakdown

The breakdown characteristics of an oxide film are often described in terms of its *time to breakdown*, which measures the time needed for the film to reach breakdown, or its *charge to breakdown*, which measures the integrated total tunneling charge leading up to breakdown. In product design, we want to ensure that the gate current of a CMOS device will not grow to the point of causing circuit failure before the product end of life. Therefore, in product design, we want to know the time to breakdown. However, it appears easier to develop physical models relating charge to breakdown to the physical mechanisms involved in the dielectric breakdown process, such as hole current, trapping, trap generation, and interface state generation (Schuegraf and Hu, 1994; DiMaria and Stathis, 1997; Stathis and DiMaria, 1998), than to develop physical models relating time to breakdown to these physical mechanisms. Most publications on the physics of dielectric breakdown discuss the breakdown process in terms of charge to breakdown instead of time to breakdown. Therefore, we will not discuss time to breakdown any further here. The reader is referred to the literature for discussions on time to breakdown and the breakdown statistics in the time domain (see e.g. Suñé *et al.*, 2004, and the references therein).

As discussed in Section 2.5.3.3, a tunneling electron current can generate a hole current. Thus, the charge to breakdown, Q_{BD}, is the sum of the charges due to electrons and holes. If an MOS capacitor structure is used to measure Q_{BD}, then, owing to the two-terminal nature of the device, only the total charge can be measured. However, if an n-channel MOSFET or an n^+–p gated-diode structure is used to measure Q_{BD}, then both the total charge and the hole-charge component can be determined. For the case of an n-channel MOSFET, the bias configuration for such measurements is illustrated in Fig. 2.73. The basic concept of this *charge separation* method is that electron current is measured at the n-type terminal and hole current is measured at the p-type terminal. Integration of the gate current gives the total charge, and integration of the substrate current gives the charge due to the holes. It is shown that, charge for charge, hot holes are much more effective than tunnel electrons in generating defects that lead to oxide breakdown (Li *et al.*, 1999).

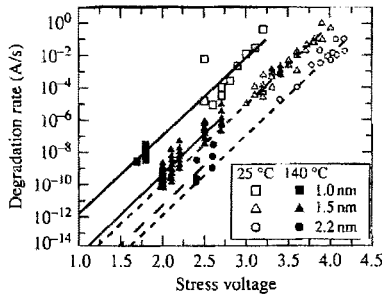


Figure 2.74. Measured rate of increase of breakdown current for typical thin oxides. (After Linder *et al.*, 2002.)

2.5.6.3 Progressive Breakdown and Successive Breakdown

Referring to Fig. 2.72, the evolution of the tunneling current can be described as a process of positive feedback between defect generation and trap-assisted tunneling. When a stress voltage is applied across an oxide layer, at first there are few defects in the oxide and the tunneling current is relatively low, decreasing with time as electrons are trapped. As trapped holes start to accumulate in the oxide, the current will start increasing. The tunneling current generates defects which in turn assist in the tunneling process (DiMaria, 1987; Stathis and DiMaria, 1998). At some point, soft breakdown starts when the defects in the oxide become dense enough such that an electron can tunnel relatively easily from one defect center to another across the oxide layer. This trap-assisted tunneling current tends to be noisy (Depas *et al.*, 1996). As the defect density continues to grow, hard breakdown (a breakdown event or a series of breakdown events) starts when there is a connected path of overlapping defects all the way across the oxide layer (Degraeve *et al.*, 1995; Stathis, 1999). This connected path of defects acts as a low-resistance conduction path for the electrons. Once hard breakdown starts, the electron current is completely dominated by the flow along this low-resistance path and the magnitude of the current is more-or-less independent of the device area. The electron current causes the diameter of the path of connected defects to grow, which in turn causes the current to grow. The current does not grow smoothly, but in a staircase manner. Each time the tunneling current jumps, it represents a breakdown event. Eventually, catastrophic breakdown of the oxide layer occurs. Thus, a thin oxide can go through many successive breakdown events before it breaks down catastrophically (Suñé and Wu, 2002). The oxide degradation rate (the rate at which the average breakdown current increases with time) is a strong function of the stress voltage (Linder *et al.*, 2002; Lombardo *et al.*, 2003). Figure 2.74 is a plot of typical measured degradation rates for thin oxides. It suggests that even after hard breakdown has commenced, the tunneling current in a thin oxide at a low voltage can take a long time to grow to a value sufficiently large to cause circuit failure.

In the literature, most charge to breakdown measurements are made by integrating the tunneling current until the current shows a sudden jump in magnitude, or until the first breakdown event. For a given oxide film, the charge to breakdown Q_{BD} is often plotted

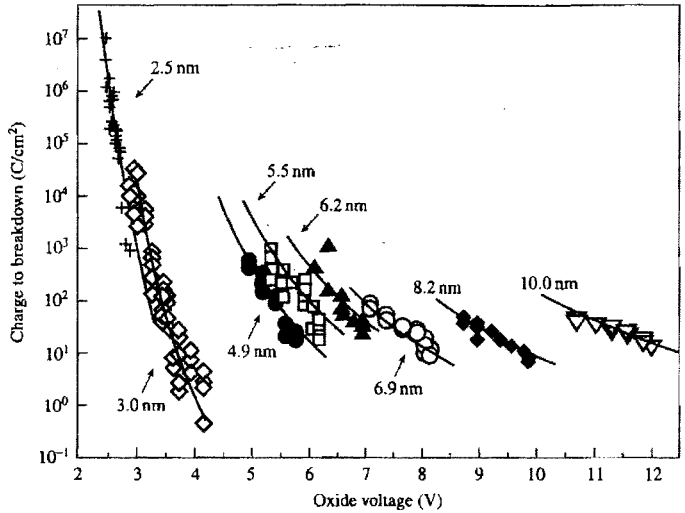


Figure 2.75. Typical plot of charge to breakdown versus oxide voltage for several oxide thickness values. (After Schuegraf and Hu, 1994.)

as a function of oxide voltage. Figure 2.75 is a typical plot for oxide thickness in the 2.5–10 nm range (Schuegraf and Hu, 1994). It shows that, for these relatively thick oxides, Q_{BD} decreases with increasing oxide voltage. It has also been shown that Q_{BD} is about the same for n-channel and p-channel MOSFETs (DiMaria and Stathis, 1997). Since these published Q_{BD} values do not take into account the progressive nature of the breakdown process, they project a lower allowed voltage for an oxide than is justified from a device reliability point of view. The progressive breakdown and the successive breakdown models, which take into account the oxide degradation rate after hard breakdown has commenced (stage 3 in Fig. 2.72), project a larger but more accurate allowed voltage (Linder *et al.*, 2002). The reader is referred to the literature on thin oxide reliability for more details (Degraeve *et al.*, 1998; Suehle, 2002; Suñé *et al.*, 2004).

As illustrated in Fig. 2.72, the gate leakage current of a MOSFET in a circuit has to reach a certain critical level, indicated by I_{fail}, before the circuit ceases to function properly. Thus, in circuit applications, what designers really need to know is the time to critical current instead of time to first breakdown or charge to breakdown.

Exercises

2.1 Show that the values of the Fermi–Dirac distribution function, Eq. (2.4), at a pair of energies symmetric about the Fermi energy E_f, are complementary, i.e., show that $f_D(E_f - \Delta E) + f_D(E_f + \Delta E) = 1$, independent of temperature.

2.2 For a given donor level E_d and concentration N_d of an n-type silicon, solve the Fermi energy E_f from the charge neutrality condition, Eq. (2.19) (neglecting the hole term). Show that $E_c - E_f$ approaches the complete ionization value, Eq. (2.20), under the condition of shallow donor level with low to moderate concentration. What happens if the condition is not satisfied?

2.3 Use the density of states $N(E)$ derived in Section 2.1.1.2 to evaluate the average kinetic energy of electrons in the conduction band:

$$\langle \text{K.E.} \rangle = \frac{\int_{E_c}^{\infty} (E - E_c) N(E) f_D(E) dE}{\int_{E_c}^{\infty} N(E) f_D(E) dE}.$$

(a) For a nondegenerate semiconductor in which can $f_D(E)$ be approximated by the Maxwell–Boltzmann distribution, Eq. (2.5), show that $\langle \text{K.E.} \rangle = \frac{3}{2}kT$.

(b) For a degenerate semiconductor at 0 K, show that $\langle \text{K.E.} \rangle = \frac{3}{5}(E_f - E_c)$.

2.4 The 3-D Gauss's law is obtained after a volume integration of the 3-D Poisson's equation and takes the form

$$\oint_S \mathscr{E} \cdot dS = \frac{Q}{\varepsilon_{si}},$$

where the left-hand side is an integral of the normal electric field over a closed surface S, and Q is the net charge enclosed within S. Use it to derive the electric field at a distance r from a point charge Q (Coulomb's law). What is the electric potential in this case?

2.5 (a) Use Gauss's law to show that the electric field at a point above a uniformly charged sheet of charge density Q_s per unit area is $Q_s/2\varepsilon$, where ε is the permittivity of the medium.

(b) For two oppositely charged parallel plates with surface charge densities Q_s and $-Q_s$, show that the electric field is uniform and equals Q_s/ε in the region between the two plates and is zero in the regions outside the two plates.

2.6 The total depletion charge and inversion charge densities of a p-type MOS capacitor can be expressed as

$$Q_d = -qN_a \int_0^{W_d} (1 - e^{-q\psi/kT}) dx = -qN_a \int_0^{\psi_s} \frac{1 - e^{-q\psi/kT}}{\mathscr{E}} d\psi,$$

and

$$Q_i = -q\frac{n_i^2}{N_a} \int_0^{W_d} (e^{q\psi/kT} - 1) dx = -q\frac{n_i^2}{N_a} \int_0^{\psi_s} \frac{e^{q\psi/kT} - 1}{\mathscr{E}} d\psi,$$

using Eqs. (2.177) and (2.178). Here $\mathscr{E} = -d\psi/dx$ is given by Eq. (2.181).

(a) Write down the expressions for the small-signal depletion capacitance, $C_d = -dQ_d/d\psi_s$ and the small-signal inversion capacitance (low frequency), $C_i = -dQ_i/d\psi_s$, in silicon as represented in the equivalent circuit in Fig. 2.37.

(b) Show that $C_d + C_i = C_{si}$, where $C_{si} = -dQ_s/d\psi_s$ is evaluated using Eq. (2.182).

(c) Show that $C_d \approx C_i$ at the condition of strong inversion, $\psi_s = 2\psi_B$. (This allows one to use a split C–V measurement to determine the gate voltage where $\psi_s = 2\psi_B$).

(d) From the behavior of C_d beyond strong inversion, explain the "screening" of depletion charge (incremental) by the inversion layer.

2.7 Near the surface of an MOS capacitor biased well into strong inversion, only the $\exp(q\psi/kT)$ term in the square-root expression of Eq. (2.191) needs to be kept (classical model). Solve $\psi(x)$ under the boundary condition $\psi(0) = \psi_s$. Express the inversion electron concentration $n(x)$ in terms of the surface concentration $n(0)$ given by Eq. (2.193).

2.8 Solve the gate voltage equation (2.195) for $\psi_s(V_g)$ under the depletion condition in which $Q_s(\psi_s) = Q_d(\psi_s)$ given by Eq. (2.189). Show that for incremental changes, $\Delta\psi_s = \Delta V_g/(1 + C_d/C_{ox})$, where C_d is the depletion charge capacitance given by Eq. (2.201).

2.9 When the gate voltage greatly exceeds the threshold for strong inversion, a first-order solution of $\psi_s(V_g)$ can be obtained from the coupled equations (2.195) and (2.182), by keeping only the inversion charge term. Show that

$$\psi_s \approx 2\psi_B + \frac{2kT}{q} \ln\left(\frac{C_{ox}(V_g - V_{fb} - 2\psi_B)}{\sqrt{2\varepsilon_{si} kT N_a}}\right)$$

under these circumstances. Estimate how much higher ψ_s can be over $2\psi_B$ by substituting some typical values in the logarithmic expression.

2.10 In the split C–V measurement in Fig. 2.37(b), show that the n^+ channel part of the small-signal gate capacitance is

$$\frac{dQ_i}{dV_g} = \frac{C_{ox}C_i}{C_{ox} + C_i + C_d}.$$

Sketch the functional behavior of dQ_i/dV_g versus V_g, and from it describe the behavior of Q_i versus V_g.

2.11 The multiplication factors for holes and for electrons are given by Eqs. (2.256) and (2.257), respectively. For the special case of constant α_p and α_n, show that $M_p \to \infty$ occurs when the depletion-layer width approaches the value of $W = \ln(\alpha_n/\alpha_p)/(\alpha_n - \alpha_p)$. Also show that the condition for $M_n \to \infty$ gives the same result for W.

2.12 Prove the following mathematical identities:

$$\int_0^W f(x) \exp\left(-\int_0^x f(x')\,dx'\right) dx = 1 - \exp\left(-\int_0^W f(x)\,dx\right)$$

and

$$\int_0^W f(x) \exp\left(-\int_x^W f(x')\,dx'\right) dx = 1 - \exp\left(-\int_0^W f(x)\,dx\right).$$

These identities are used in Appendix 8 to show that the condition for hole-initiated avalanche breakdown, namely $1/M_p \to 0$, is the same as that for electron-initiated avalanche breakdown, namely $1/M_n \to 0$.

2.13 The depletion-layer capacitance per unit area C_d of a uniformly doped abrupt p–n diode and its dependence on doping concentration and applied voltage are given in Eqs. (2.80), (2.81), and (2.83). Sketch $1/C_d^2$ as a function of the applied reverse-bias voltage V_{app}. Show how this plot can be used to determine N_a and N_d.

2.14 The depletion-layer capacitance of a one-sided p–n diode is often used to determine the doping profile of the lightly doped side. Consider an n^+–p diode, with a nonuniform p-side doping concentration of $N_a(x)$. If $Q_d(V)$ is the depletion-layer charge per unit area at bias voltage V, the capacitance per unit area at bias voltage V is $C = dQ_d/dV$. In terms of the depletion-layer width W, we have $C(V) = \varepsilon_{si}/W$, where W is a function of V. (For simplicity, we have dropped the subscripts in C, W, and V here.) Show that the doping concentration at the depletion-layer edge is given by

$$N_a(W) = \frac{2}{q\varepsilon_{si} d(1/C^2)/dV}.$$

2.15 The charge distribution of a p–i–n diode is shown schematically in Fig 2.17. The i-layer thickness is d. The depletion-layer capacitance is given by Eq. (2.96), namely $C_d = \varepsilon_{si}/W_d$, where $W_d = x_n + x_p$ is the total depletion-layer width. Derive this result from $C_d = dQ_d/dV$.

2.16 Consider a p–n diode. Assume the junction is located at $x = 0$, with the n-region to the left (i.e., $x < 0$) and the p-region to the right (i.e., $x > 0$) of the junction. The distribution of the excess electrons is given by Eq. (2.119), and the electron current density entering the p-region is given by Eq. (2.120). Derive the equation for the distribution of the excess holes in the n-region and the equation for the hole current density entering the n-region.

2.17 The minimum leakage current of a reverse-biased diode is determined by its saturation current components. The saturation currents depend on the dopant concentrations of the diode, as well as on the widths of the quasineutral p- and n-regions. They also depend on whether or not heavy-doping effect is included. This exercise is designed to show the magnitude of these effects.

(a) Consider an n^+–p diode, with an emitter doping concentration of 10^{20} cm^{-3} and a base doping concentration of 10^{17}cm^{-3}. Assume both the emitter and the base to be wide compared with their corresponding minority-carrier diffusion lengths. Ignore heavy-doping effect and calculate the electron and hole saturation current densities [see Eq. (2.129)].

(b) In most modern MOSFET and bipolar devices, the n^+–p diodes have the n^+–region width small compared with its hole diffusion length. If we assume the quasineutral n^+ region to have a width of 0.1 μm, again ignoring heavy-doping effect, estimate the hole saturation current density [see Eq. (2.133)].

(c) It is discussed in Section 6.1.2 and shown in Fig 6.3 that the effect of heavy doping should be included once the doping concentration is larger than about 10^{17} cm^{-3}. Heavy-doping effect is usually included simply by replacing the intrinsic-carrier concentration n_i by an effective intrinsic-carrier concentration n_{ie}, where n_i and n_{ie} are related by

$$n_{ie}^2 = n_i^2 \exp(\Delta E_g/kT).$$

The empirical parameter ΔE_g is called the apparent bandgap narrowing due to heavy-doping effect, and its values are plotted in Fig. 6.3. Repeat (b) including the effects of heavy doping.

2.18 Consider an n^+–p diode, with the n^+ emitter side being wide compared with its hole diffusion length and the p base side being narrow compared with its electron diffusion length. The diffusion capacitance due to electron storage in the base is C_{Dn}, and that due to hole storage in the emitter is C_{Dp}. Assume the emitter to have a doping concentration of 10^{-20} cm^{-3} and the base to have a width of 100 nm and a doping concentration of 10^{17} cm^{-3}.

(a) If heavy-doping effect is ignored, the capacitance ratio is [see Eq. (2.168)]

$$\frac{C_{Dn}}{C_{Dp}} = \frac{2}{3}\frac{N_E}{N_B}\frac{W_B}{L_{pE}} \quad \text{(heavy-doping effect ignored)}.$$

Evaluate this ratio for the n^+–p diode.

(b) When heavy-doping effect cannot be ignored, it is usually included simply by replacing the intrinsic-carrier concentration n_i by an effective intrinsic-carrier concentration n_{ie} [see part (c) of Exercise 2.17]. Show that when heavy-doping effect is included, the capacitance ratio becomes

$$\frac{C_{Dn}}{C_{Dp}} = \frac{2}{3}\left(\frac{n_{ieB}^2}{n_{ieE}^2}\right)\left(\frac{N_E}{N_B}\right)\frac{W_B}{L_{pE}} \quad \text{(heavy-doping effect included)},$$

where the subscript B denotes quantities in the base and the subscript E denotes quantities in the emitter. Evaluate this ratio for the n^+–p diode.

(This exercise demonstrates that heavy-doping effects cannot be ignored in any quantitative modeling of the switching speed of a diode.)

2.19 As electrons are injected from silicon into silicon dioxide, some of these electrons become trapped in the oxide. Let N_T be the electron trap density, n_T be the density of trapped electrons, and j_G/q be the injected electron particle current density. The rate equation governing $n_T(t)$ is

$$\frac{dn_T}{dt} = \frac{j_G}{q}\sigma(N_T - n_T),$$

where σ is the capture cross section of the traps. If the initial condition for n_T is $n_T(t=0)=0$, show that the time dependence of the trapped electron density is given by

$$n_T(t) = N_T\{1 - \exp[-\sigma N_{inj}(t)]\},$$

where

$$N_{inj}(t) \equiv \int_0^t \frac{j_G(t')}{q} dt'$$

is the number of injected electrons per unit area. Assume $N_T = 5 \times 10^{12}$ cm^{-3} and $\sigma = 1 \times 10^{-13}$ cm^2, sketch a log-log plot of n_T as a function of N_{inj}. (The capture cross section is often measured by fitting to such a plot.)

2.20 The avalanche multiplication factors M_p and M_n are given by Eqs. (2.256) and (2.257). Assume α_p and α_n are constant, independent of distance or electric field. Show that avalanche breakdown occurs when the width W of the high-field region (the region where impact ionization occurs) approaches $[\ln(\alpha_p/\alpha_n)]/(\alpha_p - \alpha_n)$.

2.21 Show that the band-to-band tunneling exponent in Eq. (2.259) can be derived from the WKB approximation, i.e. Eq. (2.245), for tunneling through a triangular barrier of height E_g, slope $q\mathscr{E}$ and tunneling distance $E_g/q\mathscr{E}$.

2.22 Assume silicon, room temperature, complete ionization. An abrupt p–n junction with $N_a = N_d = 10^{17}$ cm^{-3} is reversed biased at 2.0 V.

(a) Draw the band diagram. Label the Fermi levels and indicate where the voltage appears.

(b) What is the total depletion layer width?

(c) What is the maximum field in the junction?

2.23 For an abrupt n^+–p diode in Si, the n^+ doping is 10^{20} cm^{-3}, the p-type doping is 3×10^{16} cm^{-3}. Assume room temperature and complete ionization.

(a) Draw the band diagram at zero bias. Indicate $x=0$ as the boundary where the doping changes from n^+ to p. Also indicate where the Fermi level is with respect to the midgap.

(b) Write the equation and calculate the built-in potential.

(c) Write the equation and calculate the depletion width.

(d) Will the built-in potential increase or decrease if the temperature goes up and why?

2.24 Sketch the C–V curve (high frequency) of an MOS capacitor consisting of n^+ poly gate on n-type Si doped to $N_d = 10^{16}$ cm^{-3}. Calculate and show the flatband voltage on the C–V. Draw the band diagram for $V_g = 0$. Given $t_{ox} = 10$ nm, what is V_{ox} (potential across oxide) at the onset of inversion ($\psi_s = 2\psi_B$)? Ignore quantum and poly depletion effects.

2.25 Consider an MOS device with 20 nm thick gate oxide and uniform p-type substrate doping of 10^{17} cm^{-3}. The gate work function is that of n^+ Si.

(a) What is the flatband voltage? What is the threshold voltage for strong inversion?

(b) Sketch the high frequency C–V curve. Label where the flatband voltage and threshold voltage are.

(c) Calculate the maximum and the minimum capacitance (per area) values.

2.26 If the device in Exercise 2.25 is biased at zero gate voltage, determine the surface potential and the electron and hole densities at the surface.

3 MOSFET Devices

The metal–oxide–semiconductor field-effect transistor (MOSFET) is the building block of VLSI circuits in microprocessors and dynamic memories. Because the current in a MOSFET is transported predominantly by carriers of one polarity only (e.g., electrons in an n-channel device), the MOSFET is usually referred to as a unipolar or majority-carrier device. Throughout this chapter, n-channel MOSFETs are used as an example to illustrate device operation and derive drain-current equations. The results can easily be extended to p-channel MOSFETs by exchanging the dopant types and reversing the voltage polarities.

The basic structure of a MOSFET is shown in Fig. 3.1. It is a four-terminal device with the terminals designated as *gate* (subscript *g*), *source* (subscript *s*), *drain* (subscript *d*), and *substrate* or *body* (subscript *b*). An n-channel MOSFET, or nMOSFET, consists of a p-type silicon substrate into which two n^+ regions, the source and the drain, are formed (e.g., by ion implantation). The gate electrode is usually made of metal or heavily doped polysilicon and is separated from the substrate by a thin silicon dioxide film, the *gate oxide*. The gate oxide is usually formed by thermal oxidation of silicon. In VLSI circuits, a MOSFET is surrounded by a thick oxide called the *field oxide* to isolate it from the adjacent devices. The surface region under the gate oxide between the source and drain is called the *channel* region and is critical for current conduction in a MOSFET. The basic operation of a MOSFET device can be easily understood from the MOS capacitor discussed in Section 2.3. When there is no voltage applied to the gate or when the gate voltage is zero, the p-type silicon surface is either in accumulation or in depletion and there is no current flow between the source and drain. The MOSFET device acts like two back-to-back p–n junction diodes with only low-level leakage currents present. When a sufficiently large positive voltage is applied to the gate, the silicon surface is inverted to n-type, which forms a conducting channel between the n^+ source and drain. If there is a voltage difference between them, an electron current will flow from the source to the drain. A MOSFET device therefore operates like a switch ideally suited for digital circuits. ***Since the gate electrode is electrically insulated from the substrate, there is effectively no dc gate current, and the channel is capacitively coupled to the gate via the electric field in the oxide*** (hence the name *field-effect transistor*).

3.1 Long-Channel MOSFETs

This section describes the basic characteristics of a long-channel MOSFET, which will serve as the foundation for understanding the more important but more complex

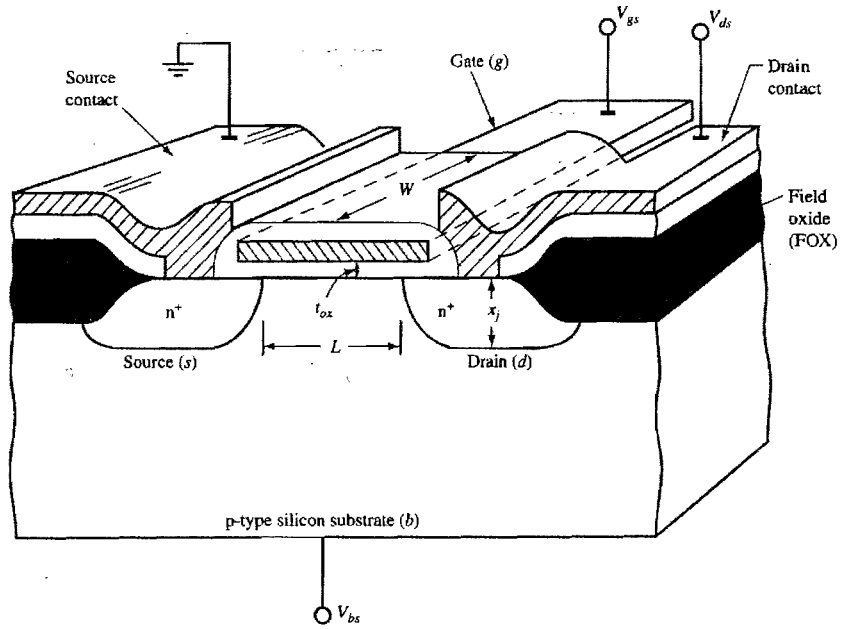


Figure 3.1. Three-dimensional view of basic MOSFET device structure. (After Arora, 1993.)

short-channel MOSFETs in Section 3.2. First, a general MOSFET current model based on the *gradual channel approximation* (GCA) is formulated in Section 3.1.1. The GCA is valid for most regions of MOSFET operation except beyond the pinch-off or saturation point. A *charge-sheet model* is then introduced to obtain implicit equations for the source–drain current. Regional approximations are applied in Section 3.1.2 to derive explicit *I–V* expressions for the linear and parabolic regions. Current characteristics in the subthreshold region are discussed in Section 3.1.3. Section 3.1.4 addresses the threshold-voltage dependence on substrate bias and temperature. Section 3.1.5 presents an empirical model for electron and hole mobilities in a MOSFET channel. Lastly, intrinsic MOSFET capacitances and inversion-layer capacitance effects (neglected in the regional approximation) are covered in Section 3.1.6.

3.1.1 Drain-Current Model

In this subsection, we formulate a general drain-current model for a long-channel MOSFET. The model will then be simplified using *charge-sheet approximation*, leading to an analytical expression for the source–drain current. Figure 3.2 shows the schematic cross section of an n-channel MOSFET in which the source is the n^+ region on the left, and the drain is the n^+ region on the right. A thin oxide film separates the gate from the channel region between the source and drain. We choose an *x–y* coordinate system

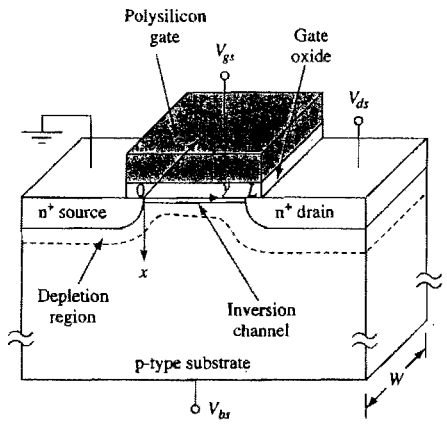


Figure 3.2. A schematic MOSFET cross section, showing the axes of coordinates and the bias voltages at the four terminals for the drain-current model.

consistent with Section 2.3 on MOS capacitors, namely, the x-axis is perpendicular to the gate electrode and is pointing into the p-type substrate with $x=0$ at the silicon surface. The y-axis is parallel to the channel or the current flow direction, with $y=0$ at the source and $y=L$ at the drain. L is called the *channel length* and is a key parameter in a MOSFET device. The MOSFET is assumed to be uniform along the z-axis over a distance called the *channel width*, W, determined by the boundaries of the thick field oxide.

Conventionally, the source voltage is defined as the ground potential. The drain voltage is V_{ds}, the gate voltage is V_{gs}, and the p-type substrate is biased at V_{bs}. Initially, we assume $V_{bs}=0$, i.e., the substrate contact is grounded to the source potential. Later on, we will discuss the effect of substrate bias on MOSFET characteristics. The p-type substrate is assumed to be uniformly doped with an acceptor concentration N_a.

3.1.1.1 Gradual-Channel Approximation

One of the key assumptions in any 1-D MOSFET model is the ***gradual channel approximation (GCA), which assumes that the variation of the electric field in the y-direction (along the channel) is much less than the corresponding variation in the x-direction (perpendicular to the channel)*** (Pao and Sah, 1966). This allows us to reduce the 2-D Poisson equation to 1-D slices (x-component only) as in Eq. (2.175). The GCA is valid for most of the channel regions except beyond the *pinch-off* point, which will be discussed later.

As defined in Section 2.3.2, $\psi(x, y)$ is the band bending, or intrinsic potential, at (x, y) with respect to the intrinsic potential of the bulk substrate. We further assume that $V(y)$ is the electron quasi-Fermi potential at a point y along the channel with respect to the Fermi potential of the n^+ source. The assumption that V is independent of x in the direction perpendicular to the surface is justified by the consideration that current is proportional to

the gradient of the quasi-Fermi potential and that the MOSFET current flows predominantly in the source-to-drain, or y-direction. At the source end of the channel, $V(y=0)=0$. At the drain end of the channel, $V(y=L)=V_{ds}$, the reverse bias of the drain-to-substrate junction since $V_{bs}=0$. For a vertical slice between the source and drain, the channel-to-substrate diode is reverse biased at $V(y)$ which plays the same role as V_R in Section 2.3.5 on MOS capacitors under nonequilibrium. As depicted in Fig. A4.5 for a reverse biased p–n junction, the electron quasi-Fermi potential is essentially flat in the vertical direction across the n-type inversion layer, and is displaced by $V(y)$ from the Fermi potential of the p-type substrate.

From Eq. (2.178) and Eq. (2.214), the electron concentration at any point (x, y) is given by

$$n(x,y)=\frac{n_i^2}{N_a}e^{q(\psi-V)/kT}. \tag{3.1}$$

Following the same approach as in Section 2.3.2, one obtains an expression for the electric field similar to that of Eq. (2.181):

$$\mathscr{E}^2(x,y)=\left(\frac{d\psi}{dx}\right)^2=\frac{2kTN_a}{\varepsilon_{si}}\left[\left(e^{-q\psi/kT}+\frac{q\psi}{kT}-1\right)+\frac{n_i^2}{N_a^2}\left(e^{-qV/kT}(e^{q\psi/kT}-1)-\frac{q\psi}{kT}\right)\right]. \tag{3.2}$$

The condition for surface inversion, Eq. (2.217), becomes

$$\psi(0,y)=V(y)+2\psi_B, \tag{3.3}$$

which is a function of y. From Eq. (2.218), the maximum depletion layer width is

$$W_{dm}(y)=\sqrt{\frac{2\varepsilon_{si}[V(y)+2\psi_B]}{qN_a}}, \tag{3.4}$$

which is also a function of y.

3.1.1.2 Pao and Sah's Double Integral

Under the assumption that both the hole current and the generation and recombination current are negligible, the current continuity equation can be applied to the electron current in the y-direction. In other words, the total drain-to-source current I_{ds} is the same at any point along the channel. From Eq. (2.63), the electron current density at a point (x, y) is

$$J_n(x,y)=-q\mu_n n(x,y)\frac{dV(y)}{dy}, \tag{3.5}$$

where $n(x, y)$ is the electron density, and μ_n is the electron mobility in the channel. The carrier mobility in the channel is generally much lower than the mobility in the bulk, due to additional surface scattering mechanisms, as will be addressed in

Section 3.1.5. ***With V(y) defined as the quasi-Fermi potential, i.e., playing the role of ϕ_n in Eq. (2.63), Eq. (3.5) includes both the drift and diffusion currents.*** The total current at a point y along the channel is obtained by multiplying Eq. (3.5) with the channel width W and integrating over the depth of the current-carrying layer. The integration is carried out from $x=0$ to x_i, where x_i is a depth into the p-type substrate but not infinity:[1]

$$I_{ds}(y) = qW\int_0^{x_i} \mu_n n(x,y)\frac{dV}{dy}dx. \tag{3.6}$$

There is a sign change, as we define $I_{ds}>0$ to be the drain-to-source current in the $-y$ direction. Since V is a function of y only, dV/dy can be taken outside the integral. We also assume that μ_n can be taken outside the integral by defining an *effective mobility*, μ_{eff}, at some average gate and drain fields. What remains in the integral is the electron concentration, $n(x, y)$. Its integration over the inversion layer gives the inversion charge per unit gate area, Q_i:

$$Q_i(y) = -q\int_0^{x_i} n(x,y)dx. \tag{3.7}$$

Equation (3.6) then becomes

$$I_{ds}(y) = -\mu_{eff}W\frac{dV}{dy}Q_i(y) = -\mu_{eff}W\frac{dV}{dy}Q_i(V). \tag{3.8}$$

In the last step, Q_i is expressed as a function of V; V is interchangeable with y, since V is a function of y only. Multiplying both sides of Eq. (3.8) by dy and integrating from 0 to L (source to drain) yield

$$\int_0^L I_{ds}dy = \mu_{eff}W\int_0^{V_{ds}}[-Q_i(V)]dV. \tag{3.9}$$

Current continuity requires that I_{ds} be a constant, independent of y. Therefore, the drain-to-source current is

$$I_{ds} = \mu_{eff}\frac{W}{L}\int_0^{V_{ds}}[-Q_i(V)]dV. \tag{3.10}$$

An alternative form of $Q_i(V)$ can be derived if $n(x, y)$ is expressed as a function of (ψ, V) using Eq. (3.1), i.e.,

$$n(x,y) = n(\psi,V) = \frac{n_i^2}{N_a}e^{q(\psi-V)/kT}, \tag{3.11}$$

[1] It does not matter what the exact choice of x_i is. See further discussions after Eq. (3.12).

and substituted into Eq. (3.7):

$$\begin{aligned} Q_i(V) &= -q\int_{\psi_s}^{\delta} n(\psi,V)\frac{dx}{d\psi}d\psi \\ &= -q\int_{\delta}^{\psi_s}\frac{(n_i^2/N_a)e^{q(\psi-V)/kT}}{\mathscr{E}(\psi,V)}d\psi. \end{aligned} \tag{3.12}$$

Here, ψ_s is the surface potential at $x=0$ and $\mathscr{E}(\psi, V) = -d\psi/dx$ is given by the square root of Eq. (3.2). The lower integration limit, δ, represents any small potential $<<kT/q$, but not zero as the integral is unbounded at $\psi = 0$.[2] Substituting Eq. (3.12) into Eq. (3.10) yields

$$I_{ds} = q\mu_{eff}\frac{W}{L}\int_0^{V_{ds}}\left(\int_{\delta}^{\psi_s}\frac{(n_i^2/N_a)e^{q(\psi-V)/kT}}{\mathscr{E}(\psi,V)}d\psi\right)dV. \tag{3.13}$$

This is referred to as *Pao and Sah's double integral* (Pao and Sah, 1966). The boundary value ψ_s is determined by two coupled equations: Eq. (2.195) and $Q_s = -\varepsilon_{si}\mathscr{E}_s(\psi_s)$ or Gauss's law, where $\mathscr{E}_s(\psi_s)$ is obtained by letting $\psi = \psi_s$ in Eq. (3.2). In depletion and inversion where $q\psi_s/kT >> 1$, only two of the terms in Eq. (3.2) are significant and need to be kept. The merged equation is then

$$\begin{aligned} V_{gs} &= V_{fb} + \psi_s - \frac{Q_s}{C_{ox}} \\ &= V_{fb} + \psi_s + \frac{\sqrt{2\varepsilon_{si}kTN_a}}{C_{ox}}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_a^2}e^{q(\psi_s-V)/kT}\right]^{1/2}, \end{aligned} \tag{3.14}$$

which is an implicit equation for $\psi_s(V)$. Equations (3.14) and (3.13) can only be solved numerically.

3.1.1.3 Charge-Sheet Model

Pao and Sah's double integral can be simplified to a single integral if the inversion charge density Q_i can be expressed as a function of ψ_s. This is accomplished by the *charge-sheet model* (Brews, 1978) which is based on the fact that the inversion layer is located very close to the silicon surface like a thin sheet of charge. There is an abrupt increase of the field (spatial integration of volume charge density) across the thin inversion layer, but very little change of the potential (spatial integration of field). As shown in the example in Fig. 2.36, neither the surface potential nor the depletion charge density changes much after strong inversion. The central assumption of the charge-sheet model is that Eq. (2.189) for the depletion charge density,

$$Q_d = -qN_aW_d = -\sqrt{2\varepsilon_{si}qN_a\psi_s}, \tag{3.15}$$

can be extended to beyond strong inversion. (Actually, once the inversion charge dominates, Q_d hardly changes with ψ_s. See Ex. 2.6.) Since the total silicon charge density

[2] For very small ψ, the integral yields Q_i/q of $(n_i^2/N_a)L_D$ ln ψ where L_D is the Debye length (Eq. (2.53)). This is many orders of magnitude below the level of interest even for a factor of 10^{10} change in ψ. But $\psi=0$ is equivalent to x_i = infinity in Eq. (3.7), in which case Q_i diverges.

Q_s is given by Eq. (3.14) or Eq. (2.195), Eq. (3.15) allows the inversion charge density to be expressed as

$$Q_i = Q_s - Q_d = -C_{ox}(V_{gs} - V_{fb} - \psi_s) + \sqrt{2\varepsilon_{si}qN_a\psi_s}. \tag{3.16}$$

It should be noted that the charge sheet model does not literally assume all the inversion charge is located at the silicon surface with a zero depth. That would mean $d|Q_i|/dV_{gs} = C_{ox}$, which is not the case with Eq. (3.16) since ψ_s also increases with V_{gs} as described by Eq. (3.14).

The variable in the drain current integral, Eq. (3.10), can be transformed from V to ψ_s,

$$I_{ds} = \mu_{eff}\frac{W}{L}\int_{\psi_{s,s}}^{\psi_{s,d}}(-Q_i(\psi_s))\frac{dV}{d\psi_s}d\psi_s, \tag{3.17}$$

where $\psi_{s,s}$, $\psi_{s,d}$ are the values of the surface potential at the source and the drain ends of the channel. For given V_{gs} and V_{ds}, they can be solved numerically from the implicit equation (3.14) by setting $V = 0$ (for $\psi_{s,s}$) and $V = V_{ds}$ (for $\psi_{s,d}$), respectively. Equation (3.14) can also be used to solve for $V(\psi_s)$,

$$V = \psi_s - \frac{kT}{q}\ln\left\{\frac{N_a^2}{n_i^2}\left[\frac{C_{ox}^2(V_{gs} - V_{fb} - \psi_s)^2}{2\varepsilon_{si}kTN_a} - \frac{q\psi_s}{kT}\right]\right\}, \tag{3.18}$$

and evaluate its derivative:

$$\frac{dV}{d\psi_s} = 1 + \frac{2kT}{q}\frac{C_{ox}^2(V_{gs} - V_{fb} - \psi_s) + \varepsilon_{si}qN_a}{C_{ox}^2(V_{gs} - V_{fb} - \psi_s)^2 - 2\varepsilon_{si}qN_a\psi_s}. \tag{3.19}$$

Substituting Eqs. (3.16) and (3.19) into Eq. (3.17) yields

$$I_{ds} = \mu_{eff}\frac{W}{L}\int_{\psi_{s,s}}^{\psi_{s,d}}\left[C_{ox}(V_{gs} - V_{fb} - \psi_s) - \sqrt{2\varepsilon_{si}qN_a\psi_s} + \frac{2kT}{q}\frac{C_{ox}^2(V_{gs} - V_{fb} - \psi_s) + \varepsilon_{si}qN_a}{C_{ox}(V_{gs} - V_{fb} - \psi_s) + \sqrt{2\varepsilon_{si}qN_a\psi_s}}\right]d\psi_s, \tag{3.20}$$

which expresses the drain current in a single integral.

It is too tedious to carry out the integral in Eq. (3.20) exactly. A second approximation is introduced in the charge sheet model (Brews, 1978) to obtain an analytical expression for the drain current. Note that the first two terms in the square bracket of Eq. (3.20) is simply $-Q_i$. Because of the kT/q multiplier, the last term in the square brackets is usually much smaller than the first two unless $Q_i \approx 0$ which happens when $C_{ox}(V_{gs} - V_{fb} - \psi_s) \approx \sqrt{2\varepsilon_{si}qN_a\psi_s}$. It is then a good approximation to apply this relation to the last term in the square brackets so that the integral can be carried out analytically:

$$I_{ds} = \mu_{eff}\frac{W}{L}\left\{C_{ox}\left(V_{gs} - V_{fb} + \frac{kT}{q}\right)\psi_s - \frac{1}{2}C_{ox}\psi_s^2 - \frac{2}{3}\sqrt{2\varepsilon_{si}qN_a}\psi_s^{3/2} + \frac{kT}{q}\sqrt{2\varepsilon_{si}qN_a\psi_s}\right\}\Bigg|_{\psi_{s,s}}^{\psi_{s,d}}. \tag{3.21}$$

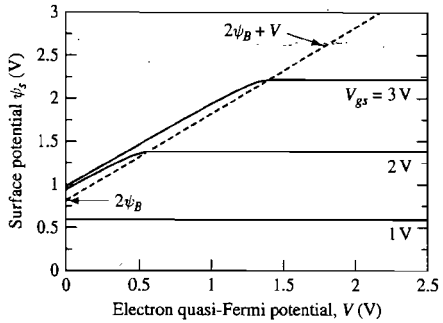


Figure 3.3. Numerical solutions of the implicit Eq. (3.14) for three values of V_{gs}. The dotted line represents the regional approximation used in Section 3.1.2. The MOS device parameters are $N_a = 10^{17}$ cm^{-3}, $t_{ox} = 10$ nm, and $V_{fb} = 0$.

Because Eq. (3.21) covers all regions of MOSFET operation: subthreshold, linear, and saturation in a single, continuous function, it has become the basis of all surface potential based compact models for circuit (SPICE) simulations (Gildenblat *et al.*, 2006). Many numerical methods have been developed to solve the implicit Eq. (3.14) for $\psi_{s,s}$ and $\psi_{s,d}$, given V_{gs} and V_{ds}. They employ either explicit approximations or iterative procedures. The general behavior of the solution $\psi_s(V)$ is shown in an example in Fig. 3.3. The device parameters are the same as those of Fig. 2.36. For $V_{gs} = 1$ V, the device is below threshold where the inversion charge is negligible, i.e., the $(n_i^2/N_a^2)e^{q(\psi_s - V)/kT}$ term in the square brackets in Eq. (3.14) is negligible. The solution ψ_s depends on V_{gs} (see Fig. 2.36), but is totally insensitive to V. For $V_{gs} = 2$ V, the MOSFET is turned on. Here, ψ_s increases more or less linearly with V when V is not too large. As V increases (for large enough V_{ds}), ψ_s reaches a saturation value beyond which it becomes independent of V. This is called the *pinch-off* condition where Q_i given by Eq. (3.16) becomes very small. The argument of the log function in Eq. (3.18) also approaches zero. For $V_{gs} = 3$ V, the saturation value of ψ_s increases while the saturation happens at a higher V (or V_{ds}).

Because of the two simplifying approximations used, the current calculated from the charge sheet model, Eq. (3.21), deviates from that of Pao and Sah's double integral, Eq. (3.13). The error is a function of doping concentration, oxide thickness, gate and drain bias voltages. Typically, it can be of the order of 10% under certain conditions when biased above threshold (Kyung, 2005). The error is generally larger in subthreshold where the current levels are low and high accuracy is not a paramount issue.

3.1.2 MOSFET *I–V* Characteristics

In this subsection, we derive the basic *I–V* characteristics of a long-channel MOSFET in the linear and parabolic regions.

3.1.2.1 Regional Approximations

To obtain explicit equations for the drain current, it is necessary to apply regional approximations to break the charge-sheet model into piecewise models. After the onset of inversion but before saturation, the surface potential can be approximated by $\psi_s = 2\psi_B + V(y)$, or Eq. (3.3). This relation is plotted in Fig. 3.3 (dotted line) for comparison with the more exact curves. It then follows that $dV/d\psi_s = 1$ and Eq. (3.17) can be readily integrated. Applying $\psi_{s,s} = 2\psi_B$ and $\psi_{s,d} = 2\psi_B + V_{ds}$, we obtain the drain current as a function of the gate and drain voltages:

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \left\{ \left(V_{gs} - V_{fb} - 2\psi_B - \frac{V_{ds}}{2} \right) V_{ds} - \frac{2\sqrt{2\varepsilon_{si} q N_a}}{3C_{ox}} \left[(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2} \right] \right\}. \tag{3.22}$$

Equation (3.22) represents the basic I–V characteristics of a MOSFET device based on the charge-sheet model. It indicates that, for a given V_{gs}, the drain current I_{ds} first increases linearly with the drain voltage V_{ds} (called the *linear* or *triode* region), then gradually levels off to a saturated value (*parabolic* region). These two distinct regions are further examined below.

3.1.2.2 Characteristics in the Linear (Triode) Region

When V_{ds} is small, one can expand Eq. (3.22) into a power series in V_{ds} and keep only the lowest-order (first-order) terms:

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \left(V_{gs} - V_{fb} - 2\psi_B - \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}} \right) V_{ds} = \mu_{eff} C_{ox} \frac{W}{L} (V_{gs} - V_t) V_{ds}, \tag{3.23}$$

where V_t is the *threshold voltage* given by

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}}. \tag{3.24}$$

Comparing this equation with Eq. (2.202), one can see that ***V_t is simply the gate voltage when the surface potential or band bending reaches $2\psi_B$ and the silicon charge (the square root) is equal to the bulk depletion charge for that potential.*** As a reminder, $2\psi_B = (2kT/q) \ln(N_a/n_i)$, which is typically 0.6–0.9 V. When V_{gs} is below V_t, there is very little current flow and the MOSFET is said to be in the *subthreshold* region, to be discussed in Section 3.1.3. Equation (3.23) indicates that, ***in the linear region, the MOSFET simply acts like a resistor with a sheet resistivity, $\rho_{sh} = 1/[\mu_{eff} C_{ox} (V_{gs} - V_t)]$, modulated by the gate voltage.*** The threshold voltage V_t can be determined by plotting I_{ds} versus V_{gs} at low drain voltages, as shown in Fig. 3.4. The extrapolated intercept of the linear portion of the $I_{ds}(V_{gs})$ curve with the V_{gs}-axis gives the approximate value of V_t. In reality, such a *linearly extrapolated threshold voltage* (V_{on}) is slightly

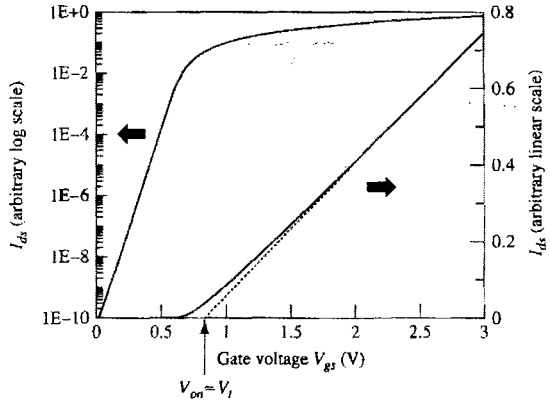


Figure 3.4. Typical MOSFET $I_{ds} - V_{gs}$ characteristics at low drain bias voltages. The same current is plotted on both linear and logarithmic scales. The dotted line illustrates the determination of the linearly extrapolated threshold voltage, V_{on}.

higher than the "$2\psi_B$" V_t due to inversion-layer capacitance and other effects, as seen in Fig. 2.36 and further addressed in Section 3.1.6. Low-drain $I_{ds}(V_{gs})$ curves are also used to extract the *effective channel length* of a MOSFET, which is discussed in Chapter 4.

3.1.2.3 Characteristics in the Parabolic Region

For larger values of V_{ds}, the second-order terms in the power series expansion of Eq. (3.22) are also important and must be kept. A good approximation to the drain current is then

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \left((V_{gs} - V_t) V_{ds} - \frac{m}{2} V_{ds}^2 \right), \tag{3.25}$$

where

$$m = 1 + \frac{\sqrt{\varepsilon_{si} q N_a / 4\psi_B}}{C_{ox}} \tag{3.26}$$

is a factor greater than one, which is related to the *subthreshold slope* and the *body effect* to be discussed in Subsections 3.1.3 and 3.1.4. Equation (3.26) can be converted to several alternative expressions by using Eq. (2.201) for the bulk depletion capacitance C_{dm} at $\psi_s = 2\psi_B$:

$$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{dm}}. \tag{3.27}$$

The last expression follows from $C_{dm} = \varepsilon_{si}/W_{dm}$, $C_{ox} = \varepsilon_{ox}/t_{ox}$, and $\varepsilon_{si}/\varepsilon_{ox} \approx 3$. A graphical interpretation of m is given in Fig. 3.5. At the threshold condition, $\psi_s = 2\psi_B$, the MOSFET acts like two capacitors, C_{ox} and C_{dm}, in series as the inversion charge capacitance is still

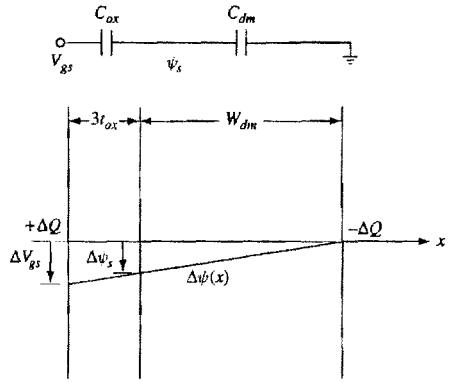


Figure 3.5. Incremental change of potential in a MOSFET due to a gate-voltage modulation near or below threshold. Grounding of the body anchored the potential on the bulk side of the depletion region where $\Delta\psi = 0$. The potential drop across the oxide, $(\Delta Q/\varepsilon_{ox})t_{ox}$, is equivalent to $(\Delta Q/\varepsilon_{si})[(\varepsilon_{si}/\varepsilon_{ox})t_{ox}]$. The factor m is defined as $\Delta V_{gs}/\Delta\psi_s$, which equals $(W_{dm} + 3t_{ox})/W_{dm}$.

negligible. The factor m equals $\Delta V_{gs}/\Delta\psi_s$, where $\Delta\psi_s$ is the incremental change of surface potential due to ΔV_{gs}, an incremental change of gate voltage. ΔV_{gs} induces sheet charge densities $+\Delta Q$ at the gate and $-\Delta Q$ at the far edge of the depletion region. They cause a field change of $\Delta\mathscr{E} = \Delta Q/\varepsilon_{si}$ in the silicon and $\Delta Q/\varepsilon_{ox}$ in the oxide, which give rise to an incremental change of potential $\Delta\psi(x)$ as shown in Fig. 3.5. Here, the oxide width is expanded to $\varepsilon_{si}/\varepsilon_{ox} \approx 3$ times its physical width so there is no change of slope at the silicon–oxide interface. ***While Eq. (3.26) is only valid for uniform bulk doping, Eq. (3.27) is more generally valid for nonuniform doping profiles*** to be discussed in Section 4.2.2. Since $1/m$ is a measure of the efficiency of the gate in modulating the surface potential, m should be kept close to one, e.g., between 1.1 and 1.4, in MOSFET design.

Equation (3.25) indicates that as V_{ds} increases, I_{ds} follows a parabolic curve, as shown in Fig. 3.6, until a maximum or saturation value is reached. This occurs when $V_{ds} = V_{dsat} = (V_{gs} - V_t)/m$, at which

$$I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2m}. \tag{3.28}$$

Equation (3.28) reduces to the well-known expression for the MOSFET saturation current when the bulk depletion charge is neglected (valid for low substrate doping) so $m = 1$. The dashed curve in Fig. 3.6 shows the trajectory of V_{dsat} through the various $I_{ds} - V_{ds}$ curves for different V_{gs}. Because of the regional approximation, $\psi_s = 2\psi_B + V$, used in the derivation, Eq. (3.22) and therefore Eq. (3.25) are valid only for $V_{ds} \le V_{dsat}$. Beyond V_{dsat}, one must go back to the more general Eq. (3.21) coupled with Eq. (3.14). Since $\psi_{s,d}$ saturates at large V_{ds} as depicted in Fig. 3.3, ***I_{ds} stays constant at I_{dsat}, independent of V_{ds} for $V_{ds} \ge V_{dsat}$.***

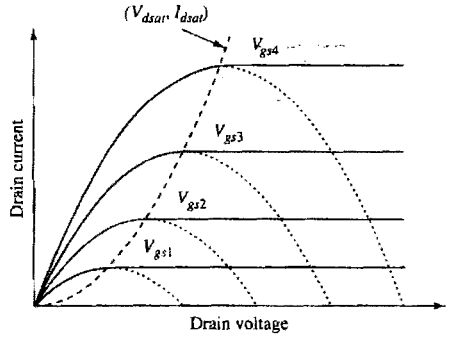


Figure 3.6. Long-channel MOSFET $I_{ds} - V_{ds}$ characteristics (solid curves) for several different values of V_{gs}. The dashed curve shows the trajectory of drain voltage beyond which the current saturates. The dotted curves help to illustrate the parabolic behavior of the characteristics before saturation.

3.1.2.4 The Onset of Pinch-Off and Current Saturation

The saturation of drain current can be understood from the inversion charge density, Eq. (3.16). For $\psi_s = 2\psi_B + V$ and $V \lesssim 2\psi_B$, one can expand the square-root term of Eq. (3.16) into a power series in V and keep only the two lowest terms,

$$Q_i(V) = -C_{ox}(V_{gs} - V_t - mV). \tag{3.29}$$

$Q_i(V)$ is plotted in Fig. 3.7. Equation (3.10) states that the drain current is proportional to the area under the $-Q_i(V)$ curve between $V = 0$ and V_{ds}. When V_{ds} is small (linear region), the inversion charge density at the drain end of the channel is only slightly lower than that at the source end. As the drain voltage increases (for a fixed gate voltage), the current increases, but the inversion charge density at the drain decreases until finally it goes to zero when $V_{ds} = V_{dsat} = (V_{gs} - V_t)/m$. At this point, I_{ds} reaches its maximum value. In other words, ***the surface channel vanishes at the drain end of the channel when saturation occurs.*** This is called *pinch-off* and is illustrated in Fig. 3.8. When V_{ds} increases beyond saturation, the pinch-off point moves toward the source, but the drain current remains essentially the same. This is because for $V_{ds} > V_{dsat}$, the voltage at the pinch-off point remains at V_{dsat} and the current, given by

$$\int_0^{L'} I_{ds} dy = \mu_{eff} W \int_0^{V_{dsat}} [-Q_i(V)] dV, \tag{3.30}$$

stays the same apart from a slight decrease in L (to L'), as shown in Fig. 3.8. This phenomenon is called *channel length modulation* and will be discussed in association with short-channel MOSFETs in Section 3.2.

Further insight into the MOSFET behavior at pinch-off can be gained by examining the function $V(y)$. Integrating from 0 to y after multiplying both sides of Eq. (3.8) by dy yields

$$\begin{aligned} I_{ds} y &= \mu_{eff} W \int_0^V [-Q_i(V)] dV \\ &= \mu_{eff} C_{ox} W \left[(V_{gs} - V_t)V - \frac{m}{2} V^2\right], \end{aligned} \tag{3.31}$$

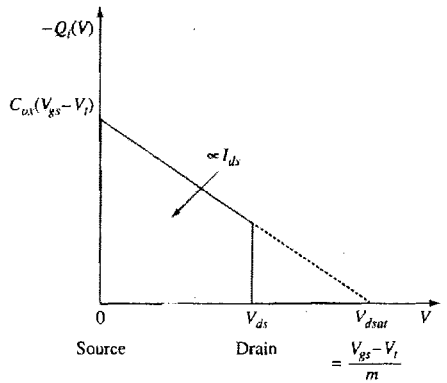


Figure 3.7. Inversion charge density as a function of the quasi-Fermi potential of a point in the channel. Before saturation, the drain current is proportional to the shaded area integrated from zero to the drain voltage.

where the simplified $Q_i(V)$, Eq.(3.29), has been used.

Substituting I_{ds} from Eq. (3.25) into Eq. (3.31), one can solve for $V(y)$:

$$V(y) = \frac{V_{gs}-V_t}{m} - \sqrt{\left(\frac{V_{gs}-V_t}{m}\right)^2 - 2\frac{y}{L}\left(\frac{V_{gs}-V_t}{m}\right)V_{ds} + \frac{y}{L}V_{ds}^2}. \tag{3.32}$$

Both $V(y)$ and $-Q_i/mC_{ox} = (V_{gs}-V_t)/m - V(y)$ are plotted in Fig. 3.9 for several values of V_{ds}. At low V_{ds}, $V(y)$ varies almost linearly between the source and drain. As V_{ds} increases, the inversion charge density at the drain decreases due to the lowering of the electron quasi-Fermi level. This is accompanied by a corresponding increase of dV/dy to maintain current continuity. When V_{ds} reaches $V_{dsat} = (V_{gs} - V_t)/m$, we have $Q_i\ (y=L) = 0$ and $V(y)$ exhibits a singularity at the drain, where $dV/dy = \infty$. This implies that ***the electric field in the y-direction changes more rapidly than the field in the x-direction and the gradual channel approximation breaks down***. In other words, beyond the pinch-off point, carriers are no longer confined to the surface channel, and a 2-D Poisson equation must be solved for carrier injection from the pinch-off point into the drain depletion region (El-Mansy and Boothroyd, 1977).

Strictly speaking, if $V_{ds} > 2\psi_B$, Eq. (3.22) cannot be expanded into a power series in V_{ds}. A more general form of the saturation voltage is obtained by letting $Q_i = 0$ in Eq. (3.16) with $\psi_s = 2\psi_B + V$ and solving for $V = V_{dsat}$ [equivalent to solving $dI_{ds}/dV_{ds} = 0$ by differentiating Eq. (3.22)]:

$$V_{dsat} = V_{gs} - V_{fb} - 2\psi_B + \frac{\varepsilon_{si} q N_a}{C_{ox}^2} - \sqrt{\frac{2\varepsilon_{si} q N_a}{C_{ox}^2}\left(V_{gs} - V_{fb} + \frac{\varepsilon_{si} q N_a}{2C_{ox}^2}\right)}. \tag{3.33}$$

The corresponding saturation current can be found by substituting Eq. (3.33) for V_{ds} in Eq. (3.22). The mathematics is rather tedious (Brews, 1981). A few selected curves are plotted in Fig. 3.10 and compared with those calculated from Eq. (3.25). It turns out that Eq. (3.25) serves as a good approximation to the drain current over a much wider range of

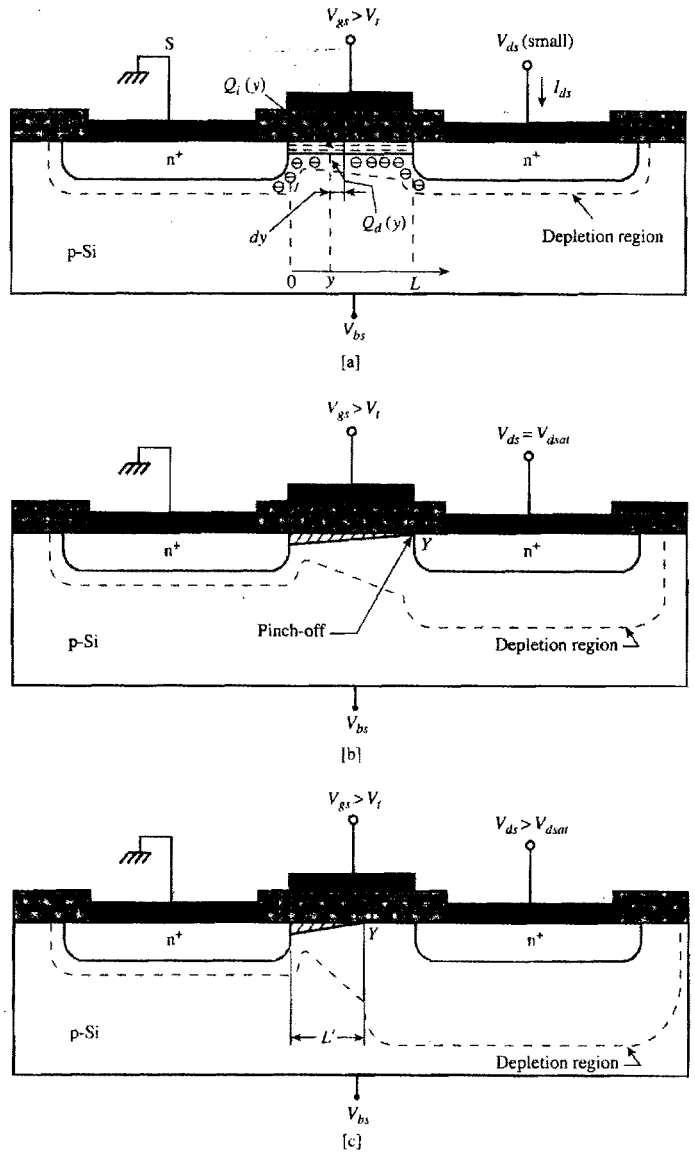


Figure 3.8. (a) MOSFET operated in the linear region (low drain voltage). (b) MOSFET operated at the onset of saturation. The pinch-off point is indicated by Y. (c) MOSFET operated beyond saturation where the channel length is reduced to L'. (After Sze, 1981.)

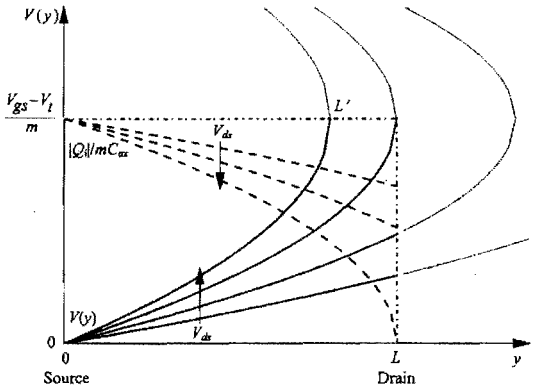


Figure 3.9. Quasi-Fermi potential versus distance between the source and the drain for several V_{ds}-values from the linear region to beyond saturation. The dashed curves show the corresponding variation of inversion charge density along the channel. The dotted curves help visualize the parabolic behavior of the characteristics.

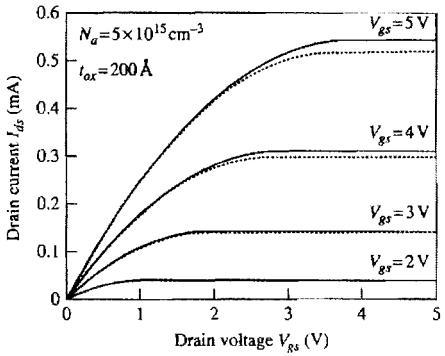


Figure 3.10. I_{ds}–V_{ds} curves calculated from the full equation (3.22) (solid curves), compared with the parabolic approximation (3.25) (dotted curves).

voltages than expected. Even for a drain voltage several times greater than $2\psi_B$, the current is only slightly (≈5%) underestimated.

3.1.2.5 pMOSFET *I*–*V* Characteristics

So far we have used an n-channel device as an example to discuss MOSFET operation and *I*–*V* characteristics. A p-channel MOSFET operates similarly, except that it is

fabricated inside an n-well with implanted p^+ source and drain regions (cf. Fig. 3.2), and that the polarities of all the voltages and currents are reversed. For example, I_{ds}–V_{ds} characteristics for a pMOSFET (cf. Fig. 3.10) have negative gate and drain voltages with respect to the source terminal for a hole current to flow from the source to the drain.

Since the source of a pMOSFET is at the highest potential compared with the other terminals, it is usually connected to the power supply V_{dd} in a CMOS circuit so that all the voltages are positive (or zero). In that case, the device conducts if the gate voltage is lower than $V_{dd} - V_t$, where $V_t(>0)$ is the magnitude of the threshold voltage of the pMOSFET. The ohmic contact to the n-well is also connected to V_{dd}, in contrast to an nMOSFET, where the p-type substrate is usually tied to the ground potential. This leaves the n-well-to-p-substrate junction reverse biased. More about nMOSFET and pMOSFET bias conditions in a CMOS circuit configuration will be given in Section 5.1.

3.1.3 Subthreshold Characteristics

Depending on the gate and source–drain voltages, a MOSFET device can be biased in one of the three regions shown in Fig. 3.11. Linear (including parabolic) and saturation region characteristics have been described in the previous subsection. In this subsection, we discuss the characteristics of a MOSFET device in the subthreshold region where $V_{gs} < V_t$. In Fig. 3.4, the drain current on a linear scale appears to approach zero immediately below the threshold voltage. On a logarithmic scale, however, the descending drain current remains at nonnegligible levels for several tenths of a volt below V_t. This is because the inversion charge density does not drop to zero abruptly. Rather, it follows an exponential dependence on ψ_s or V_g, as is evident from Eq. (3.11). Subthreshold behavior is of particular importance in low-voltage, low-power applications, such as in digital logic and memory circuits, because it describes how a MOSFET device switches off. The subthreshold region immediately below V_t, in which $\psi_B \le \psi_s \le 2\psi_B$, is also called the *weak inversion* region.

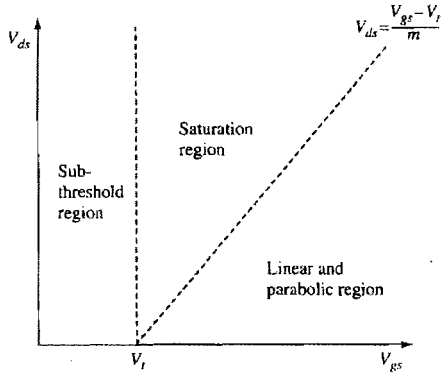


Figure 3.11. Three regions of MOSFET operation in the V_{ds}–V_{gs} plane.

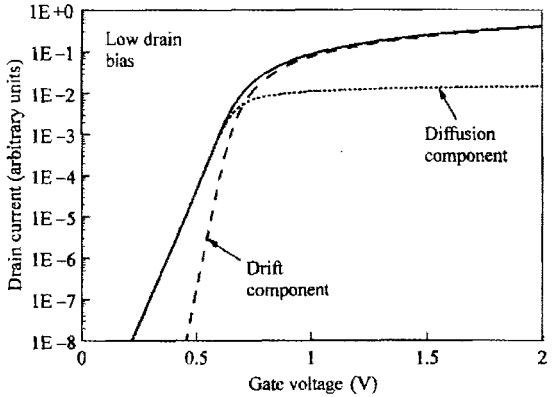


Figure 3.12. Drift and diffusion components of current in an I_{ds}–V_{gs} plot. Their sum is the total current represented by the solid curve.

3.1.3.1 Drift and Diffusion Components of Drain Current

Unlike the strong inversion region, in which the drift current dominates, subthreshold conduction is dominated by the diffusion current. Both current components are included in Pao and Sah's double integral, Eq. (3.13). In general, current continuity only applies to the total current, not to its individual components. In other words, the fractional ratio between the drift and the diffusion components may vary from one point of the channel to another. At low drain bias voltages, however, it is possible to separate the drift and diffusion components using the implicit $\psi_s(V)$ relation, Eq. (3.14). When $qV/kT \ll 1$, only the first-order terms of V need to be kept. In Eq. (3.8), $Q_i(V)$ can be replaced by its zeroth-order value, $Q_i(V=0)$; hence V must vary linearly from the source to the drain, as required by current continuity. Since the total current is proportional to dV/dy and the drift current is proportional to the electric field or $d\psi_s/dy$, the drift fraction of the current is given by the change of surface potential (band bending) with respect to the quasi-Fermi potential, i.e., $d\psi_s/dV$. This can be evaluated from Eq. (3.19) while making use of Eq. (3.14) in the limit of $V \to 0$:

$$\frac{d\psi_s}{dV} = \frac{(n_i^2/N_a^2)e^{q\psi_s/kT}}{1 + (n_i^2/N_a^2)e^{q\psi_s/kT} + (C_{ox}^2/\varepsilon_{si}qN_a)(|Q_s|/C_{ox})}, \tag{3.34}$$

where $|Q_s|/C_{ox}$ is the voltage drop across the oxide given by the last term of Eq. (3.14). It is clear that in weak inversion where $\psi_B < \psi_s < 2\psi_B$, the numerator is much less than unity and the diffusion component dominates. Conversely, beyond strong inversion, $d\psi_s/dV \approx 1$ and the drift current dominates. These kinds of behavior are further illustrated in Fig. 3.12.

3.1.3.2 Subthreshold Current Expression

To find an expression for the subthreshold current, we note from Eq. (3.14) that the total charge density in silicon is,

$$-Q_s = \varepsilon_{si}\mathscr{E}_s = \sqrt{2\varepsilon_{si}kTN_a}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_a^2}e^{q(\psi_s - V)/kT}\right]^{1/2}. \tag{3.35}$$

In weak inversion, the second term in the brackets arising from the inversion charge density is much less than the first term from the depletion charge density. Equation (3.35) can then be expanded into a power series: the zeroth-order term is identified as the depletion charge density $-Q_d$ by Eq. (3.15), and the first-order term gives the inversion charge density,

$$-Q_i = \sqrt{\frac{\varepsilon_{si}qN_a}{2\psi_s}}\left(\frac{kT}{q}\right)\left(\frac{n_i}{N_a}\right)^2 e^{q(\psi_s - V)/kT}. \tag{3.36}$$

The surface potential ψ_s is related to the gate voltage through Eq. (3.14). Since the inversion charge density is small, ψ_s is a function of V_{gs} only, independent of V (the case of $V_{gs} = 1$ V in Fig. 3.3). This also means that the electric field along the channel direction is small; hence the drift current is negligible.

Substituting Q_i into Eq. (3.10) and carrying out the integration, we obtain the drain current in the subthreshold region:

$$I_{ds} = \mu_{eff}\frac{W}{L}\sqrt{\frac{\varepsilon_{si}qN_a}{2\psi_s}}\left(\frac{kT}{q}\right)^2\left(\frac{n_i}{N_a}\right)^2 e^{q\psi_s/kT}\left(1 - e^{-qV_{ds}/kT}\right). \tag{3.37}$$

ψ_s can be expressed in terms of V_{gs} using Eq. (3.14), where only the depletion charge term needs to be kept:

$$V_{gs} = V_{fb} + \psi_s + \frac{\sqrt{2\varepsilon_{si}qN_a\psi_s}}{C_{ox}}. \tag{3.38}$$

It is straightforward to solve a quadratic equation for ψ_s. To further simplify the result, we consider ψ_s as only slightly deviated from the threshold value, $2\psi_B$ (Swanson and Meindl, 1972). Using the concept of $m = \Delta V_{gs}/\Delta\psi_s$ in Fig. 3.5, one can approximate V_{gs} as $V_{gs} = V_t + m(\psi_s - 2\psi_B)$. Solving for ψ_s and substituting it into Eq. (3.37) yield the subthreshold current as a function of V_{gs}:

$$I_{ds} = \mu_{eff}\frac{W}{L}\sqrt{\frac{\varepsilon_{si}qN_a}{4\psi_B}}\left(\frac{kT}{q}\right)^2 e^{q(V_{gs} - V_t)/mkT}\left(1 - e^{-qV_{ds}/kT}\right), \tag{3.39}$$

or

$$I_{ds} = \mu_{eff}C_{ox}\frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{q(V_{gs} - V_t)/mkT}\left(1 - e^{-qV_{ds}/kT}\right). \tag{3.40}$$

3.1.3.3 Subthreshold Slope

The subthreshold current is independent of the drain voltage once V_{ds} is larger than a few kT/q, as would be expected for diffusion-dominated current transport. The dependence on gate voltage, on the other hand, is exponential with an *inverse subthreshold slope* (Fig. 3.12),

$$S = \left(\frac{d(\log_{10} I_{ds})}{dV_{gs}}\right)^{-1} = 2.3\frac{mkT}{q} = 2.3\frac{kT}{q}\left(1 + \frac{C_{dm}}{C_{ox}}\right), \tag{3.41}$$

of typically 70–100 mV/decade. Here $m = 1 + (C_{dm}/C_{ox})$ from Eq. (3.27). If the Si–SiO_2 interface trap density is high, the subthreshold slope may be more graded than that given by Eq. (3.41), since the capacitance associated with the interface trap is in parallel with the depletion-layer capacitance C_{dm}. It should be noted that for ψ_s substantially below $2\psi_B$, e.g., when V_{gs} is a few tenths of a volt below V_t, Eq. (3.41) tends to underestimate the inverse subthreshold slope by 5–10%. As a result, the subthreshold current can be 2 to 4 times higher than that given by Eq. (3.40). For VLSI circuits, a steep subthreshold slope is desirable for the ease of switching the transistor current off. In MOSFET design, therefore, the gate oxide thickness and the bulk doping concentration should be chosen such that the factor m is not too much larger than unity, e.g., between 1.1 and 1.4. ***Nevertheless, the inverse subthreshold slope has a lower bound of 2.3 kT/q, or 60 mV/decade at room temperature, that does not change with device parameters.*** This has significant implications on device scaling as will be discussed in Chapter 4.

3.1.4 Substrate Bias and Temperature Dependence of Threshold Voltage

The threshold voltage is one of the key parameters of a MOSFET device. In this subsection, we examine the dependence of threshold voltage on substrate bias and temperature.

3.1.4.1 Substrate Sensitivity (Body Effect)

The drain-current equation in Section 3.1.2 was derived assuming zero substrate bias (V_{bs}). If $V_{bs} \neq 0$, one can modify the previously discussed MOSFET equations by considering that applying V_{bs} to the substrate is equivalent to subtracting all other voltages (namely, gate, source, and drain voltages) by V_{bs} while keeping the substrate grounded. This is shown in Fig. 3.13.

Using the charge-sheet model with $\psi_s = 2\psi_B + V$ as before, Eq. (3.16) becomes

$$Q_i = -C_{ox}(V_{gs} - V_{bs} - V_{fb} - 2\psi_B - V) + \sqrt{2\varepsilon_{si} q N_a (2\psi_B + V)}, \tag{3.42}$$

where V is the reverse bias voltage between a point in the channel and the substrate.

The current is obtained by integrating Q_i from $V = -V_{bs}$ (source) to $V_{ds} - V_{bs}$ (drain):

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}\left\{\left(V_{gs} - V_{fb} - 2\psi_B - \frac{V_{ds}}{2}\right)V_{ds} - \frac{2\sqrt{2\varepsilon_{si} q N_a}}{3C_{ox}}\left[(2\psi_B - V_{bs} + V_{ds})^{3/2} - (2\psi_B - V_{bs})^{3/2}\right]\right\}. \tag{3.43}$$

At low drain voltages (linear region), the current is still given by Eq. (3.23), except that the threshold voltage is now

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B - V_{bs})}}{C_{ox}}. \tag{3.44}$$

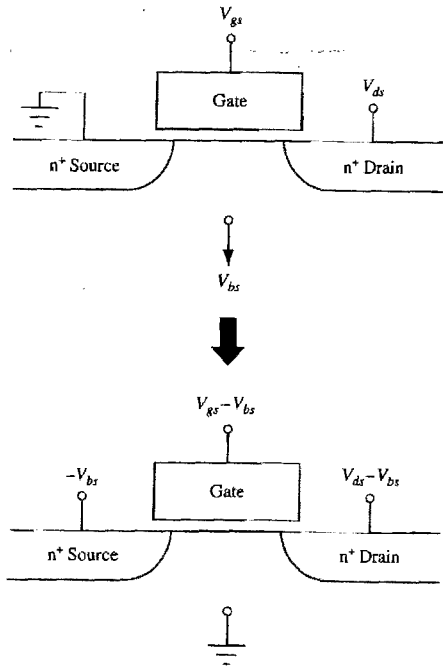


Figure 3.13. Equivalent circuits used to evaluate the effect of substrate bias on MOSFET I–V characteristics.

It can be seen from Eq. (3.44) that ***the effect of a reverse substrate bias ($V_{bs} < 0$) is to widen the bulk depletion region and raise the threshold voltage.*** Figure 3.14 plots V_t as a function of $-V_{bs}$. The slope of the curve,

$$\frac{dV_t}{d(-V_{bs})} = \frac{\sqrt{\varepsilon_{si} q N_a / [2(2\psi_B - V_{bs})]}}{C_{ox}}, \tag{3.45}$$

is referred to as the *substrate sensitivity*. At $V_{bs} = 0$, the slope equals C_{dm}/C_{ox}, or $m - 1$ [Eq. (3.26)]. The substrate sensitivity is higher for a higher bulk doping concentration. It is clear from Fig. 3.14 that the substrate sensitivity decreases as the substrate reverse bias increases. From Eq. (3.41), a reverse substrate bias also makes the subthreshold slope slightly steeper, since it widens the depletion region and lowers C_{dm}.

3.1.4.2 Temperature Dependence of Threshold Voltage

Next, we examine the temperature dependence of the threshold voltage. The flat-band voltage of an nMOSFET with n^+ polysilicon gate is $V_{fb} = -E_g/2q - \psi_B$ [Eq. (2.209)], assuming there is no oxide charge. Substituting it into Eq. (3.24) yields the threshold voltage,

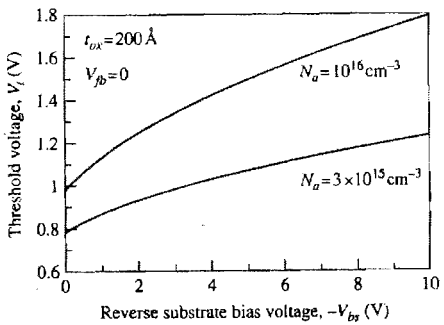


Figure 3.14. Threshold-voltage variation with reverse substrate bias for two uniform substrate doping concentrations.

$$V_t = -\frac{E_g}{2q} + \psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}}, \tag{3.46}$$

at zero substrate bias. The temperature dependence of V_t is related to the temperature dependence of E_g and ψ_B:

$$\begin{aligned}\frac{dV_t}{dT} &= -\frac{1}{2q}\frac{dE_g}{dT} + \left(1 + \frac{\sqrt{\varepsilon_{si} q N_a/\psi_B}}{C_{ox}}\right)\frac{d\psi_B}{dT} \\ &= -\frac{1}{2q}\frac{dE_g}{dT} + (2m-1)\frac{d\psi_B}{dT}.\end{aligned} \tag{3.47}$$

$d\psi_B/dT$ stems from the temperature dependence of the intrinsic carrier concentration, which can be evaluated using Eq. (2.48) and Eq. (2.13):

$$\begin{aligned}\frac{d\psi_B}{dT} &= \frac{d}{dT}\left[\frac{kT}{q}\ln\left(\frac{N_a}{\sqrt{N_c N_v e^{-E_g/2kT}}}\right)\right] \\ &= -\frac{k}{q}\ln\left(\frac{\sqrt{N_c N_v}}{N_a}\right) - \frac{kT}{q\sqrt{N_c N_v}}\frac{d\sqrt{N_c N_v}}{dT} + \frac{1}{2q}\frac{dE_g}{dT}.\end{aligned} \tag{3.48}$$

Since both N_c and N_v are proportional to $T^{3/2}$, we have $d(N_c N_v)^{1/2}/dT = \frac{3}{2}(N_c N_v)^{1/2}/T$. Substituting Eq. (3.48) into Eq. (3.47) yields

$$\frac{dV_t}{dT} = -(2m-1)\frac{k}{q}\left[\ln\left(\frac{\sqrt{N_c N_v}}{N_a}\right) + \frac{3}{2}\right] + \frac{m-1}{q}\frac{dE_g}{dT}. \tag{3.49}$$

From Section 2.1.1 and Table 2.1, $dE_g/dT \approx -2.7 \times 10^{-4}$ eV/K and $(N_c N_v)^{1/2} \approx 3 \times 10^{19}$ cm^{-3}. For $N_a \sim 10^{16}$ cm^{-3} and $m \approx 1.1$, ***dV_t/dT is typically −1 mV/K***. Note that the temperature coefficient decreases slightly as N_a increases: for $N_a \sim 10^{18}$ cm^{-3} and $m \approx 1.3$, dV_t/dT is about −0.7 mV/K. These numbers imply that, at an elevated temperature of, for example, 100°C, the threshold voltage is 55–75 mV lower than at room temperature. Since digital VLSI circuits often operate at elevated temperatures due to heat generation, this effect, plus the degradation of subthreshold slope with temperature, causes the leakage current at $V_{gs}=0$ to increase considerably over its room-temperature value. Typically, the off-state leakage current of a MOSFET at 100 °C is 30–50 times larger than the leakage current at 25 °C. These are important design considerations, to be addressed in detail in Chapter 4.

3.1.5 MOSFET Channel Mobility

The carrier mobility in a MOSFET channel is significantly lower than that in bulk silicon, due to additional scattering mechanisms. Lattice or phonon scattering is aggravated by the presence of crystalline discontinuity at the surface boundary, and surface roughness scattering severely degrades mobility at high normal felds. Channel mobility is also affected by processing conditions that alter the Si–SiO_2 interface properties (e.g., oxide charge and interface traps, as discussed in Section 2.3.6).

3.1.5.1 Effective Mobility and Effective Normal Field

In Section 3.1.1, the channel mobility was taken out of the integral by defining an effective mobility as

$$\mu_{eff} = \frac{\int_0^{x_i} \mu_n n(x)dx}{\int_0^{x_i} n(x)dx}, \tag{3.50}$$

which is essentially an average value weighted by the carrier concentration in the inversion layer. Empirically, it has been found that ***when μ_{eff} is plotted against an effective normal field $\mathscr{E}_{eff}$, there exists a universal relationship independent of the substrate bias, doping concentration, and gate oxide thickness*** (Sabnis and Clemens, 1979). The effective normal field is defined as the average electric field perpendicular to the Si–SiO_2 interface experienced by the carriers in the channel. Using Gauss's law, one can express $\mathscr{E}_{eff}$ in terms of the depletion and inversion charge densities:

$$\mathscr{E}_{eff} = \frac{1}{\varepsilon_{si}}\left(|Q_d| + \frac{1}{2}|Q_i|\right), \tag{3.51}$$

where $|Q_d| + \frac{1}{2}|Q_i|$ is the total silicon charge inside a Gaussian surface through the middle of the inversion layer. Using Eq. (2.189) and Eq. (3.24), the depletion charge can be expressed as

$$|Q_d| = \sqrt{4\varepsilon_{si} q N_a \psi_B} = C_{ox}(V_t - V_{fb} - 2\psi_B). \tag{3.52}$$

Substituting this expression and $|Q_i| \approx C_{ox}(V_{gs} - V_t)$ into Eq. (3.51) yields

$$\mathscr{E}_{eff} = \frac{V_t - V_{fb} - 2\psi_B}{3t_{ox}} + \frac{V_{gs} - V_t}{6t_{ox}}, \tag{3.53}$$

where $C_{ox} = \varepsilon_{ox}/t_{ox}$ and $\varepsilon_{si} \approx 3\varepsilon_{ox}$ were used. Equation (3.53) can further be simplifed if the gate electrode is n^+ polysilicon (for nMOSFETs) such that $V_{fb} = -E_g/2q - \psi_B$. For submicron CMOS technologies, ψ_B = 0.30–0.42 V. Therefore, the effective normal field can be expressed in terms of explicit device parameters as

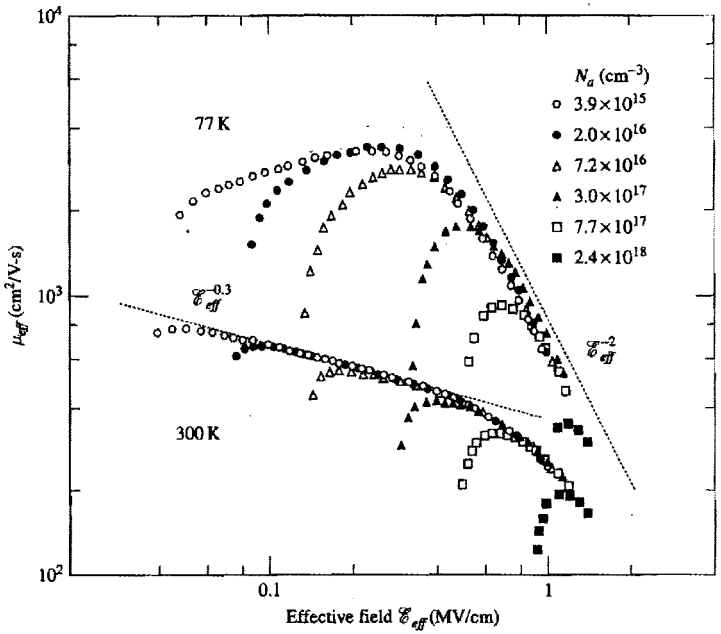


Figure 3.15. Measured electron mobility at 300 and 77 K versus effective normal field for several substrate doping concentrations. (After Takagi *et al.*, 1988).

$$\mathscr{E}_{eff} = \frac{V_t + 0.2}{3t_{ox}} + \frac{V_{gs} - V_t}{6t_{ox}}. \tag{3.54}$$

Equation (3.54) is valid for low drain voltages. At high drain voltages, Q_i decreases toward the drain end of the channel. To estimate the average effective field in that case, the second term in Eq. (3.54) should be reduced accordingly.

3.1.5.2 Electron Mobility Data

A typical set of data on mobility versus effective normal field for nMOSFETs is shown in Fig. 3.15 (Takagi *et al.*, 1988). At room temperature, the mobility follows a $\mathscr{E}_{eff}^{-1/3}$ dependence below 5×10^5 V/cm. A simple, approximate expression for this case is (Baccarani and Wordeman, 1983)

$$\mu_{eff} \approx 32500 \times \mathscr{E}_{eff}^{-1/3}. \tag{3.55}$$

Beyond $\mathscr{E}_{eff} = 5 \times 10^5$ V/cm, μ_{eff} decreases much more rapidly with increasing $\mathscr{E}_{eff}$ because of increased surface roughness scattering as carriers are distributed closer to the surface under high normal fields. For each doping concentration, there exists an effective field below which the mobility falls off the universal curve. This is believed to be due to Coulomb (or impurity) scattering, which becomes more important when the doping

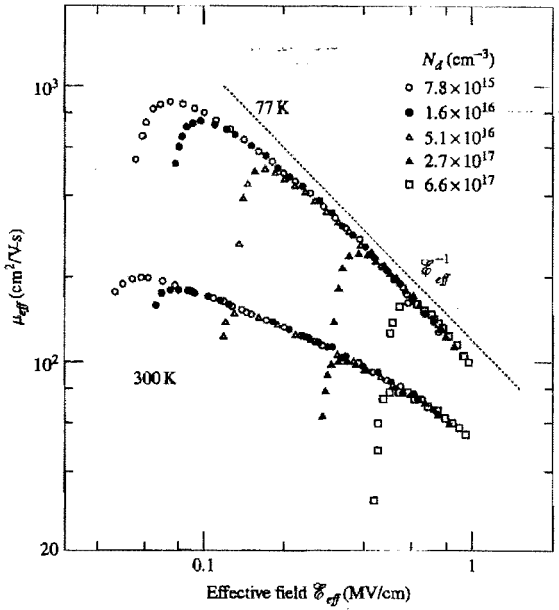


Figure 3.16. Measured hole mobility at 300 K and 77 K versus effective normal field (with a factor $\frac{1}{3}$) for several substrate doping concentrations. (After Takagi *et al.*, 1988).

concentration is high and the gate voltage or the normal field is low. There is less effect of Coulomb scattering on mobility when the inversion charge density is high because of charge screening effects. At 77 K, μ_{eff} is an even stronger function of $\mathscr{E}_{eff}$ and N_a. At low temperatures, surface scattering is the dominant mechanism at high fields, while Coulomb scattering dominates at low fields.

3.1.5.3 Hole Mobility Data

Similar mobility–field data for pMOSFETs are shown in Fig. 3.16. In this case, however, the effective normal field is defined by

$$\mathscr{E}_{eff} = \frac{1}{\varepsilon_{si}}\left(|Q_d| + \frac{1}{3}|Q_i|\right), \tag{3.56}$$

which has been found necessary in order for the measured hole mobilities to fall on a universal curve when plotted against $\mathscr{E}_{eff}$ (Arora and Gildenblat, 1987). Note that the factor $\frac{1}{3}$ is entirely empirical with no physical reasoning behind it. Like electron mobility, hole mobility is also influenced by Coulomb scattering at low fields, depending on the doping concentration. The field dependence is also stronger at 77 K, but not quite as strong as in the electron case. It should be noted that the hole mobility data were taken from surface-channel pMOSFETs with p^+ polysilicon gate. Buried-channel pMOSFETs with n^+ polysilicon gate have a higher mobility (by about 30%) for the same threshold

voltage. This is because the normal field in a buried-channel device, given by Eq. (3.53) with $V_{fb} = +E_g/2q - \psi_B$ (p^+ gate on nMOSFET as an analogy), is much lower. However, buried-channel MOSFETs cannot be scaled to as short a channel length as surface-channel MOSFETs, as will be discussed in Chapter 4.

At higher temperatures, the MOSFET channel mobility decreases because of increased phonon scattering. The temperature dependence is similar to that of bulk mobility discussed in Section 2.1.3, i.e., $\mu_{eff} \propto T^{-3/2}$.

3.1.6 MOSFET Capacitances and Inversion-Layer Capacitance Effect

In this subsection, we discuss the intrinsic capacitances of a MOSFET device in different regions of operation and the effect of finite inversion-layer capacitance on linear $I_{ds} - V_{gs}$ characteristics, which has been neglected in the regional model.

3.1.6.1 Intrinsic MOSFET Capacitances

The capacitance of a MOSFET device plays a key role in the switching delay of a logic gate, since for a given current, the capacitance determines how fast the gate can be charged (or discharged) to a certain potential which turns on (or off) the source-to-drain current. MOSFET capacitances can be divided into two main categories: intrinsic capacitances and parasitic capacitances. This sub-subsection focuses on the intrinsic MOSFET capacitances arising from the inversion and depletion charges in the channel region. Parasitic capacitances are discussed in Section 5.2. As in the earlier drain-current discussions, gate capacitances are also considered separately in the three regions of MOSFET operation: subthreshold region, linear region, and saturation region, as shown in Fig. 3.11.

- *Subthreshold region.* In the subthreshold region, the inversion charge is negligible. Only the depletion charge needs to be supplied when the gate potential is changed. Therefore, the intrinsic gate-to-source–drain capacitance is essentially zero (the extrinsic gate-to-source–drain overlap capacitance is discussed in Section 5.2.2), while the gate-to-body capacitance is given by the serial combination of C_{ox} and C_d (Fig. 2.37), i.e.,

$$C_G = WL\left(\frac{1}{C_{ox}} + \frac{1}{C_d}\right)^{-1} \approx WLC_d, \tag{3.57}$$

  where C_d is the depletion capacitance per unit area given by Eq. (2.201). Here, the upper-case subscript is used to distinguish the total gate capacitance C_G from the gate capacitance per unit area, C_g. For high drain biases, the surface potential and therefore the depletion width at the drain end of the channel become larger, according to Eq. (3.4). The average C_d to be used in Eq. (3.57) should then be slightly lower than that evaluated at the source end.
- *Linear region.* Once the surface channel forms, there is no more capacitive coupling between the gate and the body due to screening by the inversion charge. All the gate capacitances are to the channel, i.e., to the source and drain terminals. Within the framework of the regional charge-sheet model, Eq. (3.29), the inversion charge density Q_i at low drain biases varies linearly from $-C_{ox}(V_{gs} - V_t)$ at the source end to $-C_{ox}(V_{gs} - V_t - mV_{ds})$ at the drain end. The total inversion charge under the gate is then $-WLC_{ox}(V_{gs} - V_t - mV_{ds}/2)$, and the gate to channel capacitance is simply given by the oxide capacitance,

$$C_G = WLC_{ox}. \tag{3.58}$$

- *Saturation region.* When V_{ds} is appreciable, the inversion charge density $Q_i(y)$ varies parabolically along the channel, as shown in Fig. 3.9. At the pinch-off condition, $V_{ds} = V_{dsat} = (V_{gs} - V_t)/m$, and $Q_i = 0$ at the drain. In this case,

$$Q_i(y) = -C_{ox}(V_{gs} - V_t)\sqrt{1 - \frac{y}{L}}, \tag{3.59}$$

  from Eqs. (3.29) and (3.32). The total inversion charge obtained by integrating Eq. (3.59) in both the channel length (y) and the channel width directions is then $-\frac{2}{3}WLC_{ox}(V_{gs} - V_t)$, and the gate-to-channel capacitance in the saturation region is

$$C_G = \frac{2}{3}WLC_{ox}. \tag{3.60}$$

3.1.6.2 Inversion-Layer Capacitance

In Section 3.1.2 and in the discussions above, MOSFET I–V relations and capacitances were derived based on the regional approximation that all the inversion charge is located at the silicon surface and the surface potential is pinned at ψ_s (inv) $= 2\psi_B$ once the inversion layer forms. In reality, the inversion layer has a finite thickness (Fig. 2.34), and the surface potential still increases slightly with V_{gs} even beyond $2\psi_B$ (Fig. 2.36). In other words, there is a finite inversion-layer capacitance, $C_i = -dQ_i/d\psi_s$, in series with the oxide capacitance. As a result, the inversion charge density is less than that given by Eq. (3.29). The error is illustrated in the $Q_i - V_{gs}$ curves in Fig. 3.17. The dashed line represents $|Q_i| = C_{ox}(V_{gs} - V_t)$ from the piecewise model. The solid curve is a more exact solution calculated numerically from Eq. (3.12) and Eq. (3.14). The discrepancy extends to high gate voltages but is more serious for low-voltage operation. An approximate expression for the inversion charge density taking this effect into account can be derived by considering the small-signal capacitances in Fig. 2.37(b) (Wordeman, 1986),

$$\frac{d(-Q_i)}{dV_{gs}} = \frac{C_{ox}C_i}{C_{ox} + C_i + C_d} \approx C_{ox}\left(1 - \frac{1}{1 + C_i/C_{ox}}\right). \tag{3.61}$$

Here $C_d \approx 0$ after the onset of strong inversion because of screening by the inversion charge. From Eq. (2.206), $C_i \approx |Q_i|/(2kT/q)$. Since $|Q_i| \approx C_{ox}(V_{gs} - V_t)$, one can write $C_i/C_{ox} = (V_{gs} - V_t)/(2kT/q)$. Substituting it into Eq. (3.61) and integrating with respect to V_{gs}, one obtains

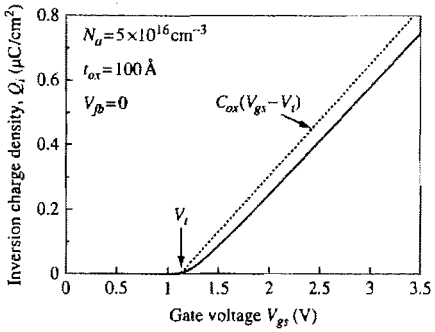


Figure 3.17. $Q_i - V_{gs}$ curve (solid line) calculated from Pao and Sah's model for zero drain voltage, compared with that of the regional approximation (dotted line). V_t indicates the $2\psi_B$ threshold.

$$-Q_i = C_{ox}\left[(V_{gs}-V_t) - \frac{2kT}{q}\ln\left(1+\frac{q(V_{gs}-V_t)}{2kT}\right)\right], \tag{3.62}$$

which agrees well with the numerically calculated curve in Fig. 3.17.

3.1.6.3 Effect of Polysilicon-Gate Depletion on Inversion Charge

Depletion of polysilicon gates, discussed in Section 2.3.4, can also have an effect on the $Q_i - V_{gs}$ curve if the gate is not doped highly enough. To first order, the depletion region in polysilicon acts like a large capacitor in series with the oxide capacitor, which further degrades inversion charge density for a given applied gate voltage. In contrast to the inversion-layer capacitance effect, however, the gate depletion effect becomes more severe at high gate voltages. Assuming an n^+ polysilicon gate on nMOSFET (and vice versa for pMOSFET) and following a similar approach to that in Eq. (2.213), one can add an additional term to Eq. (3.62) for polysilicon depletion effects:

$$-Q_i = C_{ox}\left[(V_{gs}-V_t) - \frac{2kT}{q}\ln\left(1+\frac{q(V_{gs}-V_t)}{2kT}\right) - \frac{C_{ox}^2(V_{gs}-V_t)^2}{2\varepsilon_{si}qN_p}\right]. \tag{3.63}$$

Here N_p is the electrically active doping concentration of the polysilicon gate, and the gate charge density Q_g has been approximated by $C_{ox}(V_{gs}-V_t)$ (ignoring the depletion charge in bulk silicon). Note that the factor of $\frac{1}{2}$ in the polysilicon depletion term arises from integrating a gate-voltage-dependent capacitance with respect to the gate voltage. In order to keep the last degradation term negligible, N_p should be in the range of 10^{20} cm^{-3}, especially for thin-oxide MOSFETs.

3.1.6.4 Linear $I_{ds} - V_{gs}$ Characteristics

Given $Q_i(V_{gs})$ and $\mu_{eff}(V_{gs})$ [Eqs. (3.55) and (3.54)], the low-drain-bias (linear) $I_{ds} - V_{gs}$ curve is simply

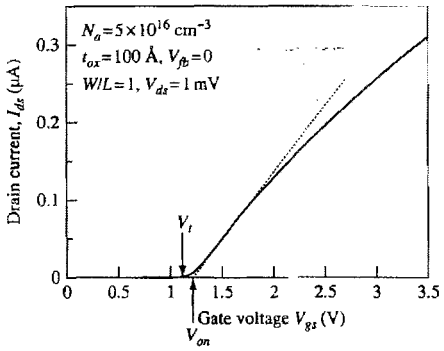


Figure 3.18. Calculated low-drain $I_{ds} - V_{gs}$ curve with inversion-layer capacitance and mobility degradation effects. The dotted line shows the linearly extrapolated threshold voltage V_{on}.

$$I_{ds}(V_{gs}) = \mu_{eff}(V_{gs})\frac{W}{L}[-Q_i(V_{gs})]V_{ds}. \tag{3.64}$$

An example is shown in Fig. 3.18, where I_{ds} is calculated assuming no polysilicon depletion with $Q_i(V_{gs})$ from Fig. 3.17 (solid curve). Both the drain current and the *linear transconductance*, defned by $g_m \equiv dI_{ds}/dV_{gs}$, are degraded signifcantly at high gate voltages because of the decrease of mobility with increasing normal field. There is a point of maximum slope or linear transconductance about 0.5 V above the threshold voltage. It is conventional to define the linearly extrapolated threshold voltage, V_{on}, by the intercept of a tangent through this point. For a second-order correction in V_{ds} based on Eq. (3.25), V_{on} is obtained by subtracting $mV_{ds}/2$ from the intercept. Because of the combined inversion-layer capacitance and mobility degradation effects, the linearly extrapolated threshold voltage, V_{on}, is typically (2–4)kT/q higher than the threshold voltage V_t at ψ_s(inv) = $2\psi_B$. One should be careful not to mix up V_{on} with V_t, which is used in Eq. (3.40) for estimating subthreshold currents. At $V_{gs} = V_{on}$, the extrapolated subthreshold current (along the same subthreshold slope in a semilog plot) is about 10× of that at $V_{gs} = V_t$. This current is rather insensitive to temperature but does depend on the technology generation.

The inversion layer capacitance in Eq. (3.62) was derived assuming classical density of states with Boltzmann statistics. The inversion layer is actually deeper than the classical value due to quantum effects (Section 4.2.4) which further degrade Q_i from that of Eq. (3.63). It is a common practice to lump all these effects into a parameter called t_{inv} by defining $-dQ_i/dV_{gs} = C_{inv} = \varepsilon_{ox}/t_{inv}$. In general, t_{inv} is a function of V_{gs} and is 5–10 Å thicker than the physical t_{ox}. A split C–V measurement like that described in Fig. 2.39 is needed to separate the t_{inv}-factor from $\mu_{eff}(V_{gs})$ in Eq. (3.64).

3.2 Short-Channel MOSFETs

It is clear from Section 3.1 that for a given supply voltage, the MOSFET current increases with decreasing channel length. The intrinsic capacitance of a short-channel MOSFET is

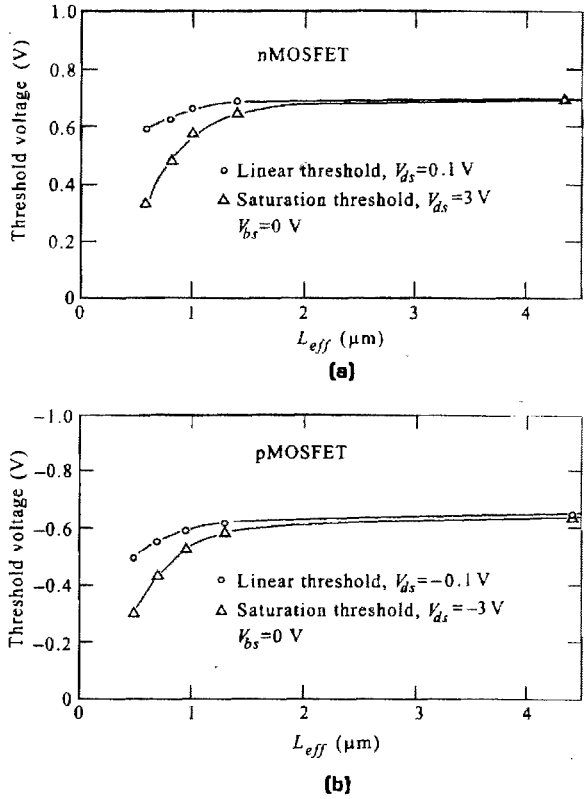


Figure 3.19. Short-channel threshold rolloff: Measured low- and high-drain threshold voltages of n- and p-MOSFETs versus channel length. (After Taur *et al.*, 1985.)

also lower, which makes it easier to switch. However, for a given process, the channel length cannot be arbitrarily reduced even if allowed by lithography. Short-channel MOSFETs differ in many important aspects from long-channel devices discussed in Section 3.1. This section covers the basic features of short-channel devices that are important for device design consideration. These features are: (a) short-channel effect, (b) velocity saturation, (c) channel length modulation, (d) source–drain series resistance, and (e) MOSFET degradation and breakdown.

3.2.1 Short-Channel Effect

The *short-channel effect* (SCE) is the decrease of the MOSFET threshold voltage as the channel length is reduced. An example is shown in Fig. 3.19 (Taur *et al.*, 1985). The short-channel effect is especially pronounced when the drain is biased at a voltage equal

to that of the power supply (high drain bias). In a CMOS VLSI technology, channel length varies statistically from chip to chip, wafer to wafer, and lot to lot due to process tolerances. The short-channel effect is therefore an important consideration in device design; one must ensure that the threshold voltage does not become too low for the minimum-channel-length device on the chip.

3.2.1.1 2-D Potential Contours in a MOSFET

The key difference between a short-channel and a long-channel MOSFET is that the field pattern in the depletion region of a short-channel MOSFET is two-dimensional, as shown in Fig. 3.20. The constant-potential contours in a long-channel device in Fig. 3.20(a) are largely parallel to the oxide–silicon interface or along the channel length direction (y-axis), so that the electric field is one-dimensional (along the vertical direction or x-axis) over the most part of the device. The constant-potential contours in a short-channel device in Fig. 3.20(b), however, are more curvilinear, and the resulting electric field pattern is of a two-dimensional nature. In other words, both the x- and y-components of the electric field are appreciable in a short-channel MOSFET. It is also important to note that, for a given gate voltage, there is more band bending (higher ψ) at the silicon–oxide interface in a short-channel device than in a long-channel device. Specifically, the maximum surface potential is slightly over 0.65 V (the fourth contour from the bottom) in Fig. 3.20(b), but below 0.65 V in Fig. 3.20(a). The depletion region width, as indicated by the depth of the first contour (ψ = 0.05 V) from the bottom, is also wider in the short-channel case. These all point to a lower threshold voltage in the short-channel MOSFET.

The two-dimensional field pattern in a short-channel device arises from the proximity of source and drain regions. Just like the depletion region under an MOS gate (Section 2.3), there are also depletion regions surrounding the source and drain junctions (Section 2.2 and Fig. 2.44). In a long-channel device, the source and drain are far enough separated that their depletion regions have no effect on the potential or field pattern in most parts of the device. ***In a short-channel device, however, the source–drain distance is comparable to the MOS depletion width in the vertical direction, and the source–drain potential has a strong effect on the band bending over a significant portion of the device.***

3.2.1.2 Drain-Induced Barrier Lowering

The physics of the short-channel effect can be understood from another angle by considering the potential barrier (to electrons for an n-channel MOSFET) at the surface between the source and drain, as shown in Fig. 3.21 (Troutman, 1979). Under off conditions, this potential barrier (p-type region) prevents electron current from flowing to the drain. The surface potential is mainly controlled by the gate voltage. When the gate voltage is below the threshold voltage, there are only a limited number of electrons injected from the source over the barrier and collected by the drain (subthreshold current). In the long-channel case, the potential barrier is flat over most parts of the device. Source and drain fields only affect the very ends of the channel. As the channel length is shortened, however, the source and drain fields penetrate deeply into the middle of the channel, which lowers the potential barrier between the source and drain. This

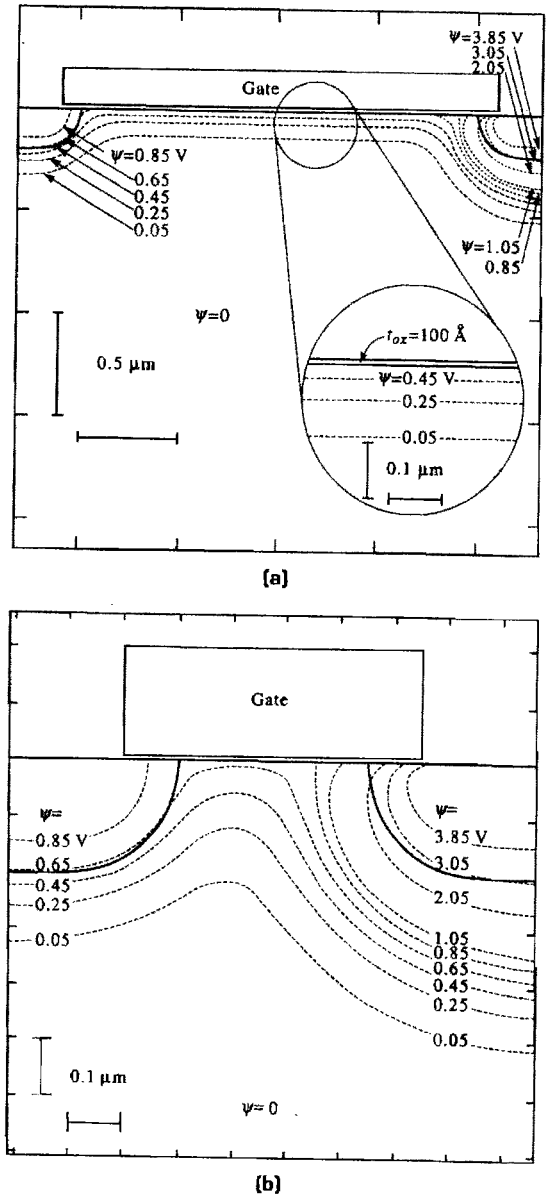


Figure 3.20. Simulated constant potential contours of (a) a long-channel and (b) a short-channel nMOSFET. The contours are labeled by the band bending with respect to the neutral p-type region. The solid lines indicate the location of the source and drain junctions (metallurgical). The drain is biased at 3.0 V. Both devices are biased at the same gate voltage slightly below the threshold.

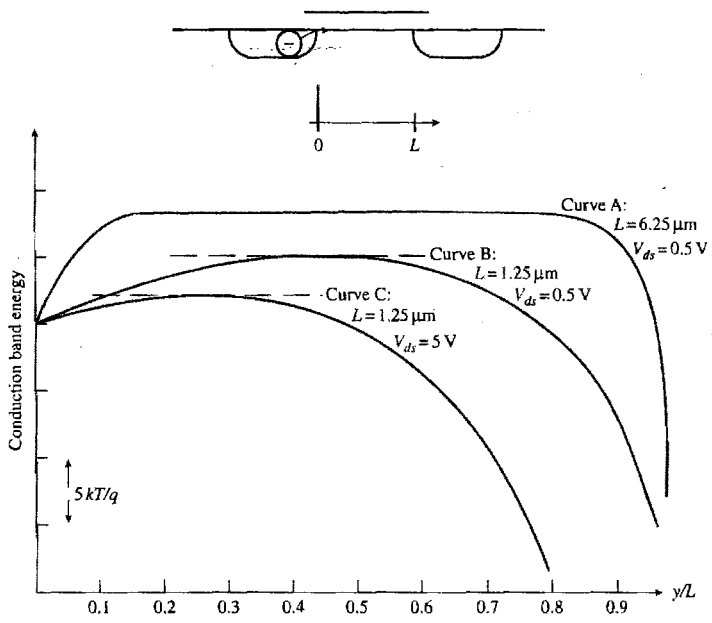


Figure 3.21. Conduction band energy versus lateral distance (normalized to the channel length L) from the source to the drain for (A) a long-channel MOSFET, (B) a short-channel MOSFET at low drain bias, and (C) a short-channel MOSFET at high drain bias. The gate voltage is the same for all three cases. (After Troutman, 1979.)

causes a substantial increase of the subthreshold current. In other words, the threshold voltage becomes lower than the long-channel value. The region of maximum potential barrier also shrinks to a single point near the center of the device.

When a high drain voltage is applied to a short-channel device, the barrier height is lowered even more, resulting in further decrease of the threshold voltage. The point of maximum barrier also shifts toward the source end as shown in Fig. 3.21. This effect is referred to as *drain-induced barrier lowering* (DIBL). It explains the experimentally observed increase of subthreshold current with drain voltage in a short-channel MOSFET. Figure 3.22 shows the subthreshold characteristics of long- and short-channel devices at different drain bias voltages. For long-channel devices, the subthreshold current is independent of drain voltage ($\geq 2kT/q$), as expected from Eq. (3.40). For short-channel devices, however, there is a parallel shift of the curve to a lower threshold voltage under high drain bias conditions. At even shorter channel lengths, the subthreshold slope starts to degrade as the surface potential is more controlled by the drain than by the gate. Eventually, the device reaches the *punch-through* condition when the gate totally loses control of the channel and high drain current persists independent of gate voltage.

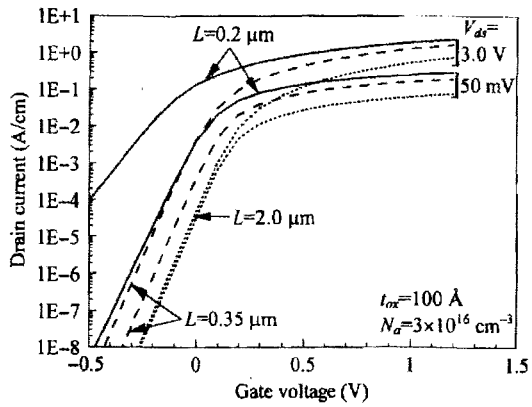


Figure 3.22. Subthreshold characteristics of long- and short-channel devices at low and high drain bias.

3.2.1.3 2-D Poisson Equation and Lateral Field Penetration

Further insight into the role of the lateral electric field, $\mathscr{E}_y = -\partial\psi_i/\partial y$, in a short-channel MOSFET can be gained by examining the two-dimensional Poisson equation,

$$\frac{\partial^2\psi_i}{\partial x^2} + \frac{\partial^2\psi_i}{\partial y^2} = -\frac{\rho}{\varepsilon_{si}}. \tag{3.65}$$

In the depletion region of an nMOSFET, mobile carrier densities are negligible. Only ionized acceptors need to be considered. For a uniformly doped background concentration N_a, Poisson's equation can be written in terms of the electric fields as

$$\frac{\partial\mathscr{E}_x}{\partial x} + \frac{\partial\mathscr{E}_y}{\partial y} = \frac{\rho}{\varepsilon_{si}} = -\frac{qN_a}{\varepsilon_{si}}, \tag{3.66}$$

where $\mathscr{E}_x = -\partial\psi_i/\partial x$ is the electric field in the vertical direction. The depletion charge density, $\rho = -qN_a$, from ionized acceptors can be considered as being split into two parts: the first part, $\varepsilon_{si}\partial\mathscr{E}_x/\partial x$, is controlled by the gate field in the vertical direction; the second part, $\varepsilon_{si}\partial\mathscr{E}_y/\partial y$, is controlled by the source–drain field in the lateral direction (Nguyen and Plummer, 1981). In a long-channel device, the lateral field is negligible over most of the channel, and almost all of the depletion charge is controlled by the gate field. In a short-channel device, the lateral field becomes appreciable. Figure 3.23(a) shows an example of the magnitude of the lateral field along the channel length direction obtained from a 2-D numerical simulation. The lateral field is highest at the source and drain junctions, decreasing exponentially toward the middle of the channel. At low drain voltages, the source and drain fields cancel each other exactly at the center of the device. ***When the channel length becomes shorter, the characteristic length of the exponential decay remains unchanged, while the magnitude of the lateral field near the middle of the device increases significantly. This depicts the penetration of source and drain fields into the channel region of a short-channel MOSFET.*** It is shown in Appendix 9 that the

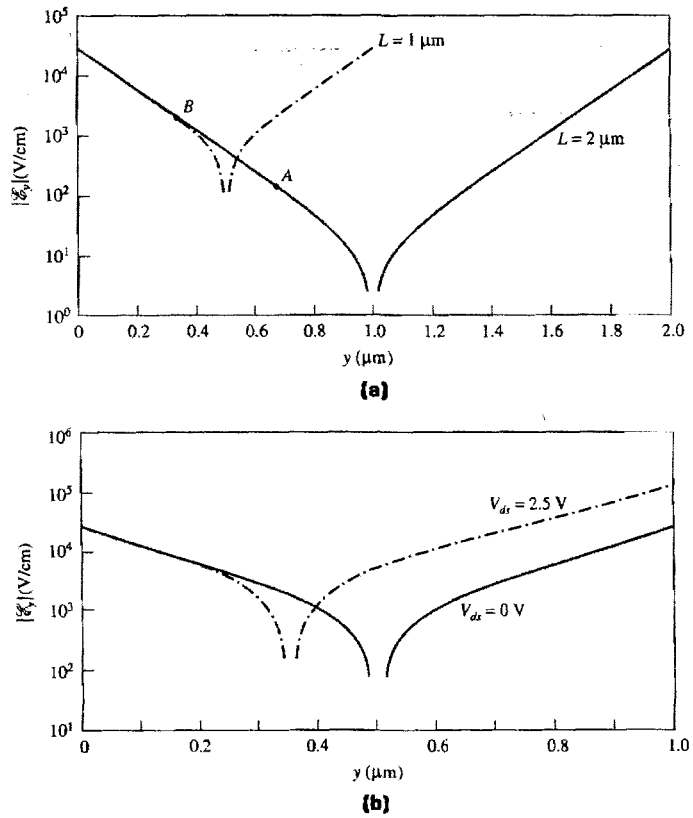


Figure 3.23. Simulated lateral field as a function of lateral distance along a horizontal cut through the gate-depletion layer for (a) long- and short-channel devices and (b) low and high drain bias voltages. (After Nguyen, 1984.)

characteristic length of the exponential decay of lateral fields is $(W_d + 3t_{ox})/\pi$ [Eq. (A9.22)], where W_d is the depth of the gate depletion region. Application of a high drain voltage [Fig. 3.23(b)] does not change the source field but does increase the drain field. This shifts the zero-field point toward the source, thus making it asymmetric, and at the same time raises the lateral field intensity even further. The zero-field point corresponds to the point of the shallowest depletion depth in Fig. 3.20(b), as well as the point of maximum energy barrier in Fig. 3.21.

3.2.1.4 An Analytical Expression for Short-Channel Threshold Voltage

With a few approximations, an analytical solution to the two-dimensional Poisson equation can be obtained using a simplifed short-channel MOSFET geometry in Appendix 9 (Nguyen, 1984). The region of interest is a rectangular box of length equal

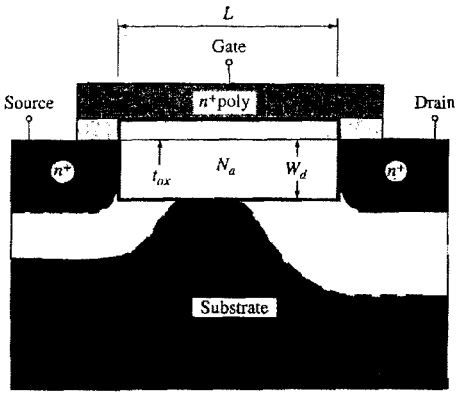


Figure 3.24. Simplified geometry for analytically solving Poisson's equation in a short-channel MOSFET. (After Nguyen, 1984.)

to the channel length L defined as the distance between the source and the drain (Fig. 3.24). In the vertical direction, the box consists of an oxide region of thickness t_{ox} and a silicon region of depth given by the depletion-layer width W_d [Eq. (2.188)]. To eliminate the discontinuity of $\partial\psi_i/\partial x$ across the silicon–oxide boundary, the oxide is replaced by an equivalent region of the same dielectric constant as silicon, but with a thickness equal to $(\varepsilon_{si}/\varepsilon_{ox})t_{ox} = 3t_{ox}$. The entire rectangular region can then be treated as a homogeneous material of height $W_d + 3t_{ox}$ and dielectric constant ε_{si}. This is a good approximation when the oxide is thin compared with the depletion depth W_d, as is the case with most practical CMOS technologies.

The boundary conditions of the electrostatic potential at the source and drain boundaries are ψ_{bi} and $\psi_{bi} + V_{ds}$, respectively, with the potential in the neutral p-type region defined as zero. Here ψ_{bi} is the built-in potential of the source- or drain-to-substrate junction, and V_{ds} is the drain voltage. For an abrupt n^+–p junction, $\psi_{bi} = E_g/2q + \psi_B$, where ψ_B is given by Eq. (2.48). Typically, $\psi_{bi} \approx 0.8$–0.9 V.

Under subthreshold conditions, current conduction is dominated by diffusion and is mainly controlled by the point of highest barrier for electrons along the channel, as shown in Fig. 3.20(b) and Fig. 3.21. The threshold voltage of a short-channel device is defined as the gate voltage at which the minimum electrostatic potential (maximum barrier for electrons) at the surface equals $2\psi_B$. It is shown in Appendix 9 [Eq. (A9.25)] that this occurs at a gate voltage lower than the long-channel threshold voltage by an amount

$$\Delta V_t = \frac{24t_{ox}}{W_{dm}}\left[\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})} - a(2\psi_B)\right]e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}. \tag{3.67}$$

Here $a \approx 0.4$ and W_{dm} is the maximum depletion width at the threshold condition, $\psi_s = 2\psi_B$. For typical values of $\psi_{bi} \approx V_{ds} \approx 2\psi_B \approx 1$ V, and $3t_{ox}/W_{dm} = m - 1 \approx 0.3$, Eq. (3.67) gives a short-channel V_t rolloff of 100 mV for $L = 2(W_{dm} + 3t_{ox})$. Since 100 mV is the

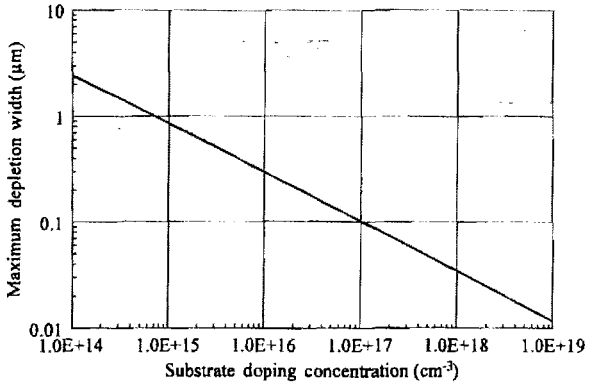


Figure 3.25. Depletion region width at $2\psi_B$ threshold condition versus doping concentration for uniformly doped substrates.

generally accepted worst-case V_t rolloff in a modern CMOS technology, ***the minimum allowable channel length is*** $L_{min} \approx 2(W_{dm} + 3t_{ox})$. Note that $W_{dm} + 3t_{ox}$ is the effective height of the rectangular box in Fig. 3.24, and L is its width. Qualitatively, the severity of the short-channel effect is measured by the aspect ratio of the rectangular box. A low aspect ratio such that $L/(W_{dm} + 3t_{ox}) > 2$ assures acceptable short-channel effects.

Because of the exponential factor, the threshold voltage rolloff with channel length is very sensitive to $W_{dm} + 3t_{ox}$, ***which can be defined as the scale length*** λ ***of the MOSFET.*** The criterion for the minimum channel length is then $L_{min} \approx 2\lambda$. To scale a MOSFET to shorter channel lengths with acceptable short-channel effects, the scale length λ needs to be reduced accordingly. This means scaling down both W_{dm} and t_{ox} by the same factor as the channel length. Note that for a uniformly doped substrate,

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a/n_i)}{q^2 N_a}} \tag{3.68}$$

from Eq. (2.190). W_{dm} is plotted in Fig. 3.25 versus N_a.

The above W_{dm} is that of a long-channel device, independent of L. This is a good approximation for $L \geq 2(W_{dm} + 3t_{ox})$. For shorter channel lengths, W_{dm} tends to increase as L decreases (see Fig. 3.20). When that happens, ΔV_t does not increase as rapidly with decreasing L as indicated by the exponential factor in Eq. (3.67). For $L \approx 1.5(W_{dm} + 3t_{ox})$, Eq. (3.67) with the long-channel W_{dm} tends to over estimate the threshold rolloff by a factor of ≈ 1.3 (Kannan, 2005). In aggressively scaled, high-performance CMOS logic technologies, L_{min} is often pushed to $\approx 1.5(W_{dm} + 3t_{ox})$. For $L \leq W_{dm} + 3t_{ox}$, the assumption that higher order terms in u_L and u_R series (Section A9.2) are negligible is no longer valid. Such devices have too severe a short-channel effect to be of practical use anyway.

All of the above discussions assume that the source and drain junction depth, x_j, is larger than the depletion region width, W_{dm}. It led to the result that V_t rolloff is controlled

by W_{dm}, insensitive to the junction depth. This is also the technologically relevant case since in practice it is difficult to scale down the junction depth without degrading the device current due to increased series resistance. But if a MOSFET with $x_j < W_{dm}$ can be made (e.g., by raised source–drain process), V_t rolloff will be linearly improved in proportion to x_j/W_{dm} (Sleva and Taur, 2005).

While analytical results like Eq. (3.67) give us key insights to the short-channel effect, in general short-channel device design is carried out with a two-dimensional device simulator for more accurate results. Further details on channel profile and threshold design are discussed in Section 4.2.

3.2.1.5 Generalized Scale Length with High-κ Gate Dielectrics

When the CMOS channel length is scaled to 20–30 nm, gate oxides of ≈1 nm thickness (Section 4.2.3) become necessary for control of short-channel effects. While a combination of improvement in process technology and better understanding of the breakdown process of gate oxides (Section 2.5.6) has made this possible, gate tunneling currents can be unacceptably high for such atomically thin oxides (Fig. 2.62). This problem can be mitigated using high-permittivity (high-κ) dielectrics to replace SiO_2 as gate insulators. From the normal field (in silicon) and gate capacitance point of view, a high-κ gate dielectric of permittivity ε_i and thickness

$$t_i = \frac{\varepsilon_i}{\varepsilon_{ox}} t_{ox} \tag{3.69}$$

is equivalent to an oxide layer of permittivity ε_{ox} and thickness t_{ox}. If $\varepsilon_i/\varepsilon_{ox} >> 1$, the physical thickness of the high-κ gate dielectric t_i is much thicker than t_{ox}, thus significantly reducing the gate tunneling current (quantum mechanical tunneling has nothing to do with the dielectric constant of the material). In practice, it is rather difficult to develop a high-κ gate insulator with acceptable characteristics for use in CMOS products. High-κ gate insulator is currently one of the intensely researched subjects in the field of VLSI.

From the one-region scale length model, one would expect that $\lambda = W_{dm} + (\varepsilon_{si}/\varepsilon_i)t_i$ for high-κ gate dielectrics. However, that is correct only for $t_i << \lambda$ when the normal fields dominate. As both ε_i and t_i increase by the same factor, i.e., at constant capacitance ε_i/t_i, tangential fields become more important for which the high dielectric constant does not help. For arbitrary gate dielectric constant and thickness, it is necessary to apply the generalized two-region scale length model described in Appendix 10. By matching the boundary conditions for both the normal and the tangential fields at the silicon–insulator interface, an eigenvalue equation for the scale length λ is obtained [Eq. (A10.7)]:

$$\frac{1}{\varepsilon_i} \tan\left(\frac{\pi t_i}{\lambda}\right) + \frac{1}{\varepsilon_{si}} \tan\left(\frac{\pi W_{dm}}{\lambda}\right) = 0. \tag{3.70}$$

This equation has an infinite number of solutions, in descending order of λ. The lowest order eigenvalue or the longest λ dominates because the short-channel potential

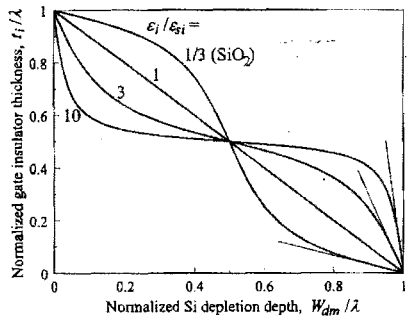


Figure 3.26. Numerical solutions to Eq. (3.70) for different values of $\varepsilon_i/\varepsilon_{si}$. The dotted lines at the lower right corner depict the asymptotic solution behavior, $\lambda \approx W_{dm} + (\varepsilon_{si}/\varepsilon_i)t_i$, for $t_i << W_{dm}$.

component is proportional to $\exp(-\pi L/2\lambda)$ as in the one-region model. Equation (3.70) cannot be solved in closed forms. The numerical solution for the longest λ is shown in normalized units in Fig. 3.26 for several representative values of $\varepsilon_i/\varepsilon_{si}$. ***The significance of λ remains that it dictates the minimum channel length, $L_{min} \approx 2\lambda$***, as in the one-region case discussed before.

The following characteristics of the solution to Eq. (3.70) are observed in Fig. 3.26.

- $\lambda \geq W_{dm}$ and $\lambda \geq t_i$, i.e., λ is larger than the larger of W_{dm}, t_i.
- In the special case of $\varepsilon_i = \varepsilon_{si}$, $\lambda = W_{dm} + t_i$, the physical height of the box in Fig. A10.1.
- In the special case of $W_{dm} = t_i$, $\lambda = 2W_{dm} = 2t_i$, regardless of ε_i, ε_{si}.
- If $t_i << W_{dm}$ (lower right corner of Fig. 3.26), $\lambda \approx W_{dm} + (\varepsilon_{si}/\varepsilon_i)t_i$. This is the approximate solution obtained in the one-region model.
- If $W_{dm} << t_i$ (upper left corner of Fig. 3.26), $\lambda \approx t_i + (\varepsilon_i/\varepsilon_{si})W_{dm}$.

While Eq. (3.70) and thus Fig. 3.26 are symmetric with respect to t_i and W_{dm}, consideration of $\Delta V_{gs}/\Delta\psi_s$ or the m-factor in Fig. 3.5 requires that $t_i/\varepsilon_i < W_{dm}/\varepsilon_{si}$. In other words, only the λ solutions in the lower right corner of Fig. 3.26 are acceptable. In that region, high-κ gate dielectric helps because $\lambda \approx W_{dm} + (\varepsilon_{si}/\varepsilon_i)t_i$ and for the same $W_{dm}/\lambda \lesssim 1$, higher $\varepsilon_i/\varepsilon_{si}$ allows a larger t_i/λ. However, because of the extreme nonlinearity of the curves for $\varepsilon_i/\varepsilon_{si} >> 1$, the λ solution quickly departs from the above one-region, linear approximation (dotted lines in Fig. 3.26) as t_i/λ increases. There exists a limit of $t_i/\lambda \leq \frac{1}{2}$, or $\lambda \geq 2t_i$, where t_i is physical thickness of the insulator no matter how high the dielectric constant is. Physically, this is caused by the lateral fields which, unlike the vertical fields, are not affected by the dielectric constant of the material (Section 2.1.4.2). In the devices with thick, very high-κ gate insulators, the short-channel effect is dominated by the lateral fields such that the scale length is determined mainly by the physical thickness of the film.

Notice that for $\varepsilon_i/\varepsilon_{si} < 1$, e.g., SiO_2, the curvature of the curve in Fig. 3.26 is opposite to those of $\varepsilon_i/\varepsilon_{si} > 1$. This means that λ is somewhat lower (better) than the one-region approximation, $W_{dm} + 3t_{ox}$, as t_{ox} increases.

3.2.2 Velocity Saturation and High-Field Transport

As discussed in Section 3.1.2, when the drain voltage increases in a long-channel MOSFET, the drain current first increases, then becomes saturated at a voltage equal to $V_{dsat} = (V_{gs} - V_t)/m$ with the onset of pinch-off at the drain. ***In a short-channel device, the saturation of drain current may occur at a much lower voltage due to velocity saturation***. This causes the saturation current I_{dsat} to deviate from the $1/L$ dependence depicted in Eq. (3.28) for long-channel devices. Velocity–field relationships in bulk silicon are plotted in Fig. 2.10. Saturation velocities of electrons and holes in a MOSFET channel are slightly lower than their bulk values. $v_{sat} \approx 7$–8×10^6 cm/s for electrons and $v_{sat} \approx 6$–7×10^6 cm/s for holes have been reported in the literature (Coen and Muller, 1980; Taur *et al.*, 1993a). Figure 3.27 shows the experimentally measured $I_{ds} - V_{ds}$ curves of a 0.25-μm nMOSFET. The dashed curve represents the long-channel-like current given by Eq. (3.28) for $V_{gs} = 2.5$ V. Due to velocity saturation, the drain current saturates at a drain voltage much lower than $(V_{gs} - V_t)/m$, thus severely limiting the saturation current of a short-channel device.

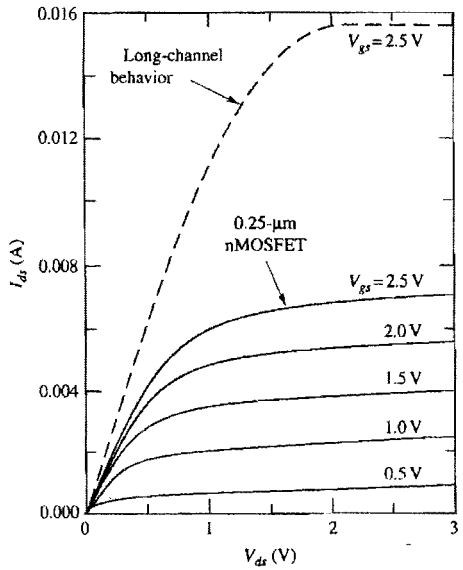


Figure 3.27. Experimental I–V curves of a 0.25-μm nMOSFET (solid lines). The device width is 9.5 μm. The dashed curve shows the long-channel-like drain current expected for this channel length if there were no velocity saturation. (After Taur *et al.*, 1993a.)

3.2.2.1 Velocity–Field Relationship

Experimental measurements show that the velocity–field relationship for electrons and holes takes the empirical form (Caughey and Thomas, 1967)

$$v = \frac{\mu_{eff}\mathscr{E}}{[1 + (\mathscr{E}/\mathscr{E}_c)^n]^{1/n}}, \tag{3.71}$$

where $n = 2$ for electrons and $n = 1$ for holes. n (≥ 1) is a measure of how rapidly the carriers approach saturation. The parameter $\mathscr{E}_c$ is called the *critical field*. When the field strength is comparable to or greater than $\mathscr{E}_c$, velocity saturation becomes important. At low fields, $v = \mu_{eff}\mathscr{E}$, which is simply Ohm's law. As $\mathscr{E} \to \infty$, $v = v_{sat} = \mu_{eff}\mathscr{E}_c$. Therefore,

$$\mathscr{E}_c = \frac{v_{sat}}{\mu_{eff}}. \tag{3.72}$$

It was discussed in Section 3.1.5 that the effective mobility μ_{eff} is a function of the vertical (or normal) field $\mathscr{E}_{eff}$. Since v_{sat} is a constant independent of $\mathscr{E}_{eff}$, the critical field $\mathscr{E}_c$ is a function of $\mathscr{E}_{eff}$ as well. More specifically, ***for a higher vertical field, the effective mobility is lower, but the critical field for velocity saturation becomes higher*** (Sodini *et al.*, 1984). Similarly, holes have a critical field higher than that of electrons, since hole mobilities are lower.

3.2.2.2 An Analytical Solution for $n = 1$

It is more important to treat velocity saturation for electrons. However, the mathematics in solving the $n = 2$ case is rather tedious (Taylor, 1984). An insight into the velocity saturation phenomenon in a MOSFET can be gained by analyzing the $n = 1$ case, which yields a simple and continuous solution. Following similar steps to those in Section 3.1.1, one replaces the low-field drift velocity, $-\mu_{eff}\,dV/dy$, in Eq. (3.8) with Eq. (3.71) to allow for high-field velocity saturation effects ($n = 1$):

$$I_{ds} = -WQ_i(V)\frac{\mu_{eff}\,dV/dy}{1 + (\mu_{eff}/v_{sat})\,dV/dy}. \tag{3.73}$$

Here V is the quasi-Fermi potential at a point y in the channel, and $Q_i(V)$ is the integrated (vertically) inversion charge density at that point. Note that $dV/dy = -\mathscr{E} > 0$.[3] Current continuity requires that I_{ds} be a constant, independent of y.

Rearranging Eq. (3.73), one obtains

$$I_{ds} = -\left(\mu_{eff}WQ_i(V) + \frac{\mu_{eff}I_{ds}}{v_{sat}}\right)\frac{dV}{dy}. \tag{3.74}$$

Multiplying by dy on both sides and integrating from $y = 0$ to L and from $V = 0$ to V_{ds}, one solves for I_{ds}:

[3] Strictly speaking, $\mathscr{E} = -d\psi_s/dy$. Above the threshold, the current is dominated by the drift component; hence $d\psi_s/dy$ and dV/dy are interchangeable.

$$I_{ds} = \frac{-\mu_{eff}(W/L)\int_0^{V_{ds}} Q_i(V)\,dV}{1 + (\mu_{eff}V_{ds}/v_{sat}L)}. \tag{3.75}$$

The numerator is simply the long-channel current, Eq. (3.10), without velocity saturation. It is clear that if the "average" field along the channel, V_{ds}/L, is much less than the critical field $\mathscr{E}_c = v_{sat}/\mu_{eff}$, the drain current is hardly affected by velocity saturation. When V_{ds}/L becomes comparable to or greater than $\mathscr{E}_c$, however, the drain current is significantly reduced. If one uses the approximate expression (3.29) in the regional charge-sheet model for $Q_i(V)$,

$$Q_i(V) = -C_{ox}(V_{gs} - V_t - mV), \tag{3.76}$$

the integration in Eq. (3.75) can be carried out to yield

$$I_{ds} = \frac{\mu_{eff}C_{ox}(W/L)[(V_{gs} - V_t)V_{ds} - (m/2)V_{ds}^2]}{1 + (\mu_{eff}V_{ds}/v_{sat}L)}. \tag{3.77}$$

For a given V_{gs}, I_{ds} increases with V_{ds} until a maximum current is reached. Beyond this point, the drain current is saturated. The saturation voltage, V_{dsat}, can be found by solving $dI_{ds}/dV_{ds} = 0$:

$$V_{dsat} = \frac{2(V_{gs} - V_t)/m}{1 + \sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)}}. \tag{3.78}$$

This expression is always less than the long-channel saturation voltage, $(V_{gs} - V_t)/m$. Substituting Eq. (3.78) into Eq. (3.77), one finds the saturation current,

$$I_{dsat} = C_{ox}Wv_{sat}(V_{gs} - V_t)\frac{\sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)} - 1}{\sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)} + 1}. \tag{3.79}$$

Example curves of I_{dsat} versus $V_{gs} - V_t$ are plotted in Fig. 3.28 for several different channel lengths. In the long-channel case, the solid curve calculated from Eq. (3.79) is not too different from the dashed curve representing the drain current without velocity saturation. In fact, it can be shown that Eq. (3.79) reduces to the long-channel saturation current [Eq. (3.28)],

$$I_{dsat} = \mu_{eff}C_{ox}\frac{W}{L}\frac{(V_{gs} - V_t)^2}{2m}, \tag{3.80}$$

when $V_{gs} - V_t << mv_{sat}L/2\mu_{eff}$. As the channel length becomes shorter, the velocity-saturated current (solid curves) is significantly less than that of Eq. (3.80) (dashed curves) over an increasing range of gate voltage. In the limit of $L \to 0$, Eq. (3.79) becomes the velocity-saturation-limited current,

$$I_{dsat} = C_{ox}Wv_{sat}(V_{gs} - V_t), \tag{3.81}$$

as indicated by the straight line labeled $L = 0$ in Fig. 3.28. ***Note that Eq. (3.81) is independent of channel length L and varies linearly with $V_{gs} - V_t$ instead of***

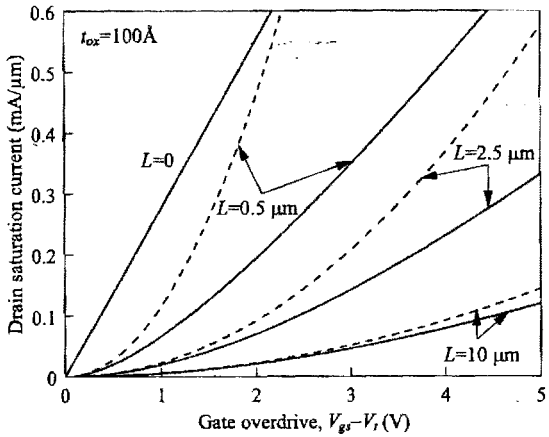


Figure 3.28. Saturation current calculated from Eq. (3.79) versus $V_{gs} - V_t$ for several different channel lengths (solid curves). The dashed curves are the corresponding "long-channel-like" saturation currents calculated from Eq. (3.80), i.e., by letting $v_{sat} \to \infty$ in Eq. (3.79). The $L = 0$ line represents the limiting case imposed by velocity saturation, Eq. (3.81).

quadratically as in the long-channel case. This is consistent with observations of the experimental curves in Fig. 3.27. For very short channel lengths, the saturation voltage, Eq. (3.78), can be approximately by

$$V_{dsat} = \sqrt{2v_{sat}L(V_{gs} - V_t)/m\mu_{eff}}, \tag{3.82}$$

which decreases with channel length.

It is instructive to examine the charge and field behavior at the drain end of the channel when $V_{ds} = V_{dsat}$. From Eq. (3.76),

$$Q_i(y = L) = -C_{ox}(V_{gs} - V_t - mV_{dsat}). \tag{3.83}$$

Substituting V_{dsat} from Eq. (3.78), one finds

$$Q_i(y = L) = -C_{ox}(V_{gs} - V_t)\frac{\sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)} - 1}{\sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)} + 1}. \tag{3.84}$$

Comparison with Eq. (3.79) yields $I_{dsat} = -Wv_{sat}Q_i(y = L)$, i.e., the carrier drift velocity at the drain end of the channel is equal to the saturation velocity. From Eq. (3.73), this means that the lateral field along the channel, dV/dy, approaches infinity at the drain. Just as in the long-channel pinch-off situation discussed in Subsection 3.1.2, such a singularity leads to the breakdown of the gradual-channel approximation which assumed that the lateral field changes slowly in comparison with the vertical field. In other words, ***beyond the saturation point, carriers which are traveling at saturation velocity are no longer confined to the surface channel.*** Their transport must then be described by a 2-D

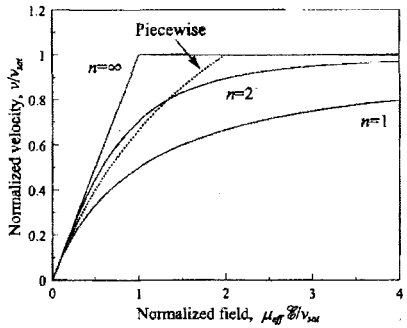


Figure 3.29. Velocity–field relationship of various velocity saturation models plotted in normalized units. The rate of approaching saturation velocity differs in different models.

Poisson equation. A key difference between pinch-off in long-channel devices and velocity saturation in short-channel devices is that in the latter case, the inversion charge density at the drain, Eq. (3.84), does not vanish.

3.2.2.3 $n = \infty$ Velocity Saturation Model

Other than the $n = 1$ velocity saturation model discussed above, analytical solutions also exist in the $n = \infty$ case and a piecewise model depicted in Fig. 3.29. The steepest approach to v_{sat} is obtained by letting $n \to \infty$ in Eq. (3.71):

$$v = \mu_{eff}\mathscr{E} \quad \text{for} \quad \mathscr{E} < v_{sat}/\mu_{eff}, \tag{3.85}$$

and

$$v = v_{sat} \quad \text{for} \quad \mathscr{E} > v_{sat}/\mu_{eff}. \tag{3.86}$$

For $v < v_{sat}$, the current expression is the same as the long-channel result, Eq. (3.25). In this case, however, before V_{ds} reaches the pinch-off value, $(V_{gs} - V_t)/m$, carrier velocity at the drain end of the channel reaches $v = v_{sat}$ and the current saturates. If this happens at $V_{ds} = V_{dsat}$, then the saturated current is

$$I_{dsat} = \mu_{eff} C_{ox} \frac{W}{L}\left[(V_{gs} - V_t)V_{dsat} - \frac{m}{2}V_{dsat}^2\right]. \tag{3.87}$$

On the other hand, at the drain end of the channel,

$$I_{dsat} = -WQ_i v_{sat} = WC_{ox} v_{sat}(V_{gs} - V_t - mV_{dsat}), \tag{3.88}$$

where Eq. (3.29) is used for Q_i. Equating Eqs. (3.87) and (3.88) allows V_{dsat} to be solved:

$$V_{dsat} = \frac{V_{gs} - V_t}{m} + \frac{Lv_{sat}}{\mu_{eff}} - \sqrt{\left(\frac{V_{gs} - V_t}{m}\right)^2 + \left(\frac{Lv_{sat}}{\mu_{eff}}\right)^2}. \tag{3.89}$$

Substituting V_{dsat} back into Eq. (3.88) yields the saturation current,

$$I_{dsat} = WC_{ox} v_{sat}\left\{\sqrt{(V_{gs} - V_t)^2 + \left(\frac{mLv_{sat}}{\mu_{eff}}\right)^2} - \frac{mLv_{sat}}{\mu_{eff}}\right\}. \tag{3.90}$$

Just like the $n = 1$ case, Eq. (3.90) is also reduced to the long-channel limit, Eq. (3.80), and the fully velocity saturated limit, Eq. (3.81), in the limits of $v_{sat} \to \infty$ and $L \to 0$, respectively.

3.2.2.4 A Piecewise-Continuous Velocity Saturation Model

It was mentioned before that while electrons behave like the $n = 2$ model, the analytic solution is too tedious to deal with. A piecewise continuous velocity saturation model was developed instead to approximate the $n = 2$ characteristics (Sodini *et al.*, 1984). At low to moderate fields, the velocity varies with the field like that of an $n = 1$ model except that v_{sat} is replaced with $2v_{sat}$. At high fields, the velocity saturates at v_{sat} once it is reached. In other words,

$$v = \frac{\mu_{eff}\mathscr{E}}{1 + (\mu_{eff}\mathscr{E}/2v_{sat})} \quad \text{for} \quad \mathscr{E} < 2v_{sat}/\mu_{eff}, \tag{3.91}$$

and

$$v = v_{sat} \quad \text{for} \quad \mathscr{E} > 2v_{sat}/\mu_{eff}. \tag{3.92}$$

The piecewise-continuous model is also shown in Fig. 3.29 (dotted curve). It has been adopted in various BSIM compact models for MOSFETs.

Analytic expressions for the saturation voltage and current can be readily developed for the piecewise model. Before saturation, the drain current is like that of Eq. (3.73) with v_{sat} replaced by $2v_{sat}$. Going through the same derivation as before, one obtains the $I_{ds}(V_{ds})$ results of Eq. (3.77) with v_{sat} replaced by $2v_{sat}$. In particular, at $V_{ds} = V_{dsat}$ when the velocity at the drain reaches v_{sat}, $I_{ds} = I_{dsat}$, i.e.,

$$I_{dsat} = \frac{\mu_{eff}C_{ox}(W/L)\left[(V_{gs} - V_t)V_{dsat} - (m/2)V_{dsat}^2\right]}{1 + (\mu_{eff}V_{dsat}/2v_{sat}L)}. \tag{3.93}$$

I_{dsat} is also related to V_{dsat} by Eq. (3.88) considering velocity saturation at the drain end of the channel. V_{dsat} is then solved by equating Eqs. (3.93) and (3.88):

$$V_{dsat} = \frac{(V_{gs} - V_t)/m}{1 + \mu_{eff}(V_{gs} - V_t)/(2mv_{sat}L)}. \tag{3.94}$$

Substituting V_{dsat} into Eq. (3.88) yields the saturation current (Sodini *et al.*, 1984),

$$I_{dsat} = \frac{\mu_{eff} C_{ox} (W/L)(V_{gs} - V_t)^2/2m}{1 + \mu_{eff}(V_{gs} - V_t)/(2mv_{sat}L)}. \tag{3.95}$$

Again, Eq. (3.95) is reduced to the long-channel limit, Eq. (3.80), and the fully velocity saturated limit, Eq. (3.81), in the limits of $v_{sat} \to \infty$ and $L \to 0$, respectively.

Owing to the piecewise-continuous nature of this model and the $n = \infty$ model, the resulting $I_{ds}(V_{ds})$ curves are also piecewise-continuous. That is, $I_{ds}(V_{ds})$ is continuous, but its derivative is not at the point of saturation. This may cause some problems when they are implemented in compact models for circuit simulation.

3.2.2.5 Velocity Overshoot

All the MOSFET current formulations discussed thus far, including the mobility definition and velocity saturation, are under the realm of the drift–diffusion approximation, which treats carrier transport in some average fashion close to thermal equilibrium with the silicon lattice. ***The drift–diffusion model breaks down in ultrashort-channel devices where high field or rapid spatial variation of potential is present.*** In such cases, the scattering events are no longer localized, and some fraction of the carriers may acquire much higher than thermal energy over a portion of the device, for example, near the drain. These carriers are not in thermal equilibrium with the silicon lattice and are generally referred to as *hot carriers*. Under these circumstances, it is possible for the carrier velocity to exceed the saturation velocity. This phenomenon is called *velocity overshoot*.

A more rigorous treatment of the carrier transport under spatially nonuniform high-field conditions has been carried out by a Monte Carlo solution of the Boltzmann transport equation for the electron distribution function (Laux and Fischetti, 1988). Figure 3.30 shows the calculated saturation transconductance of nMOSFETs versus channel length, together with experimental results (Sai-Halasz *et al.*, 1988). Local velocity overshoot near the drain starts to occur below 0.2-μm channel length. At channel lengths near 0.05 μm, velocity overshoot takes place over a substantial portion of the device such that the terminal saturation transconductance may exceed the velocity saturation limited value,

$$g_{msat} \equiv \frac{dI_{dsat}}{dV_g} = C_{ox} W v_{sat}, \tag{3.96}$$

from Eq. (3.81).

3.2.2.6 Ballistic MOSFET and Scattering Theory

It should be noted that while the carrier velocity can reach rather high values in the high-field region near the drain, it does not lead to proportionately high currents. This is illustrated in Fig. 3.31 where the band diagram of a MOSFET biased in saturation is shown. At a point near the drain, the average carrier velocity v_d is high, but the inversion charge sheet density $Q_i = C_{ox}(V_{gs} - V_t - mV_{dsat})$ is low. V_{dsat} will assume a value to maintain current continuity such that

$$I_{ds} = WC_{ox}(V_{gs} - V_t - mV_{dsat})v_d \tag{3.97}$$

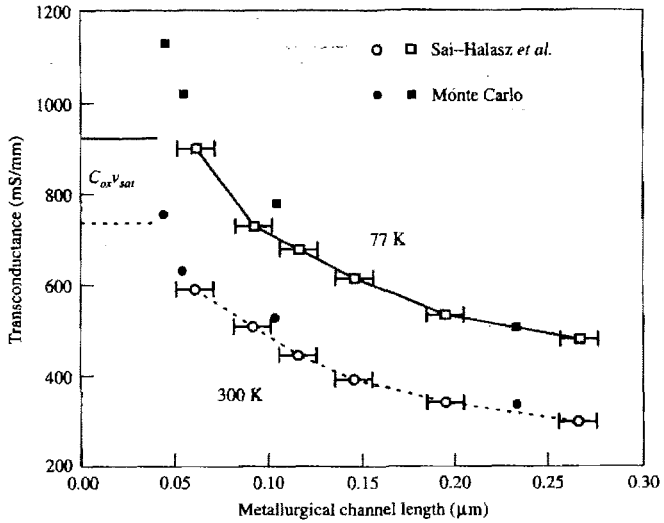


Figure 3.30. Measured (open symbols) and calculated (solid symbols) saturation transconductance versus channel length. The gate oxide is 45 Å thick. Absolute upper bounds for the transconductance in the absence of velocity overshoot at 300 and 77 K are indicated by the lines labeled $C_{ox}v_{sat}$. (After Laux and Fischetti,1998.)

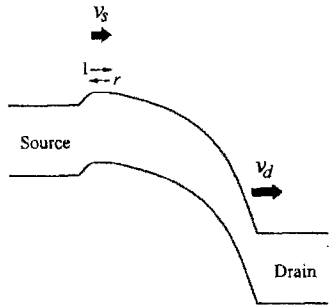


Figure 3.31. Band diagram of a MOSFET biased in saturation; v_s and v_d are the average carrier velocities near the source and the drain, respectively. Carriers near the source can be considered as made up of an incident flux (1→) and a reflected flux (←r) in the scattering theory.

at the drain equals

$$I_{ds} = WC_{ox}(V_{gs} - V_t)v_s \tag{3.98}$$

at the source where $Q_i = C_{ox}(V_{gs} - V_t)$ is high but the carrier velocity v_s is low. In this picture, MOSFET current is more directly related to the average carrier velocity

v_s at the bottleneck region near the source. Velocity overshoot near the drain helps raise MOSFET currents only to the extent that it increases V_{dsat} hence the field near the source.

In the ballistic MOSFET model discussed in Appendix 11 (Natori, 1994), the saturation current is limited by the thermal injection velocity v_T from the source. Equating I_{dsat} of Eq. (A11.11) in the one subband degenerate limit to $WC_{inv}(V_{gs}-V_t)v_T$ yields

$$v_T = \frac{2h}{3\pi m_t}\sqrt{\frac{2C_{inv}(V_{gs}-V_t)}{\pi q}}. \tag{3.99}$$

This limit is independent of the field and scattering parameters, and is not enhanced by velocity overshoot near the drain. Note that the ballistic saturation current takes the same form as the velocity-saturation-limited current, Eq. (3.81), with the parameter v_{sat} replaced by the thermal injection velocity v_T. For an electron sheet density of $C_{inv}(V_{gs}-V_t)/q \approx 10^{13}$ cm^{-2}, $v_T \approx 2\times 10^7$ cm/s, or about twice v_{sat}.

In the scattering theory (Lundstrom, 1997) for an ordinary MOSFET, carriers are injected into the channel from the source (toward the right) and from the drain (toward the left) at their respective injection velocities. They may be back scattered due to random collisions in the channel. Under high drain bias conditions, $V_{ds} >> kT/q$, carriers originated from the drain have virtually no possibility of making it uphill all the way to the source. Carriers at the point of highest barrier near the source can then be considered as made up of an incident flux (from the source) of amplitude 1 and a reflected flux of amplitude $r < 1$, as shown in Fig. 3.31. Both fluxes are moving at the thermal velocity v_T, but in opposite directions. The average carrier velocity v_s near the source is therefore

$$v_s = \frac{1-r}{1+r}v_T. \tag{3.100}$$

The drain current is obtained by substituting v_s in Eq. (3.98). The reflection coefficient r is determined by the field and scattering rates (mobility) in the low-field channel region near the source (Lundstrom, 1997). Once the carriers are a few kT below the highest energy barrier in Fig. 3.31, they are unlikely to be scattered back to the source. In a ballistic MOSFET, there is no scattering, so $r = 0$, $v_s = v_T$, and the current reaches the upper limit.

Note that r is a phenomenological parameter that cannot be predicted from the scattering theory. A Monte Carlo type of solution to the Boltzmann transport equation is needed to calculate r from the detailed physics processes. Recent experimental data suggest that $r \approx 1/3$ and $v_s \approx v_T/2$ in state-of-the-art sub-100 nm MOSFETs (Lochtefield and Antoniadis, 2001). It has been argued that r is rather insensitive to scaling because of the inevitable loss of mobility (Fig. 3.15) due to increased fields (both lateral and vertical) in the shorter device. This implies that scaling silicon MOSFETs to shorter channel lengths, e.g., 10 nm, may not result in a drain current closer to the ballistic limit.

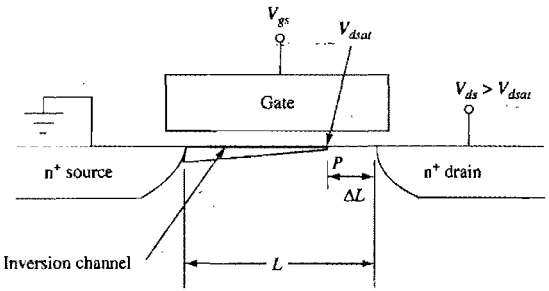


Figure 3.32. Schematic diagram showing channel length modulation when a MOSFET is biased beyond saturation. The surface channel collapses at point P, where carriers reach saturation velocity.

3.2.3 Channel Length Modulation

In this subsection, we discuss the characteristics of short-channel MOSFETs biased beyond saturation. In a long-channel device, the drain current stays constant when the drain voltage exceeds V_{dsat}, as shown in Fig. 3.10. The output conductance, dI_{ds}/dV_{ds}, is zero in the saturation region. In contrast, the drain current of a short-channel MOSFET can still increase slightly beyond the pinch-off or the velocity saturation point with a nonzero output conductance, as is evident from the experimental curves in Fig. 3.27. This arises because of two factors: the short-channel effect and channel length modulation. The short-channel effect was discussed in Section 3.2.1; when the drain voltage increases beyond saturation in a short-channel device, the threshold voltage decreases, and therefore the drain current increases. In this subsection, we describe channel length modulation.

According to the one-dimensional model in the preceding subsection, the electric field along the channel approaches infinity at the saturation point. In practice, the field remains finite. However, its magnitude becomes comparable to the vertical field, so that the gradual-channel approximation breaks down and carriers are no longer confined to the surface channel. ***As the drain voltage increases beyond the saturation voltage V_{dsat}, the saturation point where the surface channel collapses begins to move slightly toward the source***, as shown in Fig. 3.32. The voltage at the saturation point remains constant at V_{dsat}, independent of V_{ds}. The voltage difference $V_{ds}-V_{dsat}$ is dropped across the region between the saturation point and the drain. Carriers injected from the surface channel into this region travel at saturation velocity until collected by the drain junction. The distance between the saturation point and the drain, ΔL, is referred to as the amount of channel length modulation by the drain voltage. Since the one-dimensional model is still valid between the source and the saturation point where the voltage remains at V_{dsat}, the device acts as if its channel length were shortened by ΔL. The drain current is then obtained simply by replacing L with $L-\Delta L$ in Eq. (3.79). In the long-channel limit, this increases the drain current by a factor of $(1-\Delta L/L)^{-1}$, i.e.,

$$I_{ds} = \frac{I_{dsat}}{1-(\Delta L/L)}. \tag{3.101}$$

Since ΔL increases with increasing drain voltage, the drain current continues to increase in the saturation region. A 2-D device simulator is needed to numerically evaluate ΔL for a given set of device parameters and bias conditions.

3.2.4 Source–Drain Series Resistance

In the discussion of MOSFET current thus far, it was assumed that the source and drain regions were perfectly conducting. In reality, as the current flows from the channel to the terminal contact, there is a small voltage drop in the source and drain regions due to the finite silicon resistivity and metal contact resistance. In a long-channel device, the source–drain parasitic resistance is negligible compared with the channel resistance. In a short-channel device, however, the source–drain series resistance can be an appreciable fraction of the channel resistance and can therefore cause significant current degradation.

The most severe current degradation by series resistance occurs in the linear region (low V_{ds}) when the gate voltage is high. This is because the MOSFET channel resistance,

$$R_{ch} = \frac{V_{ds}}{I_{ds}} = \frac{L}{\mu_{eff} C_{inv} W (V_{gs} - V_{on})} \tag{3.102}$$

based on Eq. (3.64) and the discussion below it, is the lowest under such bias conditions. It is instructive to estimate the sheet resistivity of a MOSFET channel,

$$\rho_{ch} \equiv R_{ch} \frac{W}{L} = \frac{1}{\mu_{eff} (\varepsilon_{ox}/t_{inv})(V_{gs} - V_{on})} \approx \frac{1}{\mu_{eff} \varepsilon_{ox} \mathscr{E}_{ox}}. \tag{3.103}$$

Since the maximum oxide field $\mathscr{E}_{ox}$ is typically 2–5 MV/cm for most VLSI technologies, the minimum channel sheet resistivity is about 2000 Ω/□ for nMOSFETs and 7000 Ω/□ for pMOSFETs.

The MOSFET current in the saturation region is least affected by the resistance degradation of source–drain voltage, since I_{ds} is essentially independent of V_{ds} in saturation. The saturation current is only affected through gate-voltage degradation by the voltage drop between the source contact and the source end of the channel (Fig. 4.23).

The effect of series resistance on linear I_{ds}–V_{gs} curves used in channel-length extraction will be addressed in Section 4.3. Various contributions to the source–drain series resistance and their effect on circuit performance will be discussed in detail in Chapter 5.

3.2.5 MOSFET Degradation and Breakdown at High Fields

Besides causing dielectric breakdown, the high electric fields in a MOSFET can also cause degradation of the device characteristics. *Hot-carrier effects* (HCE) and *negative-bias-temperature instability* (NBTI) are two of the most important degradation phenomena in modern CMOS devices. Here we describe briefly the phenomena and the physical mechanisms involved.

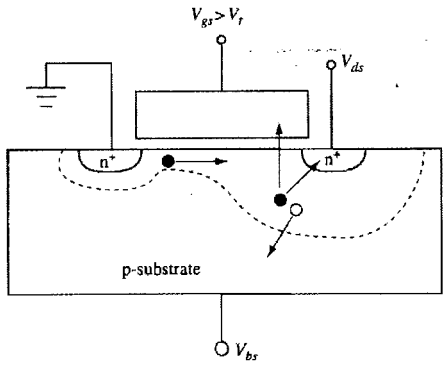


Figure 3.33. Schematic illustrating the physical processes that give rise to channel hot-electron effect in an n-channel MOSFET. The dotted line indicates the boundary of the space-charge region.

3.2.5.1 Hot-Carrier Effects

Consider an n-channel MOSFET with $V_{gs} > V_t$ applied to the gate and V_{ds} applied to the drain. A high-field space-charge region is established in the silicon near the drain, as illustrated in Fig. 3.33. As the electrons drift towards the drain, they gain energy from the electric field in the space-charge region and become hot. The hot electrons can cause impact ionization near the drain, or they can be injected into the gate insulator (see Section 2.5.4). The secondary holes from impact ionization contribute to a substrate current (Abbas, 1974). Substrate currents at drain voltages less than the silicon bandgap voltage E_g/q have been observed, suggesting that some electrons can gain additional energy from electron–electron and/or electron–phonon collisions (Chung *et al.*, 1990).

Figure 3.34 is a typical plot of the channel current and substrate current as a function of the gate voltage, in this case for an n-channel MOSFET with 0.25 μm channel length having a threshold voltage, V_t, of about 0.4 V (Chang *et al.*, 1992). Notice that the substrate current increases with the gate voltage in the subthreshold region, peaking at gate voltages between about V_t and $2V_t$, and then decreases with further increases in the gate voltage. This complex dependence on gate voltage can be understood as follows. The electrons available for initiating impact ionization, to first order, come from the drain current. At $V_{gs} < V_t$, there is no surface inversion channel and the maximum electric field in the silicon near the drain end is relatively independent of the gate voltage. As a result, the substrate current is roughly proportional to the drain current, as can be seen in Fig. 3.34. At $V_{gs} > V_t$, a surface inversion channel is formed. As discussed in Sections 3.1.2.3 and 3.1.2.4, for V_{gs} small compared with V_{ds}, the surface channel is pinched off near the drain end. [Pinch off occurs for $V_{ds} > V_{dsat} = (V_{gs} - V_t)/m$, where m is given by Eq. (3.27).] When the surface inversion channel is pinched off, there is a voltage drop V_{dsat} (drain saturation voltage) along the inversion channel between the source and the pinch-off point, and a voltage drop $V_{ds} - V_{dsat}$ in the space-charge region between the pinch-off point and the drain. That is, the maximum electric field in the silicon

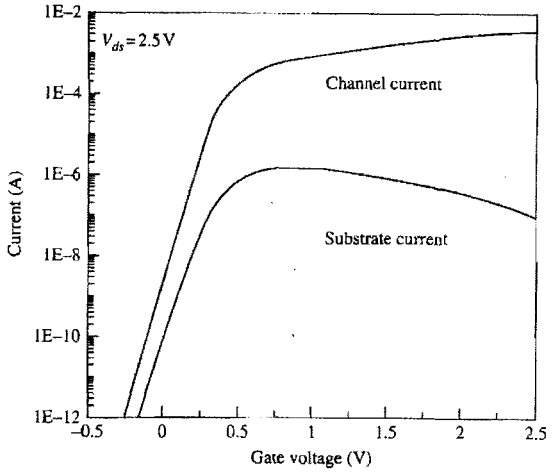


Figure 3.34. Typical plots of the channel current and substrate current of a MOSFET. The example shown here is for an n-channel FET having 0.25 μm channel length and 10 μm channel width. (After Chang *et al.*, 1992.)

space-charge region is determined by $V_{ds} - V_{dsat}$. ***For a given V_{ds}, as V_{gs} increases, so does V_{dsat}, which means the voltage drop $V_{ds} - V_{dsat}$ decreases hence the maximum electric field in the silicon decreases.*** As shown in Fig. 2.59, the rate of impact ionization decreases rapidly with decrease in electric field. The net result is that the substrate current decreases with increasing gate voltage for V_{gs} larger than about $2V_t$, as seen in Fig. 3.34. The reader interested in detailed models for the substrate current in a MOSFET is referred to the literature (e.g., Hu *et al.*, 1985, and Kolhatkar and Dutta, 2000).

The hot electrons injected into the gate insulator near the drain region contribute to a gate current. The gate current can cause bulk and interface traps to be generated (see Section 2.5.3.3), and some of the injected electrons can become trapped in the gate insulator near the drain. The trapped electrons and the interface states cause the surface potential near the drain to shift. Since the source of the hot electrons is the channel current, this device degradation is referred to as *channel hot electron (CHE) effect*. The turn on characteristics of a MOSFET are determined primarily by the surface potential near the source end of the channel (see Section 3.1.2). With the damage localized near the drain junction, a MOSFET that has suffered significant CHE damage shows a larger damage-induced threshold voltage shift when it is operated in the source–drain-reversed mode than in the normal mode (Abbas and Dockerty, 1975; Ning *et al.*, 1977b). In the case of a p-channel MOSFET, the same device degradation is referred to as *channel hot hole effect*.

In certain circuit configurations, the terminal voltages of a MOSFET are such that V_{ds} is small or zero, while both the source and the drain are reverse biased with respect to the substrate, i.e., $V_{bs} < 0$. (An example of this bias configuration is when a MOSFET is used as a pass-gate.) In this case, the channel electrons do not gain much energy traveling

from source to drain. However, the minority electrons from the p-substrate can gain energy as they drift towards the silicon–oxide interface. These hot electrons can be injected into the gate insulator as depicted in Fig. 2.68. The injected electrons can generate bulk and interface traps or become trapped in the oxide layer. Since the hot electrons come from the substrate, the associated device degradation is referred to as *substrate hot electron (SHE) effect*. By symmetry, SHE damage to the oxide is spread more or less uniformly over the entire device channel region. Also, since the minority electron density in the substrate increases with temperature, SHE effect increases with temperature (Ning *et al.*, 1976).

In the case of a p-channel MOSFET, instead of hot electrons, we have hot holes causing device degradation. Hot-carrier degradation is one of the major effects limiting the voltage that can be applied to CMOS devices. In general, the approach to minimize hot-carrier damage is to (a) reduce the peak electric field in the silicon to reduce the energy of hot carriers, and (b) minimize the trap density in the gate insulator. Designers often employ the so-called *lightly doped drain* (LDD) design to suppress CHE effect (Ogura *et al.*, 1982). In an LDD design, the drain region adjacent to the channel has a lower doping concentration than the drain region away from the channel. This laterally graded drain doping profile reduces the peak electric field near the drain. As for the substrate hot electron or substrate hot hole effects, they can be suppressed quite effectively by using a relatively lightly doped substrate (Ning *et al.*, 1979). However, both the trap density of the gate insulator and the susceptibility of the gate insulator to hot-carrier damage depend on the growth process of the gate insulator, as well as on the subsequent processes used to complete the integrated-circuit chip fabrication. As a result, hot-carrier effects cannot be predicted sufficiently accurately in advance. Instead, for each CMOS technology generation, the effects are characterized, modeled and then included in the design of circuits. The reader is referred to the literature on the subject for more details (see, e.g., Takeda *et al.*, 1995).

3.2.5.2 Negative-Bias-Temperature Instability

It was reported by Deal *et al.* (1967) that a negative voltage applied to the gate of an MOS capacitor at elevated temperatures can cause both a build up of positive charge in the oxide and an increase in the density of surface states (see Fig. 2.52). To turn on a p-channel MOSFET, the gate electrode is biased negatively with respect to the n-type body. Also, in many applications, the device temperature can be rather high, often close to 100 °C. Thus, NBTI can occur naturally in the operation of a p-channel MOSFET, causing the gate voltage needed to turn on the transistor to increase (becoming more negative with respect to the n-type body) with time. Many papers have been published on the various aspects of NBTI. It is found that device degradation due to NBTI is a function of the gate insulator process, and the degradation worsens with both the temperature and the oxide field (Jeppson and Svensson, 1977; Blat *et al.*, 1991). For short time periods, the degradation has approximately a $t^{1/4}$ time dependence (Jeppson and Svensson, 1977). It then saturates to a value depending on the oxide field and the temperature (Zafar *et al.*, 2004).

If a p-channel MOSFET is stressed with zero or relatively small voltages between the source and drain, there is only NBTI effect and the channel hot hole effect is negligible. In

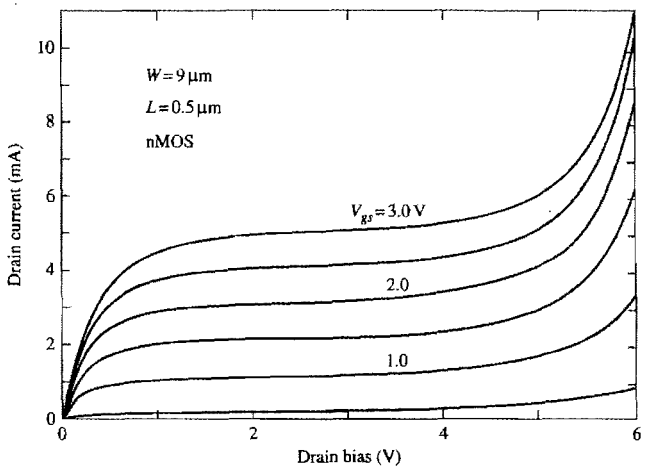


Figure 3.35. Example $I_{ds} - V_{ds}$ curves of a short-channel nMOSFET showing breakdown at high drain voltages. (After Sun *et al.*, 1987.)

this stress mode, the device degradation or NBTI by itself is relatively insensitive to channel length. However, a pMOSFET in the off state in a CMOS circuit typically has a relatively large voltage between its source and drain. Therefore, the device degradation experienced by a short-channel pMOSFET is often caused by a combination of channel hot hole effect and NBTI (La Rosa *et al.*, 1997). NBTI must be characterized for each technology so its effect can be included in the design of circuits, particularly for CMOS circuits that depend on good matching of the threshold voltage of p-channel MOSFETs (Rauch III, 2002). An excellent review of the current understanding of the physical mechanisms of NBTI in modern CMOS devices has been given by Schroder and Babcock (Schroder and Babcock, 2003).

3.2.5.3 MOSFET Breakdown

Breakdown occurs in a short-channel MOSFET when the drain voltage exceeds a certain value, as shown in Fig. 3.35. At high drain voltages, the peak electric field in the saturation region can attain large values. When the field exceeds mid-10^5 V/cm, impact ionization (Section 2.5.1) takes place at the drain, leading to an abrupt increase of drain current. The breakdown voltage of nMOSFETs is usually lower than that of pMOSFETs because electrons have a higher rate of impact ionization (Fig. 2.59) and because n^+ source and drain junctions are more abrupt than p^+ junctions. There is also a weak dependence of the breakdown voltage on channel length; shorter devices have a lower breakdown voltage.

The breakdown process in an nMOSFET is shown schematically in Fig. 3.36. Electrons gain energy from the field as they move down the channel. Before they lose energy through collisions, they possess high kinetic energy and are capable of generating secondary electrons and holes by impact ionization. The generated electrons are attracted

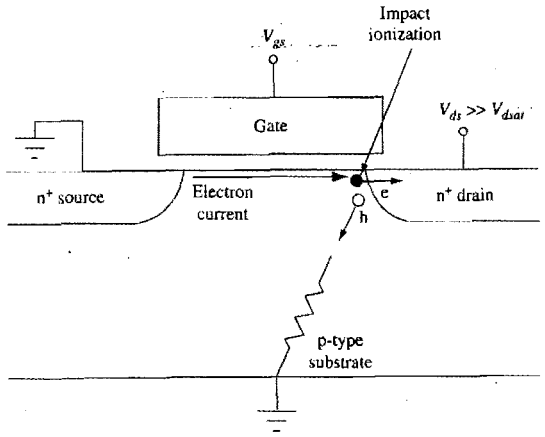


Figure 3.36. Schematic diagram showing impact ionization at the drain.

to the drain, adding to the drain current, while the holes are collected by the substrate contact, resulting in a substrate current. The substrate current in turn can produce a voltage (*IR*) drop from the spreading resistance in the bulk, which tends to forward-bias the source junction. This lowers the threshold voltage of the MOSFET and triggers a positive feedback effect, which further enhances the channel current. Substrate current (Fig. 3.34) is usually a good indicator of hot carriers generated by low-level impact ionization before runaway breakdown occurs.

Breakdown often results in permanent damage to the MOSFET as large amounts of hot carriers are injected into the oxide in the gate-to-drain overlap region. MOSFET breakdown is particularly a problem for VLSI technology during the elevated-voltage burn-in process. It can be relieved to some extent by using a lightly doped drain (LDD) structure (Ogura *et al.*, 1982), which introduces additional series resistance and reduces the peak field in a MOSFET. However, drain current and therefore device performance are traded off as a result. Ultimately, the devices should operate at a power-supply voltage far enough below the breakdown condition. This is one of the key CMOS design considerations in Chapter 4.

Exercises

3.1 Consider an n-channel MOSFET with 20 nm thick gate oxide and uniform p-type substrate doping of 10^{17} cm^{-3}. The gate work function is that of n^+ Si.
   (a) What is the threshold voltage? Sketch the band diagram at threshold condition, $\psi_s = 2\psi_B$.
   (b) What is the threshold voltage if a reverse bias of 1 V is applied to the substrate? Sketch the band diagram at threshold.
   (c) What is the scale length of this device and how short can the channel length be reduced to before severe short-channel effect takes place?

3.2 Fill in the steps that lead to Eq. (3.34), the fraction of drift current component in the limit of $V \to 0$.

3.3 The effective field $\mathscr{E}_{eff}$ plays an important role in MOSFET channel mobility. Show that the definition

$$\mathscr{E}_{eff} \equiv \frac{\int_0^{x_i} n(x)\mathscr{E}(x)dx}{\int_0^{x_i} n(x)dx}$$

leads to Eq (3.51), i.e., $\mathscr{E}_{eff} = (|Q_d| + |Q_i|/2)/\varepsilon_{si}$. Note that

$$|Q_i| = q\int_0^{x_i} n(x)\,dx$$

and

$$\mathscr{E}(x) = \frac{1}{\varepsilon_{si}}\left(|Q_d| + q\int_x^{x_i} n(x')\,dx'\right)$$

from Gauss's law. The inversion-layer depth x_i is assumed to be much smaller than the bulk depletion width.

3.4 An alternative threshold definition is based on the rate of change of inversion charge density with gate voltage. Equation (3.61) from Fig. 2.37(b) states that $d|Q_i|/dV_{gs}$ is given by the serial combination of C_{ox} and $C_i \equiv d|Q_i|/d\psi_s$. Below threshold, $C_i \ll C_{ox}$, so that $d|Q_i|/dV_{gs} \approx C_i$ and Q_i increases exponentially with V_{gs}. Above threshold, $C_i \gg C_{ox}$, so that $d|Q_i|/dV_{gs} \approx C_{ox}$ and Q_i increases linearly with V_{gs}. The change of behavior occurs at an *inversion charge threshold voltage*, V_t^{inv}, where $C_i = C_{ox}$. Show that at $V_{gs} = V_t^{inv}$ one has $d|Q_i|/dV_{gs} = C_{ox}/2$ and $Q_i \approx (2kT/q)C_{ox}$. Note that such an inversion charge threshold is independent of depletion charge and is slightly higher than the conventional $2\psi_B$ threshold.

3.5 From Eq. (3.63) (neglecting the second term from inversion charge capacitance), show that the fractional loss of inversion charge due to the polysilicon depletion effect is $\Delta Q_i/Q_i \approx C_{ox}/2C_p$ where C_p is the small-signal polysilicon-depletion capacitance defined in Eq. (2.212). Explain why there is a factor-of-two difference between the loss of charge and the loss of capacitance.

3.6 For an nMOSFET with $t_{ox} = 10$ nm, $\mu_{eff} = 500$ cm^2/V-s, $v_{sat} = 10^7$ cm/s, $W = 10\,\mu$m, and $L = 1\,\mu$m, assume $m = 1$.

(a) Use the $n = 1$ velocity saturation model to generate I_{ds} versus V_{ds} (0–5 V) curves for $V_{gs} - V_t = 1$, 2, 3, 4, and 5 V. (Note: $I_{ds} = I_{dsat}$ beyond V_{dsat}.)

(b) Now let L vary from 0.5 μm to 5 μm. Calculate and plot the saturation current for $V_{gs} - V_t = 3$ V vs. L. Compare it with the long-channel saturation current (without velocity saturation) for the same $V_{gs} - V_t$ and range of L.

3.7 The small-signal transconductance in the saturation region is defined as $g_{msat} \equiv dI_{dsat}/dV_{gs}$. Derive an expression for g_{msat} using Eq. (3.79) based on the $n = 1$ velocity saturation model. Show that g_{msat} approaches the saturation-velocity-limited value, Eq. (3.96), when $L \to 0$. What becomes of the expression for g_{msat} in the long-channel limit when $v_{sat} \to \infty$?

3.8 From Eq. (3.79) based on the $n = 1$ velocity saturation model, what is the carrier velocity at the source end of the channel? What are the limiting values when $L \to 0$ and when $v_{sat} \to \infty$?

3.9 Following a similar approach as in the text for the $n = 1$ velocity saturation model, derive an integral equation for the $n = 2$ velocity saturation model from which I_{ds} can be solved. It is very tedious to carry out the integration analytically (Taylor, 1984). Interested readers may attempt performing it numerically on a computer.

3.10 Assuming the $n = 1$ velocity saturation model, show that the total integrated inversion charge under the gate is

$$Q_i(\text{total}) = -WLC_{ox}(V_{gs} - V_t)\frac{\sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)} + \frac{1}{3}}{\sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(mv_{sat}L)} + 1}$$

in the saturation region. Evaluate the intrinsic gate-to-channel capacitance, and show that it approaches Eq. (3.60) in the long-channel limit.

3.11 The generalized MOSFET scale length is given by Eq. (3.70). For $\varepsilon_{si} = 11.7\varepsilon_0$, $\varepsilon_i = 7.8\,\varepsilon_0$, $t_i = 5.0$ nm, and $W_{dm} = 10.0$ nm, find the three longest eigenvalues λ_1, λ_2, λ_3. Take $L \approx 2\lambda_1$; what's the ratio between $\exp(-\pi L/2\lambda_1)$ and $\exp(-\pi L/2\lambda_2)$?

4 CMOS Device Design

This chapter examines the key device design issues in a modern CMOS VLSI technology. It begins with an extensive review of the concept of MOSFET scaling. Two important CMOS device design parameters, threshold voltage and channel length, are then discussed in detail.

4.1 MOSFET Scaling

CMOS technology evolution in the past thirty years has followed the path of device scaling for achieving density, speed, and power improvements. MOSFET scaling was propelled by the rapid advancement of lithographic techniques for delineating fine lines of 1 μm width and below. In Section 3.2.1, we discussed that reducing the source-to-drain spacing, i.e., the channel length of a MOSFET, led to short-channel effects. For digital applications, the most undesirable short-channel effect is a reduction in the gate threshold voltage at which the device turns on, especially at high drain voltages. Full realization of the benefits of the new high-resolution lithographic techniques therefore requires the development of new device designs, technologies, and structures which can be optimized to keep short-channel effects under control at very small dimensions. Another necessary technological advancement for device scaling is in ion implantation, which not only allows the formation of very shallow source and drain regions but also is capable of accurately introducing a sharply profiled, low concentration of doping atoms for optimum channel profile design.

4.1.1 Constant-Field Scaling

In constant-field scaling (Dennard *et al.*, 1974), it was proposed that one can keep short-channel effects under control by scaling down the vertical dimensions (gate insulator thickness, junction depth, etc.) along with the horizontal dimensions, while also proportionally decreasing the applied voltages and increasing the substrate

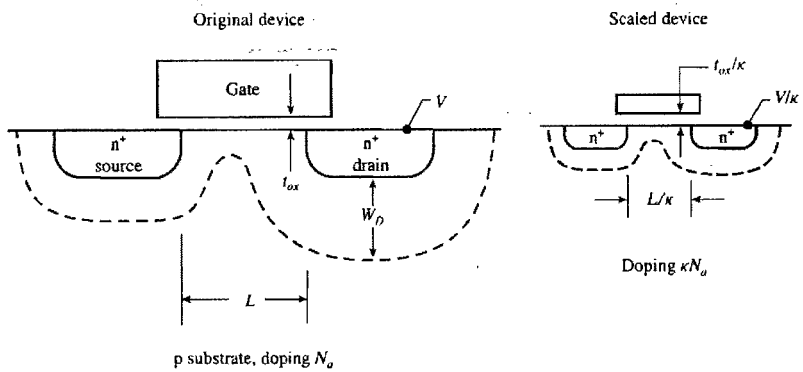


Figure 4.1. Principles of MOSFET constant-electric-field scaling. (After Dennard, 1986.)

doping concentration (decreasing the depletion width). This is shown schematically in Fig. 4.1. ***The principle of constant-field scaling lies in scaling the device voltages and the device dimensions (both horizontal and vertical) by the same factor, κ (>1), so that the electric field remains unchanged.*** This assures that the reliability of the scaled device is not worse than that of the original device.

4.1.1.1 Rules for Constant-Field Scaling

Table 4.1 shows the scaling rules for various device parameters and circuit performance factors. The doping concentration must be increased by the scaling factor κ in order to keep Poisson's equation (3.66) invariant with respect to scaling. The maximum drain depletion width,

$$W_D = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} + V_{dd})}{qN_a}}, \tag{4.1}$$

from Eq. (2.85) (with $V_{app} = -V_{dd}$) scales down approximately by κ provided that the power-supply voltage V_{dd} is much greater than the built-in potential ψ_{bi}. All capacitances (including wiring load) scale down by κ, since they are proportional to area and inversely proportional to thickness. The charge per device ($\sim C \times V$) scales down by κ^2, while the inversion-layer charge density (per unit gate area), Q_i, remains unchanged after scaling. Since the electric field at any given point is unchanged, the carrier velocity ($v = \mu \mathscr{E}$) at any given point is also unchanged (the mobility is the same for the same vertical field). Therefore, any velocity saturation effects will be similar in the original and the scaled devices.

The drift current per MOSFET width, obtained by integrating the first term of the electron current density equation (2.54) over the inversion layer thickness, is

Table 4.1 Scaling MOSFET device and circuit parameters

	MOSFET Device and Circuit Parameters	Multiplicative Factor ($\kappa > 1$)
Scaling assumptions	Device dimensions (t_{ox}, L, W, x_j)	$1/\kappa$
	Doping concentration (N_a, N_d)	κ
	Voltage (V)	$1/\kappa$
Derived scaling behavior of device parameters	Electric field ($\mathscr{E}$)	1
	Carrier velocity (v)	1
	Depletion-layer width (W_d)	$1/\kappa$
	Capacitance ($C = \varepsilon A/t$)	$1/\kappa$
	Inversion-layer charge density (Q_i)	1
	Current, drift (I)	$1/\kappa$
	Channel resistance (R_{ch})	1
Dervied scaling behavior of circuit parameters	Circuit delay time ($\tau \sim CV/I$)	$1/\kappa$
	Power dissipation per circuit ($P \sim VI$)	$1/\kappa^2$
	Power-delay product per circuit ($P\tau$)	$1/\kappa^3$
	Circuit density ($\propto 1/A$)	κ^2
	Power density (P/A)	1

$$\frac{I_{drift}}{W} = Q_i v = Q_i \mu \mathscr{E}, \tag{4.2}$$

and is unchanged with respect to scaling. This means that the drift current scales down by κ, consistent with the behavior of both the linear and the saturation MOSFET currents in Eq. (3.23) and Eq. (3.28). A key implicit assumption is that the threshold voltage also scales down by κ. Note that the velocity saturated current, Eq. (3.79), also scales the same way, since both v_{sat} and μ_{eff} are constants, independent of scaling. However, the diffusion current per unit MOSFET width, obtained by integrating the second term of the current density equation (2.54) and given by

$$\frac{I_{diff}}{W} = D_n \frac{dQ_i}{dx} = \mu_n \frac{kT}{q} \frac{dQ_i}{dx}, \tag{4.3}$$

scales up by κ, since dQ_i/dx is inversely proportional to the channel length. Therefore, the diffusion current does not scale down the same way as the drift current. This has significant implications in the nonscaling of MOSFET subthreshold currents, as will be discussed in Section 4.1.3.

4.1.1.2 Effect of Scaling on Circuit Parameters

With both the voltage and the current scaled down by the same factor, it follows that the active channel resistance [e.g., Eq. (3.102)] of the scaled-down device remains unchanged. It is further assumed that the parasitic resistance is either negligible or unchanged in scaling. The circuit delay, which is proportional to RC or CV/I, then scales

down by κ. ***This is the most important conclusion of constant-field scaling: once the device dimensions and the power-supply voltage are scaled down, the circuit speeds up by the same factor. Moreover, power dissipation per circuit, which is proportional to VI, is reduced by κ^2.*** Since the circuit density has increased by κ^2, the power density, i.e., the active power per chip area, remains unchanged in the scaled-down device. This has important technological implications in that, in contrast to bipolar devices (Chapters 6, 7, and 8), packaging of the scaled CMOS devices does not require more elaborate heat-sinking. The power–delay product of the scaled CMOS circuit shows a dramatic improvement by a factor of κ^3 (Table 4.1).

4.1.1.3 Threshold Voltage

It was assumed earlier that the threshold voltage should be decreased by the scaling factor, κ, in proportion to the power-supply voltage. This is examined using the threshold equation (3.44) for a uniformly doped substrate:

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B - V_{bs})}}{C_{ox}} \tag{4.4}$$

where V_{bs} is the substrate bias voltage. In silicon technology, the material-related parameters (energy gap, work function, etc.) do not change with scaling; hence, in general, V_t does not scale. However, in a conventional process, n^+-polysilicon gates are used for n-channel MOSFETs, and $V_{fb} = -E_g/2q - \psi_B$ from Eq. (2.209). It turns out that the first two terms on the RHS of Eq. (4.4) add up to approximately −0.15 V, which can be neglected. One can then argue (Dennard *et al.*, 1974) that by adjusting V_{bs} so that $2\psi_B - V_{bs}$ scales down by κ, the last term of Eq. (4.4) , and therefore V_t, will also scale down by κ. However, at very early stages of the technology development, V_{bs} has been reduced to zero for most logic applications, though a reverse-biased body–source junction is still used for some dynamic memory array devices. Further reduction of the $2\psi_B - V_{bs}$ term with scaling would require a forward bias on the substrate. This is not commonly used in VLSI technologies, although it has been attempted in experimental devices (Sai-Halasz *et al.*, 1990). In practice, nonuniform doping profiles have been employed to tailor the threshold voltage of scaled devices, as will be discussed in Section 4.2.

p-channel MOSFETs with p^+-polysilicon gates scale similarly to their counterparts. However, in buried-channel devices, e.g., when an n^+-polysilicon gate is used for p-channel MOSFETs, the sum of the first two terms in Eq. (4.4) is nearly 1 V (magnitude) and therefore cannot be neglected. For this reason, it is difficult to scale buried-channel devices to low threshold voltages. More about threshold voltage design can be found in Section 4.2.

4.1.2 Generalized Scaling

Even though constant-field scaling provides a basic guideline to the design of scaled MOSFETs, the requirement of reducing the voltage by the same factor as the device

Table 4.2 CMOS VLSI technology generations

Feature size (μm)	Power-supply voltage (V)	Gate oxide thickness (Å)	Oxide field (MV/cm)
2	5	350	1.4
1.2	5	250	2.0
0.8	5	180	2.8
0.5	3.3	120	2.8
0.35	3.3	100	3.3
0.25	2.5	70	3.6
0.1	1.5	30	5.0

physical dimension is too restrictive. Because of subthreshold nonscaling and reluctance to depart from the standardized voltage levels of the previous generation, the power-supply voltage was seldom scaled in proportion to channel length. Table 4.2 lists the supply voltage and device parameters of several generations of CMOS VLSI technology. It is clear that the oxide field has been increasing over the generations rather than staying constant. For device design purposes, therefore, it is necessary to develop a more general set of guidelines that allows the electric field to increase. In such a generalized scaling (Baccarani *et al.*, 1984), it is desired that both the vertical and the lateral electric fields change by the same multiplication factor so that the shape of the electric field pattern is preserved. This assures that 2-D effects, such as short-channel effects, do not become worse when scaling to a smaller dimension. Higher fields, however, do cause reliability concerns as mentioned in Section 2.5. Below 0.1 μm feature size, CMOS design space is severely constrained by power issues. Details of the performance-power tradeoff are discussed in Section 4.2.2.2 with the key parameters summarized in Fig. 4.7.

4.1.2.1 Rules for Generalized Scaling

If we assume that the electric field intensity changes by a factor of α, i.e., $\mathscr{E} \rightarrow \alpha\,\mathscr{E}$, while the device physical dimensions (both lateral and vertical) scale down by κ (>1) in generalized scaling, the potential or voltage will change by a factor equal to the ratio α/κ. If $\alpha = 1$, it reduces back to constant-field scaling. To keep Poisson's equation invariant under the transformation, $(x, y) \rightarrow (x, y)/\kappa$ and $\psi \rightarrow (\alpha/\kappa)\psi$ within the depletion region,

$$\frac{\partial^2(\alpha\psi/\kappa)}{\partial(x/\kappa)^2} + \frac{\partial^2(\alpha\psi/\kappa)}{\partial(y/\kappa)^2} = \frac{qN_a'}{\varepsilon_{si}}, \tag{4.5}$$

N_a' should be scaled to $(\alpha\,\kappa)N_a$. In other words, the doping concentration must be scaled up by an extra factor of α to control the depletion-region depth and thus avoid increased short-channel effects due to the higher electric field. Table 4.3 shows the generalized scaling rules of other device and circuit parameters.

Since the electric field intensity is usually increased in generalized scaling, the carrier velocity tends to increase as well. How much the velocity increases depends on how velocity-saturated the original device is. In the long-channel limit, carrier velocities are far from saturation and will increase by the same factor, α, as the electric field. The drift

Table 4.3 Generalized MOSFET scaling

	MOSFET device and circuit parameters	Multiplicative factor ($\kappa > 1$)	
Scaling assumptions	Device dimensions (t_{ox}, L, W, x_j)	$1/\kappa$	
	Doping concentration (N_a, N_d)	$\alpha\,\kappa$	
	Voltage (V)	α/κ	
Derived scaling behavior of device parameters	Electric field ($\mathscr{E}$)	α	
	Depletion-layer width (W_d)	$1/\kappa$	
	Capacitance ($C = \varepsilon A/t$)	$1/\kappa$	
	Inversion-layer charge density (Q_i)	α	
		Long Ch.	Vel. Sat.
	Carrier velocity (v)	α	1
	Current, drift (I)	α^2/κ	α/κ
Derived scaling behavior of circuit parameters	Circuit delay time ($\tau \sim CV/I$)	$1/\alpha\,\kappa$	$1/\kappa$
	Power dissipation per circuit ($P \sim VI$)	α^3/κ^2	α^2/κ^2
	Power-delay product per circuit ($P\,\tau$)	α^2/κ^3	
	Circuit density ($\propto 1/A$)	κ^2	
	Power density (P/A)	α^3	α^2

current, which is proportional to WQ_iv, will then change by a factor of α^2/κ. This is consistent with the scaling behavior of long-channel currents, Eq. (3.23) and Eq. (3.28). On the other hand, if the original device is fully velocity-saturated, the carrier velocity cannot increase any more, in spite of the higher field in the scaled device. The current in this case will change only by a factor of α/κ, consistent with the velocity-saturated current, Eq. (3.81). The circuit delay scales down by a factor between κ and $\alpha\kappa$, depending on the degree of velocity saturation. The most serious issue with generalized scaling is the increase of the power density by a factor of α^2 to α^3. This puts a great burden on VLSI packaging technology to dissipate the extra heat generated on the chip. The power–delay product is also a factor of α^2 higher than in constant-field scaling.

4.1.2.2 Constant-Voltage Scaling

Even though Poisson's equation within the depletion region is invariant under generalized scaling, the same is not true in the inversion layer when mobile charges are present. This is because mobile charge densities are exponential functions of potential which do not scale linearly with either physical dimensions or voltage. Furthermore, even in the depletion region, not all the boundary conditions scale consistently under generalized scaling. ***This is due to the fact that the band bending at the source junction is given by the built-in potential (Appendix 9), which does not scale with voltage. Strictly speaking, the shape of the field pattern is preserved only if $\alpha = \kappa$, i.e., constant-voltage scaling.*** Under constant-voltage scaling, the electric field scales up by κ and the doping concentration N_a scales up by κ^2. The maximum gate depletion width (long-channel) [Eq. (3.68)],

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a/n_i)}{q^2N_a}}, \tag{4.6}$$

then scales down by κ. Here $\ln(N_a/n_i)$ is a weak function of N_a and can be treated as a constant. This allows the short-channel V_t rolloff [Eq. (3.67)],

$$\Delta V_t = \frac{24 t_{ox}}{W_{dm}} \left[\sqrt{\psi_{bi}(\psi_{bi} + V_{ds})} - a(2\psi_B) \right] e^{-\frac{\pi L/2}{W_{dm} + 3t_{ox}}}, \tag{4.7}$$

to remain unchanged, as both t_{ox} and W_{dm} are scaled down by the same factor as the channel length L. Both the power-supply voltage and the threshold voltage [Eq. (4.4)],

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B - V_{bs})}}{C_{ox}}, \tag{4.8}$$

also remain unchanged. From Eq. (2.194), the inversion-layer charge per unit area is related to the electron concentration at the silicon surface, $n(0)$, by

$$Q_i = \sqrt{2\varepsilon_{si} kT n(0)}. \tag{4.9}$$

Since Q_i scales up by κ in constant-voltage scaling, $n(0)$ scales up by κ^2. Therefore, the mobile charge density scales the same way as the fixed charge density N_a. The inversion-layer thickness, being proportional to $Q_i/qn(0)$, scales down by κ just like other linear dimensions. The Debye length, $L_D = (\varepsilon_{si} kT/q^2 N_a)^{1/2}$, also scales down by κ under constant-voltage scaling.

Although constant-voltage scaling leaves the solution of Poisson's equation for the electrostatic potential unchanged except for a constant multiplicative factor in the electric field, it cannot be practiced without limit, since the power density increases by a factor of κ^2 to κ^3. Higher fields also cause hot-electron and oxide reliability problems. In reality, CMOS technology evolution has followed mixed steps of constant-voltage and constant-field scaling, as is evident in Table 4.2.

4.1.3 Nonscaling Effects

4.1.3.1 Primary Nonscaling Factors

From the above discussions, it is clear that although constant-field scaling provides a basic framework for shrinking CMOS devices to gain higher density and speed without degrading reliability and power, there are several factors that scale neither with the physical dimensions nor with the operating voltage. ***The primary reason for the nonscaling effects is that neither the thermal voltage kT/q nor the silicon bandgap E_g changes with scaling***. The former leads to subthreshold nonscaling; i.e., the threshold voltage cannot be scaled down like other parameters. The latter leads to nonscalability of the built-in potential, depletion-layer width, and short-channel effect.

From Eq. (3.40) , the *off current* of a MOSFET is given by

$$I_{ds}(V_{gs} = 0,\ V_{ds} = V_{dd}) = \mu_{eff} C_{ox} \frac{W}{L} (m-1) \left(\frac{kT}{q} \right)^2 e^{-qV_t/mkT}. \tag{4.10}$$

Because of the exponential dependence, the threshold voltage cannot be scaled down significantly without causing a substantial increase in the off current. In fact, even if the threshold voltage is held unchanged, the off current per device still increases by a factor of κ (from the C_{ox} factor) when the physical dimensions are scaled down by κ. This imposes a serious limitation on how low the threshold voltage can be, especially in dynamic circuits and random-access memories. The threshold voltage limitation in turn sets a lower limit on the power-supply voltage V_{dd}, since the circuit delay increases rapidly with the ratio V_t/V_{dd}, as will be discussed in Section 4.2.1.3.

Another nonscaling factor related to kT/q is the inversion-layer thickness, which is unchanged in constant-field scaling. Since the inversion-layer capacitance arising from the finite thickness is in series with the oxide capacitance, the total gate capacitance per unit area of the scaled device increases by a factor less than κ (Baccarani and Wordeman, 1983). This degrades the inversion charge density and therefore the current, especially at low gate voltages, as can be seen from Eq. (3.62).

Because both the junction built-in potential [Eq. (2.84)] and the maximum surface potential [Eq. (2.183)] are in the range of 0.6–1.0 V and do not change significantly with device scaling, the depletion-region widths, Eq. (4.1) and Eq. (4.6), do not scale quite as much as other linear dimensions. This results in worse short-channel effects in the scaled MOSFET, as is evident from Eq. (4.7). To compensate for these effects, the doping concentration must increase more than that suggested by constant-field scaling or generalized scaling.

4.1.3.2 Secondary Nonscaling Factors

Because of subthreshold nonscaling, the voltage level cannot be scaled down as much as the linear dimensions, and the electric field has increased as a result. This triggers several secondary nonscaling effects. First, in our discussions so far, it was implicitly assumed that carrier mobilities are constant, independent of scaling. However, as discussed in Section 3.1.5, the mobility decreases with increasing electric field:

$$\mu_{eff} \approx 32500 \mathscr{E}_{eff}^{-1/3}, \tag{4.11}$$

in units of $cm^2/V\text{-}s$ for $\mathscr{E}_{eff} \le 5 \times 10^5$ V/cm. Beyond $\mathscr{E}_{eff} = 5 \times 10^5$ V/cm, the mobility decreases even faster due to surface roughness scattering (Fig. 3.15). Since it is inevitable that the electric field increases with scaling, carrier mobilities are degraded in scaled MOSFETs. As a result, both the current and the delay improve less than the factors listed in Table 4.3 for generalized scaling. Furthermore, higher fields tend to push device operation more into the velocity-saturated regime. This means that the current gain and the delay improvement are closer to the velocity-saturated column of Table 4.3, and there is little to gain by operating at an even higher field or voltage.

The most serious problems associated with the higher field intensity are reliability and power. The power density increases by a factor of α^2 to α^3 as discussed before. Reliability problems arise from higher oxide fields, higher channel fields, and higher current densities. Even under the fully velocity-saturated condition, the current density increases by $\alpha\kappa$. This aggravates the problem of electromigration in aluminum lines, which is already becoming worse under constant-field scaling (Dennard *et al.*, 1974).

Higher fields also drive gate oxides closer to the breakdown condition, making it difficult to maintain oxide integrity. In fact, in order to curb the growing oxide field, the gate oxide thickness has been reduced less than the lateral device dimensions, e.g., the channel length, as is evident in Table 4.2. This means that the channel doping concentration must be increased more than called for in Table 4.3 to keep short-channel effects [Eq. (4.7)] under control. In other words, the maximum gate depletion width W_{dm} must be reduced more than the oxide thickness t_{ox}. This triggers another set of nonscaling effects, including the subthreshold slope $\propto m = 1 + (3t_{ox}/W_{dm})$, and the substrate sensitivity $dV_t/d(-V_{bs}) = m - 1$ [Eq. (3.45)]. These will be discussed in detail in Section 4.2.3.

4.1.3.3 Other Nonscaling Factors

In practice, there is yet another set of nonscaling factors encountered in CMOS technology evolution. One kind of nonscaling effect is related to the gate and source–drain doping levels. If not properly scaled up, they may lead to gate depletion and source–drain series resistance problems. From Eq. (2.213), polysilicon gate depletion contributes a capacitance $C_p = \varepsilon_{si} q N_p / Q_g$ in series with the oxide capacitance C_{ox}. As C_{ox} increases by a factor of κ while Q_g remains unchanged in constant-field scaling, N_p must scale up by κ also to keep C_p in step with C_{ox}. In generalized scaling, N_p must scale up even more (by $\alpha\kappa$). In reality, this cannot be done because of limitations by solid solubility. The total gate capacitance then scales up by less than C_{ox}, leading to degradation of the inversion charge density and transconductance. Similarly, it is difficult to scale up the source–drain doping level and make the profile more abrupt while scaling down the junction depth. In practice, the source–drain series resistance has not been reduced in proportion to channel resistance, Eq. (3.102). This causes loss of current drive as the parasitic component becomes a more significant fraction of the total resistance in the scaled device.

Another class of nonscaling factors arises from process tolerances. The full benefit of scaling cannot be realized unless all process tolerances are reduced by the same factor as the device parameters. These include channel length tolerance, oxide thickness tolerance, threshold voltage tolerance, etc. ***It is a key requirement and challenge in VLSI technology development to keep the tolerance to a constant percentage of the device parameter as the dimension is scaled down***. This could be a major factor in manufacturing costs as one tries to control a couple of hundred angstroms of channel length or a couple of atomic layers of gate oxide.

4.2 Threshold Voltage

This section focuses on a key design parameter in CMOS technology: threshold voltage. Although the threshold voltage was introduced in Chapter 3, the discussions there were restricted to the case of uniform doping. In this section, threshold-voltage requirements in terms of off- and on-currents are discussed, leading to the design of MOSFET channel profile with nonuniform body doping. The last two parts deal with the effects of quantum mechanics and dopant number fluctuations on threshold voltage.

4.2.1 Threshold-Voltage Requirement

4.2.1.1 Various Definitions of Threshold Voltage

First, we examine the various definitions of threshold voltage and the threshold-voltage requirement from a technology point of view. There are quite a number of different ways to define the threshold voltage of a MOSFET device. In Chapter 3 we followed the most commonly used definition [ψ_s(inv) = $2\psi_B$] of V_t. The advantage of this definition lies in its popularity and ease of incorporation into analytical solutions. However, it is not directly measurable from experimental I–V characteristics (it can be determined from a split C–V measurement; see Exercise 2.6). In Section 3.1.6, we introduced the linearly extrapolated threshold voltage, V_{on}, determined by the intercept of a tangent through the maximum-slope (linear transconductance) point of the low-drain I_{ds}–V_{gs} curve. This is easily measured experimentally, but is about $3kT/q$ higher than the $2\psi_B$ threshold voltage, due to inversion-layer capacitance effects illustrated in Fig. 3.18.

Another commonly employed definition of threshold voltage is based on the subthreshold I_{ds}–V_{gs} characteristics, Eq. (3.40). For a given constant current level I_0 (say, 50 nA/□), one can define a threshold voltage V_t^{sub} such that $I_{ds}\,(V_{gs} = V_t^{sub}) = I_0(W/L)$. The advantages of such a threshold-voltage definition are twofold. First, it is easy to extract from hardware data and is therefore suitable for automated measurement of a large number of devices. Second, the device off current, $I_{off} = I_{ds}(V_{gs} = 0)$, can be directly calculated from I_0, V_t^{sub}, and the subthreshold slope. In subsequent discussions, we will adhere to the $2\psi_B$ definition of V_t. In general, V_t depends on temperature (temperature coefficient), substrate bias (body-effect coefficient), channel length, and drain voltage (short-channel effect, or SCE).

4.2.1.2 Off-current and Standby Power

By definition, the off-current of a MOSFET is the source-to-drain subthreshold leakage current when the gate-to-source voltage is zero and the drain-to-source voltage is V_{dd}, the power supply voltage. From Eq. (3.40), the expression for the off-current with $V_{ds} = V_{dd} >> kT/q$ is

$$I_{off} = I_{ds}|_{V_{gs}=0, V_{ds}=V_{dd}} = I_{ds,Vt} e^{-qV_t/mkT}, \tag{4.12}$$

where

$$I_{ds,Vt} = \mu_{eff} C_{ox} \frac{W}{L} (m-1) \left(\frac{kT}{q}\right)^2 \tag{4.13}$$

is defined as the source-to-drain current at threshold ($V_{gs} = V_t$, $V_{ds} = V_{dd}$). In the worst case, the source–drain voltage of the transistors in the off-state equals the power supply voltage V_{dd}. The standby power dissipation due to I_{off} is then $V_{dd}I_{off}$. For order-of-magnitude estimates, $V_{dd} \approx 1$ V. If it is desired that the standby power of a VLSI chip containing 10^8 transistors be no higher than 1 W, the off-current per transistor should be kept less than 10 nA.[1]

[1] For a small fraction of transistors on the chip or for a larger standby power budget, higher off-current per transistor can be allowed.

Note that I_{ds,V_t} is rather insensitive to the temperature since $\mu_{eff} \propto T^{-3/2}$. However, it does depend on technology. For a 0.1 μm CMOS technology with $t_{ox} \approx 30$ Å, $\mu_{eff} \approx 350\,cm^2/V\text{-}s$, $m \approx 1.3$, and $W/L = 10$, I_{ds,V_t} is approximately 1 μA ($W = 1$ μm). (Note that this number is for nMOSFETs. pMOSFET current is about 3× lower due to the lower hole mobility. Also note that the extrapolated subthreshold current at the linearly extrapolated voltage V_{on} is about 10× higher than this number, as discussed at the end of Subsection 3.1.6.4.) VLSI chips are usually specified for a worst-case temperature of 100°C where the off-current is much higher than that at room temperature because not only does V_t decrease with temperature, but the slope of the $\log(I_{ds})$–V_{gs} curve also degrades in proportion to q/kT. Typically, the inverse slope of subthreshold current is 100 mV/decade at 100 °C. For the factor $\exp(-qV_t/mkT)$ in Eq. (4.12) to deliver a two-orders-of-magnitude reduction from $I_{ds,V_t} = 1$ μA to $I_{off} = 10$ nA, V_t(100 °C) needs to be at least 0.2 V. Because V_t has a negative temperature coefficient of ≈ -0.7 mV/°C (Section 3.1.4.2), this means $V_t(25\,°C) \geq 0.25$ V.[2]

The above figures are acceptable for CMOS logic technologies. In a dynamic memory technology (Dennard, 1984), however, the off-current requirement is much more stringent for the access transistor in the cell: on the order of $I_{off} \approx 10^{-13} - 10^{-14}$ A (see Section 9.2.2). This means $V_t(100\,°C) \geq 0.6$ V for a DRAM access device with $W = L = 0.1$ μm. It should be noted that Eqs. (4.12) and (4.13) are analytical expressions derived under some simplifying approximations, e.g., long channel, uniform doping, etc. They are used here for order-of-magnitude estimates. More exact values of the off-current for a particular design should be obtained by numerical simulations.

Another consideration that may further limit how low the threshold voltage can be is the *burn-in* procedure. Burn-in is required in most VLSI technologies to remove early failures and ensure product reliability. It is usually carried out at elevated temperatures and over voltages to accelerate the degradation process. Both of these conditions further lower the threshold voltage and aggravate the leakage currents. Ideally, burn-in procedure should be designed such that it does not require a compromise on the device performance.

4.2.1.3 On-current and MOSFET Performance

While the lower bound of threshold voltage is set by standby power constraints, the upper bound is imposed by considerations of on-current and switching delay. The on-current of a MOSFET is defined in the saturation region as

$$I_{on} = I_{ds}|_{V_{gs}=V_{dd}, V_{ds}=V_{dd}}. \tag{4.14}$$

Consider an nMOSFET initially in the off state with the source grounded and the drain charged to $V_{ds} = V_{dd}$ (e.g., in one of the CMOS inverter states in Fig. 5.2). If a gate voltage $V_{gs} = V_{dd}$ is applied to turn it on, the drain node will be discharged by the current I_{on} (initially) and the drain voltage will decrease at a rate given by

[2] Lower threshold voltages are allowed in the scenario under footnote 1.

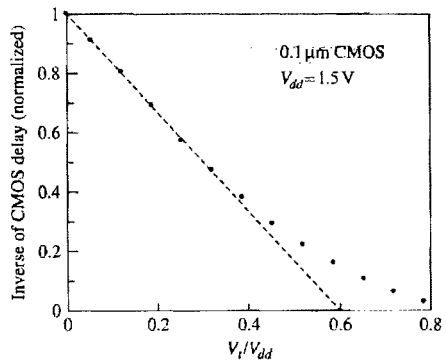


Figure 4.2. The reciprocal of CMOS delay in normalized units versus V_t/V_{dd}. The dots are from SPICE model simulations. The dashed line is a fitting proportional to $0.6 - V_t/V_{dd}$. Here V_t is defined as the gate voltage at which $I_{ds}(V_{ds} = V_{dd})$ equals that of Eq. (4.13). For a given linearly extrapolated, low-drain-bias threshold voltage V_{on}, a larger DIBL results in a lower V_t hence higher I_{off} and I_{on}.

$$C\frac{dV_{ds}}{dt} = -I_{on}, \tag{4.15}$$

where C is the total effective capacitance of the drain node. The switching delay for an incremental change of V_{ds} is then $-CdV_{ds}/I_{on} \propto 1/I_{on}$. It is evident from Chapter 3 that the lower the threshold voltage, the higher the current drive I_{on}, hence the faster the switching speed. From a CMOS performance point of view, it is desirable to have a threshold voltage as low as possible.

It will be discussed in Chapter 5 that because of the finite rise time of V_{gs} at the input, the current that goes into the discharge equation (4.15) is somewhat less than I_{on}. A circuit simulation model can be used to analyze the delay sensitivity to threshold voltage. Figure 4.2 shows a typical example of CMOS performance, defined as the reciprocal of CMOS delay, versus the normalized threshold voltage, V_t/V_{dd}. For $V_t/V_{dd} < 0.5$, the result can be fitted to an expression proportional to $0.6 - V_t/V_{dd}$. This indicates, for example, about 30% of the performance will be lost if V_t/V_{dd} is increased from 0.2 to 0.3. Because of such delay sensitivity, ***the V_t/V_{dd} ratio is usually kept ≤ 0.25 for high performance CMOS circuits.***

4.2.1.4 I_{on} versus I_{off} Characteristics

Since the choice of threshold voltage hinges on the tradeoff between I_{off} and I_{on}, it is a common practice to plot I_{off} directly against I_{on}, thus skipping the ambiguous definition of threshold voltage. Figure 4.3 plots I_{ds} versus V_{gs} for a constant $V_{ds} = V_{dd}$ in both linear and logarithmic scales for the ease of reading I_{off} and I_{on} simultaneously. In essence, adjusting the threshold voltage of the device is equivalent to parallel shifting the I_{ds}–V_{gs} curves horizontally along the V_{gs}-axis.

Note that for an incremental shift of $\Delta V_t > 0$, I_{off} decreases by a factor $\exp(q\,\Delta V_t/mkT)$ while I_{on} decreases by an amount $g_m \Delta V_t$, where $g_m = dI_{ds}/dV_{gs}$ is the saturation

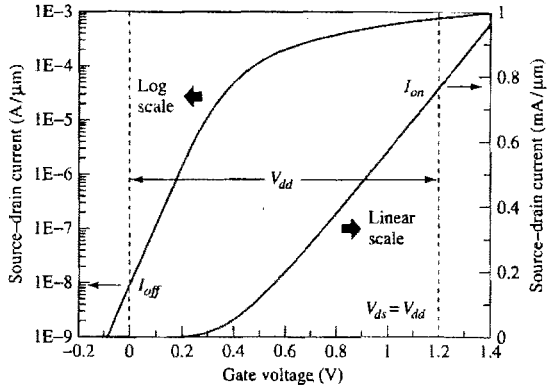


Figure 4.3. I_{ds}–V_{gs} characteristics in both linear and log scales; $V_{dd} = 1.2$ V in this example.

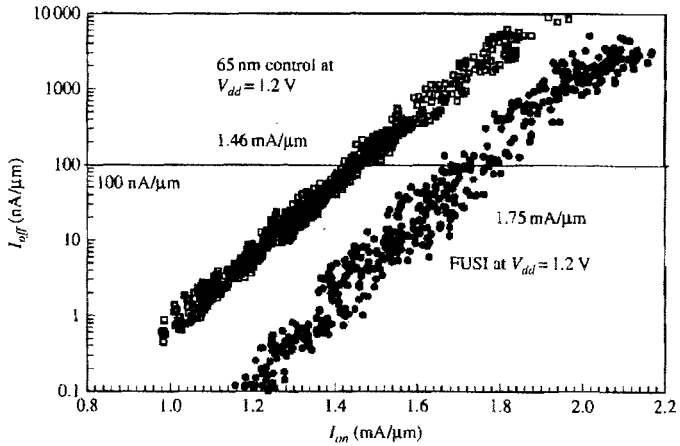


Figure 4.4. An experimental I_{off}–I_{on} plot for 65 nm nMOSFETs (Ranade *et al.*, 2005).

transconductance or the slope of the I_{ds}–V_{gs} curve at $V_{gs} = V_{dd}$. In this regard, the often cited I_{on}/I_{off} ratio is not a meaningful figure of merit because it changes constantly as ΔV_t is adjusted. In fact, to maximize the I_{on}/I_{off} ratio for a given V_{dd}, one would want to shift to as high a threshold voltage as possible so that the entire $0 \le V_{gs} \le V_{dd}$ range is in the subthreshold. That is not a desired mode of operation for high performance CMOS because then I_{on} would be so low that the delay is easily degraded by parasitic capacitances (Chapter 5).

An example of the recently published I_{on}–I_{off} characteristics for nMOSFETs is shown in Fig. 4.4 (Ranade *et al.*, 2005).

4.2.2 Channel Profile Design

In this section, we discuss the design of MOSFET doping profile that satisfies the threshold voltage and other device requirements. Parameters that come into play include the gate length, power supply voltage, and gate oxide thickness. The choice of gate work function is then addressed, leading to the channel profile requirements and trends over the CMOS technology generations.

4.2.2.1 CMOS Design Considerations

CMOS device design involves choosing a set of parameters that are coupled to a variety of circuit characteristics to be optimized. The choice of these device parameters is further subject to technology constraints and system compatibility requirements. Figure 4.5 shows a schematic diagram of the design process and the parameters involved. Because various circuit characteristics are interrelated through the device parameters, tradeoffs among them are often necessary. For example, reduction of W_{dm} improves short-channel effect, but degrades substrate sensitivity; thinner t_{ox} increases current drive, but causes reliability concerns, etc. There is no unique way of designing CMOS devices for a given technology generation. Nevertheless, we attempt here to give a general guideline of how these device parameters should be chosen.

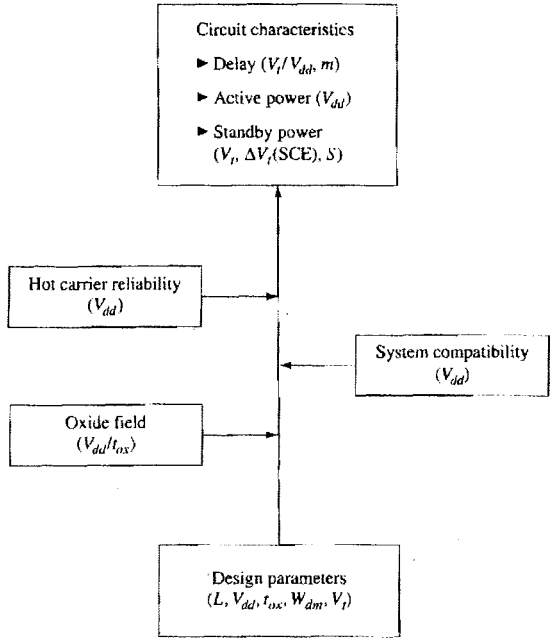


Figure 4.5. A CMOS design flowchart showing device parameters, technology constraints, and circuit objectives.

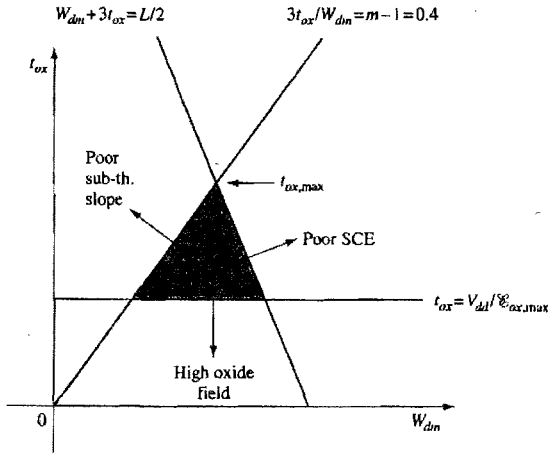


Figure 4.6. A t_{ox}–W_{dm} design plane. Some tradeoff among the various factors can be made within the parameter space bounded by SCE, body effect, and oxide field considerations.

Since threshold voltage plays a key role in determining both I_{off} and I_{on}, it is important to minimize the V_t tolerance, i.e., the spread between the high and low threshold voltages on the chip. The most dominant source of threshold voltage tolerances in a CMOS technology is from the short-channel effect. Channel length variations on a chip due to process imperfections give rise to threshold voltage variations. From Eq. (3.67) of Section 3.2.1.4, the short-channel V_t is lower than that of the long-channel by

$$\Delta V_t = \frac{24t_{ox}}{W_{dm}}\left[\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})} - a(2\psi_B)\right]e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}, \tag{4.16}$$

where $a \approx 0.4$ and W_{dm} is the maximum depletion width at the threshold condition, $\psi_s = 2\psi_B$. The sensitivity of threshold voltage to channel length variations, $\delta V_t/\delta L$, is intimately tied to ΔV_t. Since $\psi_{bi} \approx 2\psi_B \approx 1$ V, and the worst case V_{ds} equals V_{dd}, the factor in the square bracket ranges between ≈ 1 and 2 V for $V_{dd} \approx 1$ V to 5 V. The factor in front of the square bracket, $24t_{ox}/W_{dm} = 8(m-1)$, is related to the factor $m = \Delta V_{gs}/\Delta\psi_s$ illustrated in Fig. 3.5. It was discussed in Sections 3.1.2.3, 3.1.3.3, and 3.1.4.1 that from saturation current, subthreshold slope, and substrate sensitivity considerations, m should not be too much greater than unity, e.g., $m \leq 1.4$. Because of the exponential factor in Eq. (4.16), ΔV_t is very sensitive to $L/(W_{dm}+3t_{ox})$. A good choice is $L/(W_{dm}+3t_{ox}) \geq 2$, which gives $\Delta V_t \leq 0.1$ V for $V_{dd} \approx 1$ V and $\Delta V_t \leq 0.2$ V for $V_{dd} = 5$ V, assuming a median value of $m = 1.3$.[3]

These considerations are captured in a plot of the t_{ox}–W_{dm} design plane in Fig. 4.6. The intercept of the two lines, $W_{dm}+3t_{ox} = L/2$ and $3t_{ox}/W_{dm} = m-1 = 0.4$, defines an upper

[3] It can be shown from $\delta V_t/\delta L$ using Eq. (4.16) that this choice yields a δV_t spread equal to ΔV_t for a channel length tolerance $\delta L/L$ of ± 15%.

bound for the oxide thickness, $t_{ox,max} \approx L/20$. The lower limit of t_{ox} is imposed by technology constraints to $V_{dd}/\mathcal{E}_{ox,max}$, where $\mathcal{E}_{ox,max}$ is the maximum allowable oxide field from breakdown and reliability considerations. ***For a given L and V_{dd}, the allowable parameter space in a t_{ox}–W_{dm} design plane is a triangular area bounded by SCE, oxide field, and subthreshold slope (also substrate sensitivity) requirements.***

In addition to the oxide field limitation, direct quantum mechanical tunneling (Fig. 2.62) also sets a lower limit to the thickness of gate oxide. Gate current density increases sharply as t_{ox} decreases below 2 nm. From Fig. 2.62, the gate tunneling current density for a 1 nm thick oxide biased at 1 V is 10^3–10^4 A/cm^2. Assume $L \approx 30$ nm, the gate current of an individual transistor (< 3 μA/μm) is still small compared with the typical on currents (≈ 1 mA/μm) of the preceding stage so the switching delay of active transistors is hardly affected. But consider 10^8 transistors each with $W/L \approx 10$ and $L \approx 30$ nm, the total gate area per chip is of the order of 0.01 cm^2. The standby power dissipation of all the turned-on transistors[4] in the chip has reached intolerable levels of 10–100 W. ***Given the $t_{ox,max} \approx L/20$ criterion discussed above, the 1 nm t_{ox} limit translates into a channel length limit of ≈ 20 nm for SiO_2.***

If high-κ gate insulators become available, the scale length can be pushed to $\lambda \approx 2t_i$ for very high κ where t_i is the insulator thickness (Section 3.2.1.5). In that case, the minimum channel length can be extended to $2\lambda \approx 4t_i$, or ≈ 10 nm assuming a tunneling limited high-κ thickness of 2.5 nm. The last figure is thicker than that of SiO_2 because of the inherently lower barrier heights (< 3.1 eV of Fig. 2.29) of such materials.

4.2.2.2 Trends of Power Supply Voltage and Threshold Voltage

For a design window to exist in Fig. 4.6, it is required that $V_{dd}/\mathcal{E}_{ox,max} \leq t_{ox,max} \approx L/20$. This imposes an upper limit on the power supply voltage, namely,

$$V_{dd} \leq L\mathcal{E}_{ox,max}/20. \tag{4.17}$$

For $L = 1$ μm CMOS technology, the gate oxides are relatively thick and $\mathcal{E}_{ox,max} \approx 3$ MV/cm. Equation (4.17) requires $V_{dd} \leq 15$ V. There is plenty of design room to choose the power supply and threshold voltages that satisfy both the off-current and the performance requirements discussed in Sections 4.2.1.2 and 4.2.1.3. For example, $V_{dd} = 5$ V and $V_t = 0.8$–1.0 V as shown in Fig. 4.7 in which the history and trends of power supply voltage, threshold voltage, and oxide thickness are plotted for CMOS logic technologies from 1.0 μm to 0.02 μm channel lengths (Taur *et al.*, 1995a). At shorter channel lengths, V_{dd} must be reduced. It becomes increasingly more difficult to satisfy both the performance and the off-current requirements. Fortunately, $\mathcal{E}_{ox,max}$ tends to increase for thinner oxides (see Section 2.5.6) as L is scaled down. This allows V_{dd} to scale at a slower rate than the channel length. Experimentally, $\mathcal{E}_{ox,max} \approx 6$ MV/cm for oxides thinner than 3 nm. Equation (4.17) then requires, e.g., that $V_{dd} \leq 1.5$ V for $L = 50$ nm CMOS technology. With such a low supply voltage, one often faces a tradeoff of circuit speed versus leakage current. Scaling down V_t causes I_{off} to increase exponentially. Even for the

[4] The worst case gate leakage occurs with nMOSFETs biased at $V_{gs} = V_{dd}$ and $V_{ds} = 0$ (electrons tunnel from the inversion channel to the gate).

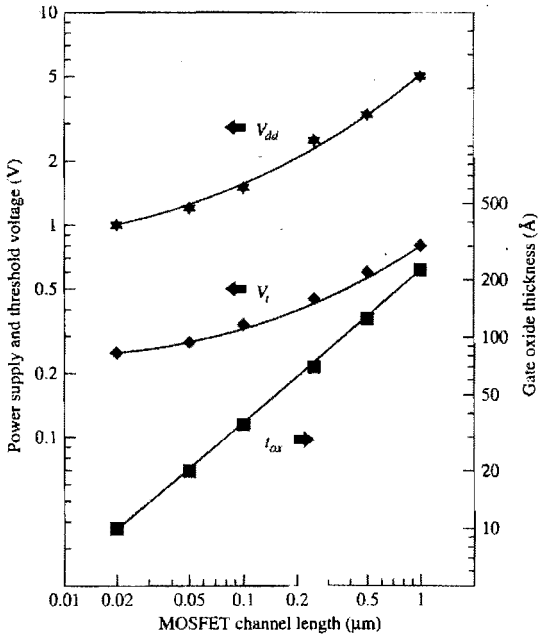


Figure 4.7. Trends of power-supply voltage, threshold voltage, and gate oxide thickness versus channel length for CMOS technologies from 1 μm to 0.02 μm. (After Taur *et al.*, 1995a.)

same V_t, I_{off} increases since $I_{ds,Vt}$ of Eq. (4.13) increases as the devices are scaled down – a manifestation of subthreshold nonscalability. For this reason and for compatibility with the standardized power supply voltage of earlier generation systems, ***the general trend is that V_{dd} has not been scaled down in proportion to L, and V_t has not been scaled down in proportion to V_{dd}***, as is evident in Fig. 4.7. At L = 20 nm, $\mathscr{E}_{ox,max}$ is pushed to 10 MV/cm for operation at V_{dd} = 1 V.

As a result of the non-scaled V_{dd}, not only does the field increase over the CMOS generations, the increasing power density (Table 4.3) also becomes more difficult to manage. It is discussed in Section 5.1.1 that the active or switching power of a CMOS circuit is given by

$$P_{ac} = CV_{dd}^2 f, \tag{4.18}$$

where C is the total equivalent capacitance being charged and discharged in a clock cycle, and f is the clock frequency. The power versus delay tradeoff can be represented conceptually in a V_{dd}–V_t design plane shown in Fig. 4.8 (Mii *et al.*, 1994). Higher performance, i.e., shorter delay, pushes for higher V_{dd} and lower V_t, which inevitably results in higher active power or higher standby power, or both. ***Depending on the specific requirements of the application, CMOS technologies can be tailored to some extent by choosing an appropriate set of power supply and threshold voltages***. High

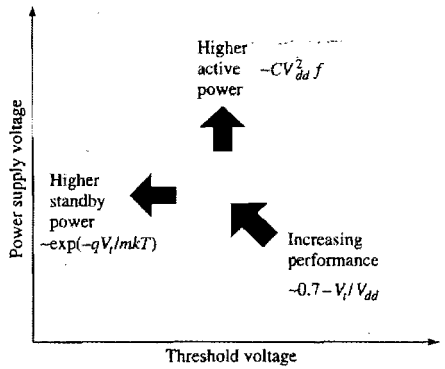


Figure 4.8. CMOS performance, active power, and standby power tradeoff in a V_{dd}–V_t design plane. The performance here is defined as the reciprocal of CMOS delay.

performance CMOS usually operates at the upper left-hand corner of the design space and pushes both power limits. Low power CMOS can operate at lower supply voltages and possibly at a higher threshold voltage if the standby power is of primary concern. It is a common practice in the state-of-the-art CMOS technologies to provide multiple threshold voltages on a chip to allow the design flexibility of using different types of devices for different functions, e.g., in memory and logic circuits. This comes, of course, at the expense of additional process complexity and cost.

4.2.2.3 Effect of Gate Work Function

To realize the threshold voltages desired from the above design considerations, it is important to use a gate material with the proper work function. Gate work function (ϕ_m) has a major impact on the threshold voltage of, e.g., nMOSFETs,

$$V_t = V_{fb} + 2\psi_B + \frac{-Q_d}{C_{ox}}, \tag{4.19}$$

since it sets the flatband voltage of the MOSFET,

$$V_{fb} = \phi_m - \phi_s = \phi_m - \left(\chi + \frac{E_g}{2q} + \psi_B\right). \tag{4.20}$$

For nMOSFETs, $2\psi_B \approx 1$ V and $Q_d < 0$, so V_t is easily larger than 1 V unless V_{fb} is substantially negative. To achieve the low threshold voltages required in Fig. 4.7, n^+-polysilicon gates have been used for n-channel MOSFETs so that $V_{fb} = -E_g/2q - \psi_B$. This results in near cancellation of the first and the second terms of Eq. (4.19). V_t is then largely determined by the third term in proportion to the depletion charge density at the $2\psi_B$ condition. How the channel doping profile should be designed in order to achieve the desired depletion charge density and therefore V_t is discussed in the next subsection.

Before p^+-polysilicon gates become technologically available, n^+-polysilicon gates are used for pMOSFETs as well in 1 μm and 0.5 μm CMOS generations. This means V_{fb} is a small negative number ($V_{fb} = -E_g/2q + \psi_B$ for n-type silicon) and the first two terms of the

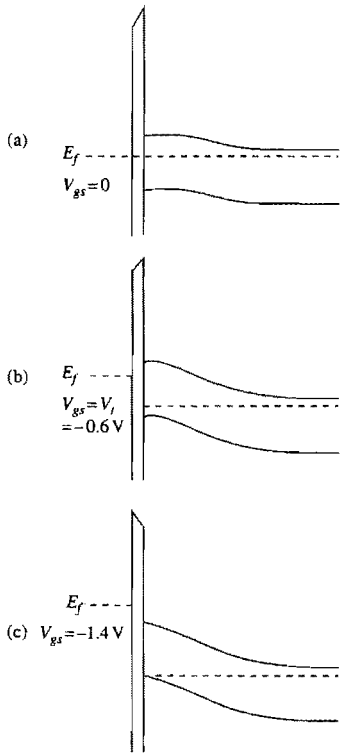


Figure 4.9. Band diagram of a buried-channel pMOSFET with n^+-polysilicon gate. A shallow p-type layer is implanted at the surface to lower the magnitude of threshold voltage. The gate is biased (a) in subthreshold, (b) at threshold, and (c) beyond threshold. (After Taur *et al.*, 1985.)

V_t equation (the second term is $-2\psi_B \approx -1\,V$ for pMOS) add up to $< -1\,V$ for pMOSFETs. To make V_t less negative, the third term of the V_t equation needs to be positive, which means p-type doping for pMOSFETs or a counterdoped channel. Since the depletion charge density is negative, the surface field at threshold is such that holes are accelerated toward the substrate, and the channel for holes is formed at a potential minimum below the surface. Such devices are called *buried-channel* MOSFETs. Figure 4.9 shows the band diagrams of a buried-channel pMOSFET at several gate voltages both below and above the threshold. As the gate voltage becomes more negative than the threshold, the field changes sign and the channel moves to the surface. But the magnitude of the effective field is still lower than that of a conventional surface-channel device. ***Although a buried-channel device has higher mobilities, its short-channel effect is inherently worse than that of a surface-channel device*** (Nguyen and Plummer, 1981). This is because the counterdoping (especially boron) at the surface tends to diffuse deeper into the silicon during subsequent thermal cycles in the process. As the channel length and the power supply voltage are scaled down, a lower magnitude of threshold

voltage is required. It becomes increasingly more difficult to build a buried-channel device since higher counterdoping in the channel invariably leads to wider gate depletion widths and poorer short-channel effects. ***For CMOS logic technologies of 0.25 μm channel length and below, dual polysilicon gates (n^+-polysilicon for nMOSFET and p^+-polysilicon for pMOSFET) are used so that both types of devices are surface-channel devices*** (Wong *et al.*, 1988).

Near the limits of CMOS scaling ($L \approx$ 10–20 nm), the threshold voltages may become too high even with dual-polysilicon gates and extreme retrograde doping (Section 4.2.3.5). In principle, one way to further reduce the threshold magnitude is by counterdoping of the channel, as will be discussed in Section 4.2.3.6.

There have been numerous research explorations (e.g., Davari *et al.*, 1987) on using metal gates with a midgap work function. The benefits are high gate conductivity, absence of polysilicon depletion effects, and the simplicity of using a single gate material for both n- and pMOSFETs. Midgap work function gates exhibit symmetric flatband voltages: $V_{fb}=-\psi_B$ (p-type) for nMOSFETs and $V_{fb}=\psi_B$ (n-type) for pMOSFETs. The resulting threshold voltage magnitudes are in the range of 0.5–1.0 V [Eq. (4.19)]. This meets the V_t requirements for 1 μm and 0.5 μm CMOS technologies in Fig. 4.7. An added benefit is that it takes much less depletion charge [the third term in Eq. (4.19)] to achieve the same V_t magnitude with a midgap gate than with an n^+-polysilicon gate for nMOSFETs. Less depletion charge means lower surface fields and therefore higher mobility. In reality, however, no midgap-work function gate material has been used in VLSI production because of technology issues such as compatibility of gate material with thin gate oxides. Gate conductivity requirement has been met with self-aligned silicide technology (Section 5.2.1.4). Once the CMOS technology is scaled to 0.25 μm and below, V_t magnitudes < 0.5 V are needed (Fig. 4.7) which are difficult to achieve with a midgap work function gate.

4.2.2.4 Channel Profile Requirement and Trends

It was discussed above that with a n^+-polysilicon gate for n-channel MOSFETs (and p^+-poly for pMOSFETs), the first and the second terms of Eq. (4.19) essentially cancel out and V_t is largely determined by the depletion charge term. For a uniform channel doping, the maximum gate depletion width at the $2\psi_B$ condition,

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a}}, \tag{4.21}$$

and the depletion charge term of the threshold voltage,

$$\frac{-Q_d}{C_{ox}} = \frac{qN_aW_{dm}}{C_{ox}} = \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}}, \tag{4.22}$$

are coupled through the parameter N_a, and therefore cannot be varied independently (for a given t_{ox}). In Section 4.2.2.1, we discussed that in order to control the short-channel effect, $W_{dm} + 3t_{ox} = m\,W_{dm}$ should be on the order of $L/2$. The doping concentration that satisfies this requirement may not give the desired threshold voltage that satisfies the on- and off-current requirements.

For a given W_{dm}, it is necessary to employ nonuniform doping to adjust the depletion charge density to obtain the desired V_t. Nonuniform channel doping gives the device designer an additional degree of freedom to tailor the profile for meeting both the SCE and the threshold requirements. Such an optimization is made possible by the ion implantation technology.

Channel profile trends can be inferred by expressing the threshold voltage in the uniformly doped case as

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}} = V_{fb} + 2\psi_B + 2(m-1)2\psi_B, \tag{4.23}$$

which does not scale much as neither m nor ψ_B changes significantly with channel length or doping. In fact, both m and ψ_B tend to increase slightly as the CMOS channel length scales down and higher doping is required. This is contrary to the downward trend of the V_t requirement depicted in Fig. 4.7. For example, for a typical $m \approx 1.3$, $V_t \approx 0.6$ V with n^+-polysilicon gates. While this value happens to meet the V_t requirement for the 0.5 μm CMOS generation, it is too low for 1 μm CMOS and too high for CMOS generations 0.25 μm and below. It is shown in the next subsection that a high-low doping profile increases the depletion charge density for a given W_{dm} and therefore raises V_t over the uniformly doped value, whereas a low-high profile reduces the depletion charge and lowers V_t.

4.2.3 Nonuniform Doping

In this subsection, analytic expressions for the maximum depletion width and the threshold voltage are derived under nonuniform doping conditions. Specific results are given for both high-low and low-high doping profiles.

4.2.3.1 Integral Solution to Poisson's Equation

Mathematically the surface potential, electric field, and threshold voltage for the case of nonuniform channel doping can be solved using the depletion approximation. For a nonuniform p-type doping profile $N(x)$ in the same x-coordinate as defined in Fig. 2.35, the electric field is obtained by integrating Poisson's equation once (neglecting mobile carriers in the depletion region):

$$\mathscr{E}(x) = \frac{q}{\varepsilon_{si}} \int_x^{W_d} N(x)dx, \tag{4.24}$$

where W_d is the depletion-layer width. Integrating again gives the surface potential,

$$\psi_s = \frac{q}{\varepsilon_{si}} \int_0^{W_d} \int_x^{W_d} N(x')\,dx'\,dx \tag{4.25}$$

Using integration by parts, one can show that Eq. (4.25) is equivalent to (Brews, 1979)

$$\psi_s = \frac{q}{\varepsilon_{si}} \int_0^{W_d} xN(x)\,dx. \tag{4.26}$$

The integral of $xN(x)$ equals the center of mass of $N(x)$ within $(0, W_d)$ times the integral of $N(x)$.

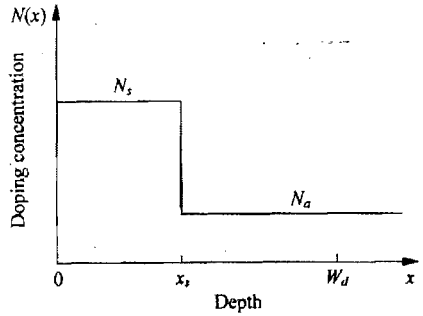


Figure 4.10. A schematic diagram showing the high-low step doping profile. $x = 0$ denotes the silicon–oxide interface.

The maximum depletion-layer width (long-channel) W_{dm} is determined by the condition $\psi_s = 2\psi_B$ when $W_d = W_{dm}$. ***The threshold voltage of a nonuniformly doped MOSFET is then determined by both the integral (depletion charge density) and the center of mass of N(x) within (0, W_{dm}).***

4.2.3.2 A High-Low Step Profile

Consider the idealized step doping profile shown in Fig. 4.10 (Rideout *et al.*, 1975). It can be formed by making one or more low-dose, shallow implants into a uniformly doped substrate of concentration N_a. After drive-in, the implanted profile is approximated by a region of constant doping N_s that extends from the surface to a depth x_s. If the entire depletion region at the threshold condition is contained within x_s, the MOSFET can be considered as uniformly doped with a concentration N_s. The case of particular interest analyzed here is when the depletion width W_d exceeds x_s, so that part of the depletion region has a charge density N_s and part of it N_a. The integration in Eq. (4.26) can be easily carried out for this profile to yield the surface potential, or the band bending at the surface,

$$\psi_s = \frac{qN_s}{2\varepsilon_{si}} x_s^2 + \frac{qN_a}{2\varepsilon_{si}} (W_d^2 - x_s^2). \tag{4.27}$$

This equation can be solved for W_d as a function of ψ_s:

$$W_d = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(\psi_s - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)}. \tag{4.28}$$

This is less than the depletion width in the uniformly doped (N_a) case for the same surface potential. The electric field at the surface is obtained by evaluating the integral in Eq. (4.24) with $x = 0$:

$$\mathscr{E}_s = \frac{qN_s x_s}{\varepsilon_{si}} + \frac{qN_a(W_d - x_s)}{\varepsilon_{si}}. \tag{4.29}$$

From Gauss's law, the total depleted charge per unit area in silicon is given by

$$Q_s = -\varepsilon_{si}\mathscr{E}_s = -qN_s x_s - qN_a(W_d - x_s), \tag{4.30}$$

as would be expected from Fig. 4.10. ***The effect of the nonuniform surface doping is then to increase the depletion charge within $0 \le x \le x_s$ by $(N_s - N_a)\,x_s$ and, at the same time, reduce the depletion layer width as indicated by Eq. (4.28).***

Substituting Eq. (4.28) into Eq. (4.30) for Q_s, the gate voltage equation (3.14) becomes

$$V_{gs} = V_{fb} + \psi_s + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si}qN_a\left(\psi_s - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)} + \frac{q(N_s - N_a)x_s}{C_{ox}}. \tag{4.31}$$

By definition, the threshold voltage is the gate voltage at which $\psi_s = 2\psi_B$, i.e.,

$$V_t = V_{fb} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si}qN_a\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)} + \frac{q(N_s - N_a)x_s}{C_{ox}}. \tag{4.32}$$

The maximum depletion width (long-channel) at threshold is given by Eq. (4.28) with $\psi_s = 2\psi_B$:

$$W_{dm} = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)}. \tag{4.33}$$

There is some ambiguity as to whether $2\psi_B$ is defined in terms of N_s or N_a. We adopt the convention that $2\psi_B$ is defined in terms of the p-type concentration at the depletion-layer edge, i.e., $2\psi_B = (2kT/q)\ln(N_a/n_i)$. In fact, it makes very little difference which concentration we use, since $2\psi_B$ is a rather weak function of the doping concentration anyway.[5] Further refinement of the threshold condition would require a numerical simulation of the specific profile.

In Section 3.1.3, we showed that the inverse subthreshold slope is given by $2.3mkT/q$ per decade where $m = dV_{gs}/d\psi_s$ at $\psi_s = 2\psi_B$. In the nonuniformly doped case, m can be evaluated from Eq. (4.31):

$$m \equiv \frac{dV_{gs}}{d\psi_s}(\psi_s = 2\psi_B) = 1 + \frac{\sqrt{2\varepsilon_{si}qN_a}}{2C_{ox}}\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)^{-1/2}. \tag{4.34}$$

[5] The "$2\psi_B$" definition of threshold voltage is only a historical convention. Actually, the channel "turns on" when the surface potential is within 0.1 V (a few kT/q) of the conduction band edge of the n^+ source, regardless of the p-type body doping. In that respect, the approximation $2\psi_B \sim 1$ V is frequently used in the discussions.

It can be expressed in terms of W_{dm} using Eq. (4.33):

$$m = 1 + \frac{\varepsilon_{si}/W_{dm}}{C_{ox}} = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{dm}}. \tag{4.35}$$

These expressions are consistent with Eq. (3.27) for a uniformly doped channel. This is to be expected from the basic concept of m in Fig. 3.5, which applies regardless of the doping specifics. Similarly, the threshold voltage in the presence of a substrate bias V_{bs} is given by Eq. (4.32) with the $2\psi_B$ term in the square root replaced by $2\psi_B - V_{bs}$. Using Eq. (4.33), one can show that the substrate sensitivity is

$$\frac{dV_t}{d(-V_{bs})} = \frac{\varepsilon_{si}/W_{dm}}{C_{ox}} = \frac{C_{dm}}{C_{ox}} = m - 1. \tag{4.36}$$

Therefore, ***all the previous expressions for the depletion capacitance, subthreshold slope, and body-effect coefficient in terms of W_{dm} for the uniformly doped case remain valid for the nonuniformly doped case.*** The only difference is that the maximum depletion layer width W_{dm} in the high-low step doping case is given by Eq. (4.33) instead of Eq. (4.21).

4.2.3.3 Generalization to a Gaussian Profile

The results of the high-low step profile discussed above can be generalized to other profiles as well. As far as the threshold voltage and depletion width are concerned, the added doping density in Fig. 4.10, $N_s - N_a$ over $(0, x_s)$, is equivalent to the delta-function profile in Fig. 4.11(b) with an equivalent dose of

$$D_I = (N_s - N_a)x_s \tag{4.37}$$

centered at $x_c = x_s/2$. This is because both the integrals of $N(x)$ [Eq. (4.24)] and the center of mass of $N(x)$ [Eq. (4.26)] over $(0, W_{dm})$ are identical between the two profiles.

Similar arguments apply to a general Gaussian (or other symmetric) profile in Fig. 4.11(a) with a dopant distribution,

$$N(x) = \frac{D_I}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x - x_c)^2}{2\sigma^2}\right), \tag{4.38}$$

where σ is the implant straggle. The effect of such an implanted profile on threshold voltage and depletion-layer width is equivalent to that of the step doping profile discussed above, independent of σ. Substituting Eq. (4.37) and $x_c = x_s/2$ into the threshold voltage equation (4.32) yields

$$V_t = V_{fb} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si}qN_a\left(2\psi_B - \frac{qD_I x_c}{\varepsilon_{si}}\right)} + \frac{qD_I}{C_{ox}}. \tag{4.39}$$

Similarly, the maximum depletion width, Eq. (4.33), becomes

$$W_{dm} = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{qD_I x_c}{\varepsilon_{si}}\right)}. \tag{4.40}$$

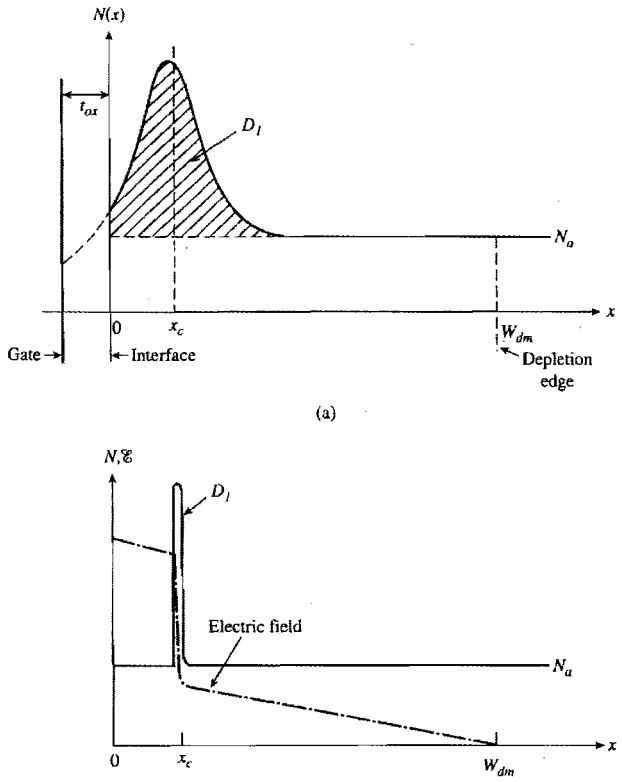


Figure 4.11. Schematic diagrams showing (a) an implanted Gaussian profile and (b) a delta-function profile equivalent to (a). The electric field is proportional to the area under the depleted charge $N(x)$ [Eq. (4.24)]. It has a step rise where the delta function doping is located. (After Brews, 1979.)

For a given implanted dose D_I, the resulting threshold voltage shift depends on the location of the implant, x_c. ***For shallow surface implants, $x_c = 0$, there is no change in the depletion width. The V_t shift is simply given by qD_I/C_{ox}, as with a sheet of charge at the silicon–oxide interface.*** All other device parameters, e.g., substrate sensitivity and subthreshold slope, remain unchanged. As x_c increases for a given dose, both the maximum depletion width and the V_t shift decrease. If x_c is not too large, one can always readjust the background doping N_a to a lower value N_a' to restore W_{dm} to its original value. The threshold voltage, in the meantime, is shifted by an amount less than the shallow implant case.

Although the above analysis on nonuniform doping assumes $N_s > N_a$, the results remain equally valid if $N_s < N_a$. Such a profile is referred to as the *retrograde channel doping*, discussed in the next subsection.

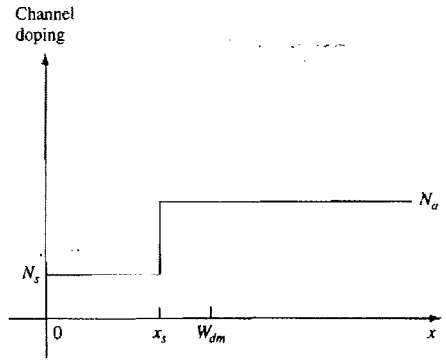


Figure 4.12. A schematic diagram showing the low-high (retrograde) step doping profile. $x = 0$ denotes the silicon–oxide interface.

4.2.3.4 Retrograde (Low–High) Channel Profile

When the channel length is scaled to 0.25 μm and below, higher doping concentration is needed in the channel to reduce W_{dm} and control short-channel effects. If a uniform profile were used, the threshold voltage [Eq. (4.23)] would be too high even with dual polysilicon gates. The problem is further aggravated by quantum effects, which, as will be discussed in Section 4.2.4, can add another 0.1–0.2 V to the threshold voltage because of the increasing fields (van Dort *et al.*, 1994).

To reduce the threshold voltage without significantly increasing the gate depletion width, a retrograde channel profile, i.e., a low–high doping profile as shown schematically in Fig. 4.12, is required (Sun *et al.*, 1987; Shahidi *et al.*, 1989). Such a profile is formed using higher-energy implants that peak below the surface. It is assumed that the maximum gate depletion width extends into the higher-doped region. All the equations in Section 4.2.3.2 remain valid for $N_s < N_a$. For simplicity, we assume an ideal retrograde channel profile for which $N_s = 0$. Equation (4.32) then becomes

$$V_t = V_{fb} + 2\psi_B + \frac{qN_a}{C_{ox}}\sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^2} - \frac{qN_a x_s}{C_{ox}}. \tag{4.41}$$

Similarly, Eq. (4.33) gives the maximum depletion width,

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^2}. \tag{4.42}$$

The net effect of low–high doping is that the threshold voltage is reduced, but the depletion width has increased, just opposite to that of high-low doping. Note that Eq. (4.42) has the same form as Eq. (2.91) for a p–i–n diode discussed in Section 2.2.2. All other expressions, such as those for the subthreshold slope and the substrate sensitivity, in Section 4.2.3.2 apply with W_{dm} replaced by Eq. (4.42).

4.2.3.5 Extreme Retrograde Profile and Ground-Plane MOSFET

Two limiting cases are worth discussing. If $x_s \ll (4\varepsilon_{si}\psi_B/qN_a)^{1/2}$, then W_{dm} remains essentially unchanged from the uniformly doped value [Eq. (4.42)], while V_t is lowered by a net amount equal to qN_ax_s/C_{ox} [Eq. (4.41)]. In the other limit, N_a is sufficiently high that $x_s \gg (4\varepsilon_{si}\psi_B/qN_a)^{1/2}$. In that case, $W_{dm} \approx x_s$, and the entire depletion region is undoped. All the depletion charge is concentrated at the edge of the depletion region. The square root term in Eq. (4.41) can be expanded into a power series to yield

$$V_t = V_{fb} + 2\psi_B + \frac{\varepsilon_{si}(2\psi_B/x_s)}{C_{ox}}. \tag{4.43}$$

The last term stems from the depletion charge density in silicon, $\varepsilon_{si}(2\psi_B/x_s)$, which can also be derived from Gauss's law by considering that the field in the undoped region is constant and equals $2\psi_B/x_s$ at threshold. Note that the work function difference that goes into V_{fb} is between the gate and the p^+ silicon at the edge of the depletion region. Using $m = 1 + 3t_{ox}/W_{dm} = 1 + 3t_{ox}/x_s$, one can write Eq. (4.43) as

$$V_t = V_{fb} + 2\psi_B + (m-1)2\psi_B. \tag{4.44}$$

Comparison with Eq. (4.23) shows that, with the extreme retrograde profile, the depletion charge (the third) term of V_t is reduced to half of the uniformly doped value. If there is a substrate bias V_{bs} present, the $2\psi_B$ factor in the last term of Eq. (4.44) is replaced by $(2\psi_B - V_{bs})$, i.e.,

$$V_t = V_{fb} + 2m\psi_B - (m-1)V_{bs}. \tag{4.45}$$

Since ψ_B is a weak function of N_a, the above results are independent of the exact value of N_a as long as it is high enough to satisfy $x_s \gg (4\varepsilon_{si}\psi_B/qN_a)^{1/2}$. All the essential device characteristics, such as SCE (W_{dm}), subthreshold slope (m), and threshold voltage, are determined by the depth of the undoped layer, x_s. ***The limiting case of retrograde channel profile therefore degenerates into a ground-plane MOSFET*** (Yan *et al.*, 1991). The band diagram and charge distribution of such a device at threshold condition are shown schematically in Fig. 4.13. Note that the field is constant (no curvature in potential) in the undoped region between the surface and x_s. There is an abrupt change of field at $x = x_s$, where a delta function of depletion charge (area $= 2\varepsilon_{si}\psi_B/x_s$) is located. Beyond x_s, the bands are essentially flat. It is desirable not to extend the p^+ region under the source and drain junctions, since that will increase the parasitic capacitance. The ideal channel doping profile is then that of a low–high–low type shown in Fig. 4.14, in which the narrow p^+ region is used only to confine the gate depletion width. Such a profile is also referred to as *pulse-shaped doping* or *delta doping* in the literature. The integrated dose of the p^+ region must be at least $2\varepsilon_{si}\psi_B/qx_s$ to provide the gate depletion charge needed. It is advisable to use somewhat higher than the minimum dose to supply additional depletion charge to temper the source–drain fields in short-channel devices. However, too high a p^+ dose or concentration may result in band-to-band tunneling leakage between the source or drain and the substrate, as mentioned in Section 2.5.2.

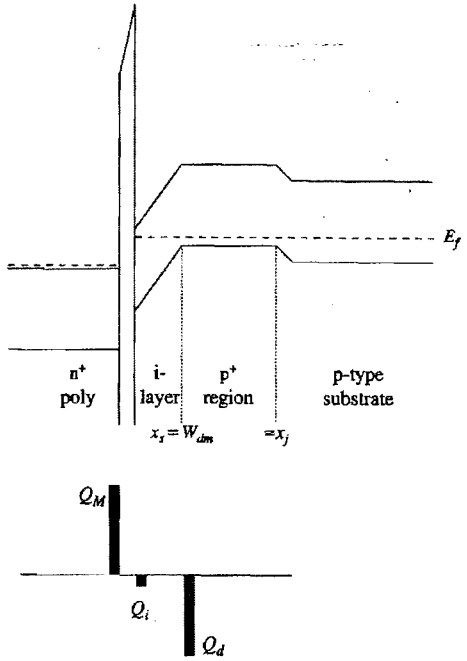


Figure 4.13. Band diagram and charge distribution of an extreme retrograde-doped or ground-plane nMOSFET at threshold condition.

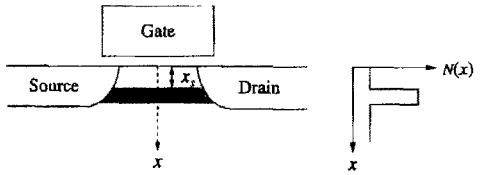


Figure 4.14. Schematic cross section of a low–high–low, or pulse-shaped, or delta-doped MOSFET. The doping concentration along the dashed line is depicted in the profile to the right. The highly doped region corresponds to the shaded area in the cross section.

4.2.3.6 Counter-Doped Channel

When CMOS devices are scaled to 20 nm channel lengths and below, the field is so high and the quantum effect so strong that even the extreme retrograde profile cannot deliver a $V_t \approx 0.2$ V with n^+ and p^+ silicon gates. Besides finding new gate materials with work functions outside of n^+ and p^+ silicon, further reduction of V_t can be accomplished, at least in principle, by either counterdoping the channel or forward biasing the substrate.

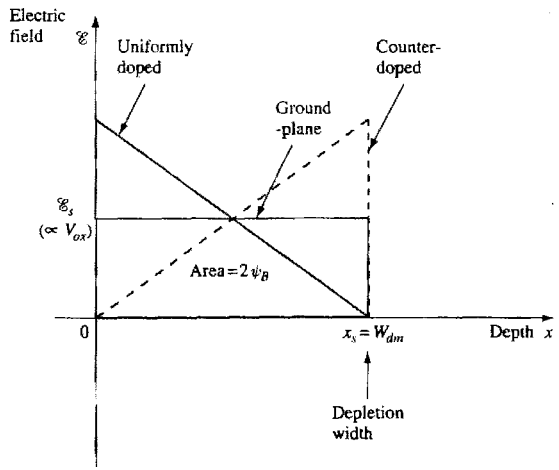


Figure 4.15. Graphical interpretation of uniformly doped, extreme retrograde or ground-plane, and counterdoped profiles. The band bending is given by the area under $\mathscr{E}(x)$ which equals $2\psi_B$ at threshold for all three cases.

A forward substrate bias also helps improve short-channel effects as it effectively reduces the built-in potential, ψ_{bi} in Eq. (3.67), between the source–drain and the p-type substrate. The flip side is that it causes source junction leakage, increases drain-to-substrate capacitance, and degrades subthreshold slope and body effect.

Instead of analyzing the counter-doped channel mathematically by letting $N_s < 0$ for the profile in Fig. 4.12, it is more instructive to give a graphical interpretation of the potential, field, and depletion charge by plotting the electric field versus depth as shown in Fig. 4.15. In the uniformly doped case, $\mathscr{E}(x)$ is a straight line with a negative slope whose magnitude is proportional to the substrate doping concentration N_a. The x-intercept gives the depletion layer width where $\mathscr{E} = 0$. The y-intercept gives the surface electric field $\mathscr{E}_s$ which from Gauss's law is proportional to the total depletion charge per unit area. Since $\mathscr{E} = -d\psi/dx$, the triangular area under $\mathscr{E}(x)$ equals the surface potential or the band bending ψ_s. As the gate voltage increases, both W_d and $\mathscr{E}_s$ increase and so does ψ_s until it reaches $2\psi_B$. At this point, surface inversion occurs and the depletion layer width has reached its maximum value. The depletion charge term of the threshold voltage, $V_{ox} = \varepsilon_{si}\mathscr{E}_s/C_{ox}$, is proportional to the y-intercept or the surface field $\mathscr{E}_s$ when $\psi_s = 2\psi_B$.

In the extreme retrograde or the ground-plane case, $\mathscr{E}(x)$ is constant within the undoped region, $0 < x < x_s$, where there is no depletion charge. At the threshold condition, the shaded rectangular area for the ground-plane case is approximately the same as the triangular area under the uniformly doped $\mathscr{E}(x)$ since $2\psi_B$ is a rather weak function of N_a and can be considered as a constant for practical purposes. It is then clear that the depletion charge term of V_t or the y-intercept of the ground-plane case is half of that of the uniformly doped case – exactly as indicated by Eqs. (4.44) and (4.23).

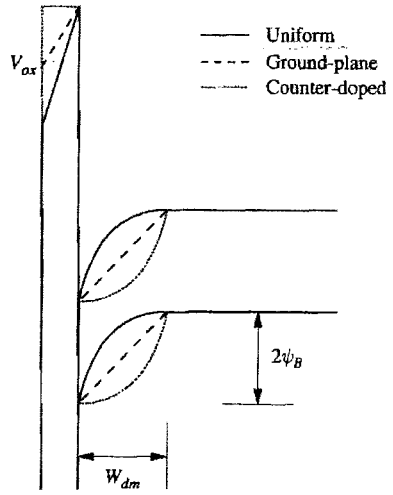


Figure 4.16. Band diagrams of uniformly doped, ground-plane (extreme retrograde), and counter-doped MOSFETs at threshold.

A specific case of the counter-doped channel is shown in Fig. 4.15. The slope $d\mathscr{E}/dx$ has the same magnitude as the uniformly doped case, but of the opposite polarity. Both the depletion width (x-intercept) and the band bending [area under $\mathscr{E}(x)$] are the same as the previous two cases. But the y-intercept ($\mathscr{E}_s$) is zero which means that the net charge in silicon is zero due to cancellation of the counter-doped charge with the depletion charge at the edge of the depletion region. This yields a very low V_t. Further counter-doping would result in $\mathscr{E}_s < 0$ or a buried channel MOSFET.

The band diagrams of these three doping cases at the threshold condition are further illustrated in Fig. 4.16. Both the depletion width and the band bending are kept the same for all three. But the surface fields (slopes) are very different, leading to dramatically different potential drops across the gate oxide.

4.2.3.7 Laterally Nonuniform Channel Doping

So far we have discussed nonuniform channel doping in the vertical direction. Another type of nonuniform doping used in very short-channel devices is in the lateral direction. For nMOSFETs, it is achieved by a medium-dose p-type implant carried out together with the n^+ source–drain implant after gate patterning. As shown in Fig. 4.17, the p-type doping peaks near the source and drain ends of the device but dips in the middle because of blocking of the implant by the gate. Such a self-aligned, laterally nonuniform channel doping is often referred to as *halo* or *pocket* implants (Ogura *et al.*, 1982). Figure 4.17 shows how halo works to counteract the short-channel effect, i.e., threshold rolloff toward the shorter devices within a spread of the channel length (or gate length). At the longer end of the spread shown in Fig. 4.17(a), the two p^+ pockets are farther apart than at the shorter end of the spread in Fig. 4.17(b). This creates a higher average p-type

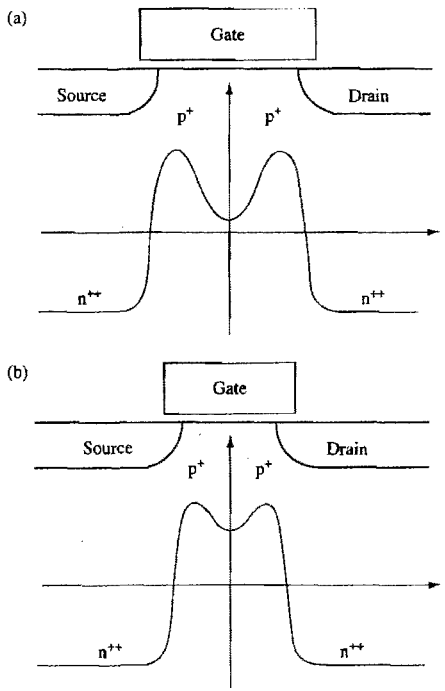


Figure 4.17. Laterally nonuniform halo doping in nMOSFETs. For a given design length on the mask, there is a spread of the actual gate lengths on the wafer. The longer end of the spread is shown in (a), the shorter in (b). The sketch below each cross section shows the schematic doping variation along a horizontal cut through the source and drain regions.

doping in the shorter device than in the longer device. Higher doping means higher threshold voltage. ***So laterally nonuniform halo doping establishes a tendency for the threshold voltage to increase toward the shorter devices, which works to offset the short-channel effect in the opposite direction***. With an optimally designed 2-D nonuniform doping profile called the *superhalo*, it is possible in principle to counteract the short-channel effect and achieve nearly identical I_{on} and I_{off} in devices of different channel lengths within the process tolerances of a 25 nm MOSFET (Taur *et al.*, 1998).

4.2.4 Quantum Effect on Threshold Voltage

It was discussed in Section 2.3.2 that in the inversion layer of a MOSFET, carriers are confined in a potential well very close to the silicon surface. The well is formed by the oxide barrier (essentially infinite except for tunneling calculations) and the silicon conduction band, which bends down severely toward the surface due to the applied gate field. Because of the confinement of motion in the direction normal to the surface,

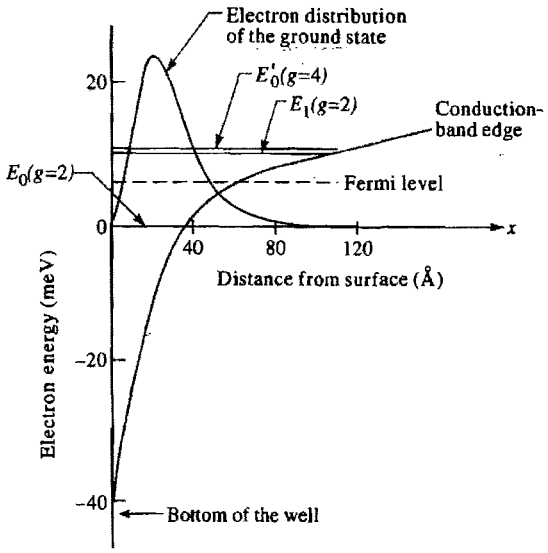


Figure 4.18. An example of quantum-mechanically calculated band bending and energy levels of inversion-layer electrons near the surface of an MOS device. The ground state is about 40 meV above the bottom of the conduction band at the surface. The dashed line indicates the Fermi level for 10^{12} electrons/cm^2 in the inversion layer. (After Stern and Howard, 1967.)

inversion-layer electrons must be treated quantum-mechanically as a 2-D gas (Stern and Howard, 1967), especially at high normal fields. Thus the energy levels of the electrons are grouped in discrete *subbands*, each of which corresponds to a quantized level for motion in the normal direction, with a continuum for motion in the plane parallel to the surface. An example of the quantum-mechanical energy levels and band bending is shown in Fig. 4.18. The electron concentration peaks below the silicon–oxide interface and goes to nearly zero at the interface, as dictated by the boundary condition of the electron wave function. This is in contrast to the classical model in which the electron concentration peaks at the surface, as shown in Fig. 4.19. Quantum-mechanical behavior of inversion-layer electrons affects MOSFET operation in two ways. ***First, at high fields, threshold voltage becomes higher, since more band bending is required to populate the lowest subband at some energy above the bottom of the conduction band. Second, once the inversion layer forms below the surface, it takes a higher gate-voltage overdrive to produce a given level of inversion charge density***. In other words, the effective gate oxide thickness is slightly larger than the physical thickness. This reduces the transconductance and the current drive of a MOSFET.

4.2.4.1 Triangular Potential Approximation for the Subthreshold Region

A full solution of the silicon inversion layer involves numerically solving coupled Poisson's and Schrödinger's equations self-consistently (Stern and Howard, 1967).

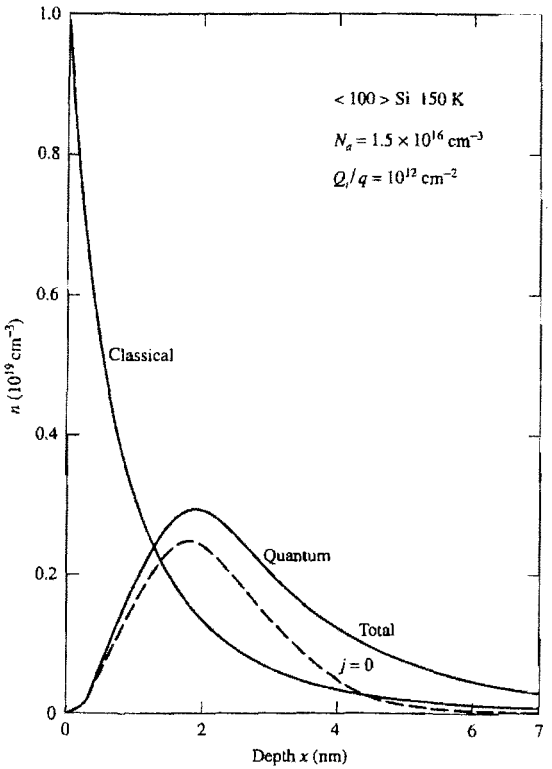


Figure 4.19. Classical and quantum-mechanical electron density versus depth for a ⟨100⟩ silicon inversion layer. The dashed curve shows the electron density distribution for the lowest subband. (After Stern, 1974.)

Under subthreshold conditions when the inversion charge density is low, band bending is solely determined by the depletion charge. It is then possible to decouple the two equations and obtain some insight into the quantum-mechanical (QM) effect on the threshold voltage. Since the inversion electrons are located in a narrow region close to the surface where the electric field is nearly constant ($\mathscr{E}_s$), it is a good approximation to consider the potential well as composed of an infinite oxide barrier for $x<0$, and a triangular potential $V(x)=q\mathscr{E}_s x$ due to the depletion charge for $x>0$. The Schrödinger equation is solved with the boundary conditions that the electron wave function goes to zero at $x=0$ and at infinity. The solutions are Airy functions with eigenvalues E_j given by (Stern, 1972)

$$E_j = \left[\frac{3hq\mathscr{E}_s}{4\sqrt{2m_x}}\left(j+\frac{3}{4}\right)\right]^{2/3}, \quad j=0,1,2,..., \tag{4.46}$$

where $h=6.63\times10^{-34}$ J-s is Planck's constant, and m_x is the effective mass of electrons in the direction of confinement. Note that MKS units are used throughout this subsection (e.g., length must be in meters, not centimeters). The average distance from the surface for electrons in the j th subband is given by

$$x_j = \frac{2E_j}{3q\mathscr{E}_s}. \tag{4.47}$$

For silicon in the ⟨100⟩ direction, there are two groups of subbands, or *valleys*. The lower valley has a twofold degeneracy ($g=2$) with $m_x=m_t=0.92m_0$, where $m_0=9.1\times10^{-31}$ kg is the free-electron mass. These energy levels are designated as E_0, E_1, E_2 .... The higher valley has a fourfold degeneracy ($g'=4$) with $m'_x=m_t=0.19m_0$. The energy levels are designated as $E'_0, E'_1, E'_2, \ldots$. Note that

$$E'_j = \left[\frac{3hq\mathscr{E}_s}{4\sqrt{2m'_x}}\left(j+\frac{3}{4}\right)\right]^{2/3}, \quad j=0,1,2,.... \tag{4.48}$$

At room temperature, several subbands in both valleys are occupied near threshold, with a majority of the electrons in the lowest subband of energy E_0 above the bottom of the conduction band. From Appendix 12, the total inversion charge per unit area is expressed as (Stern and Howard, 1967)

$$Q_i^{QM} = \frac{4\pi qkT}{h^2}\left(gm_t\sum_j \ln(1+e^{(E_f-E'_c-E_j)/kT}) + g'(m_lm_t)^{1/2}\sum_j \ln(1+e^{(E_f-E'_c-E'_j)/kT})\right), \tag{4.49}$$

where $m_t=0.19m_0$ and $(m_lm_t)^{1/2}=0.42m_0$ are the density-of-states effective masses of the two valleys, and $E_f-E'_c$ is the difference between the Fermi level and the bottom of the conduction band at the surface. It is shown in Appendix 12 that in the subthreshold region, Eq. (4.49) can be simplified to

$$Q_i^{QM} = \frac{4\pi qkTn_i^2}{h^2N_cN_a}\left(2m_t\sum_j e^{-E_j/kT} + 4(m_lm_t)^{1/2}\sum_j e^{-E'_j/kT}\right)e^{q\psi_s/kT}, \tag{4.50}$$

where N_c is the effective density of states in the conduction band.

4.2.4.2 Threshold-Voltage Shift Due to Quantum Effect

When $\mathscr{E}_s < 10^4$–10^5 V/cm at room temperature, both the lowest energy level E_0 and the spacings between the subbands are comparable to or less than kT. A large number of subbands are occupied. It is shown in Appendix 12 that in this case, Q_i^{QM} is essentially the same as the classical inversion charge density per unit area given by Eq. (3.36) for the subthreshold region,

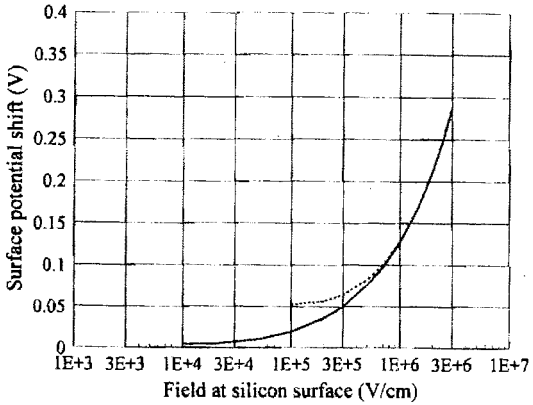


Figure 4.20. Additional band bending $\Delta\psi_s^{QM}$ (over the classical $2\psi_B$ value) required for reaching the threshold condition as a function of the surface electric field. The dotted curve is calculated by keeping only the lowest term (twofold degeneracy) in Eq. (4.50).

$$Q_i = \frac{kTn_i^2}{\mathscr{E}_s N_a} e^{q\psi_s/kT}. \tag{4.51}$$

(The expression has been generalized to cover nonuniformly doped cases where $\mathscr{E}_s$ is the electric field at the surface and N_a is the doping concentration at the edge of the depletion layer.) When $\mathscr{E}_s > 10^5$ V/cm, however, the subband spacings become greater than kT and Q_i^{QM} is significantly less than Q_i. ***The Q_i^{QM}–ψ_s curve [Eq. (4.50)] exhibits a positive parallel shift with respect to the classical Q_i–ψ_s curve [Eq. (4.51)] on a semilogarithmic scale, which means that additional band bending is required to achieve the same inversion charge per unit area as the classical value.*** The classical threshold condition, $\psi_s = 2\psi_B$, should therefore be modified to $\psi_s = 2\psi_B + \Delta\psi_s^{QM}$, where $Q_i^{QM}(\psi_s = 2\psi_B + \Delta\psi_s^{QM}) = Q_i(\psi_s = 2\psi_B)$. From this definition,

$$\Delta\psi_s^{QM} = \frac{kT}{q}\ln\left(\frac{Q_i(\psi_s = 0)}{Q_i^{QM}(\psi_s = 0)}\right) \tag{4.52}$$

can be evaluated from the preexponential factors in Eqs. (4.51) and (4.50). Figure 4.20 shows the calculated $\Delta\psi_s^{QM}$ as a function of $\mathscr{E}_s$. Beyond 10^6 V/cm, only the lowest subband is occupied by electrons and Eq. (4.52) becomes

$$\Delta\psi_s^{QM} \approx \frac{E_0}{q} - \frac{kT}{q}\ln\left(\frac{8\pi q m_t \mathscr{E}_s}{h^2 N_c}\right), \tag{4.53}$$

which is plotted as the dotted curve in Fig. 4.20. Knowing $\Delta\psi_s^{QM}$, one can easily calculate the threshold voltage shift due to the quantum effect:

$$\Delta V_t^{QM} = \frac{dV_{gs}}{d\psi_s}\Delta\psi_s^{QM} = m\Delta\psi_s^{QM}, \tag{4.54}$$

where $m = 1 + (3t_{ox}/W_{dm})$ as before.

As an example, consider a 50 nm MOSFET with a uniform doping of $N_a = 3 \times 10^{18}$ cm^{-3}, which gives $W_{dm} = 20$ nm for control of short-channel effects. For this device, $\mathscr{E}_s \approx 10^6$ V/cm, so $\Delta\psi_s^{QM} = 0.13$ V from Fig. 4.20. If $m = 1.3$, then $\Delta V_t^{QM} = 0.17$ V, resulting in a much higher threshold voltage than the classical value. A retrograde doping profile not only reduces the depletion charge density (for a given W_{dm}) but also lowers the surface field hence ΔV_t^{QM}.

4.2.4.3 Quantum Effect on Inversion-Layer Depth

After strong inversion, the inversion charge density builds up rapidly and the triangular potential-well model is no longer valid. If the separation between the minimum energies of the lowest and the first excited subbands is large enough that only the lowest subband is populated, a variational approach leads to an approximate expression for the average distance of electrons from the surface (Stern, 1972):

$$x_{av}^{QM} = \left(\frac{9\epsilon_{si}h^2}{16\pi^2 m_x q Q^*}\right)^{1/3}, \tag{4.55}$$

where $Q^* = Q_d + \frac{11}{32}Q_i$ is a combination of the depletion and inversion charge per unit area in the channel. In general, the solution must be obtained numerically. Figure 4.21 shows a comparison of the classical and QM inversion-layer depths versus the effective normal field defined in Eq. (3.51) (Ohkura, 1990). The QM value is consistently larger than the classical value by about 10–12 Å for a wide range of channel doping (uniform) and effective fields. This degrades the inversion layer capacitance, $-dQ_i/d\psi_s = \epsilon_{si}/x_{av}$ (Section 3.1.6.2),[6] and therefore the inversion charge component of the gate capacitance, $-dQ_i/dV_{gs} = \epsilon_{ox}/t_{inv}$. ***Effectively, the quantum-mechanical effect adds $\Delta t_{ox} = (\epsilon_{ox}/\epsilon_{si})\Delta x_{av} = (x_{av}^{QM} - x_{av}^{CL})/3$ or about 3–4 Å to t_{inv},*** causing lower current drive and transconductance in thin-oxide MOSFETs.

4.2.5 Discrete Dopant Effects on Threshold Voltage

As CMOS devices are scaled down, the number of dopant atoms in the depletion region of a minimum geometry device decreases. Due to the discreteness of atoms, there is a statistical random fluctuation of the number of dopants within a given volume around its average value. For example, in a uniformly doped $W = L = 0.1$-μm nMOSFET, if $N_a = 10^{18}$cm^{-3} and $W_{dm} = 350$ Å, the average number of acceptor atoms in the depletion region is $N = N_a LW W_{dm} = 350$. The actual number fluctuates from device to device with a standard deviation $\sigma_N = \langle(\Delta N)^2\rangle^{1/2} = N^{1/2} = 18.7$, which is a significant fraction of the average number N. Since the threshold voltage of a MOSFET depends on the charge of ionized dopants in the depletion region, this translates into a threshold-voltage fluctuation which could affect the operation of VLSI circuits.

[6] Strictly speaking, the x_{av} in the capacitance is the center-of mass of the differential inversion charge responding to a differential change of ψ_s. Here we neglect the subtle difference.

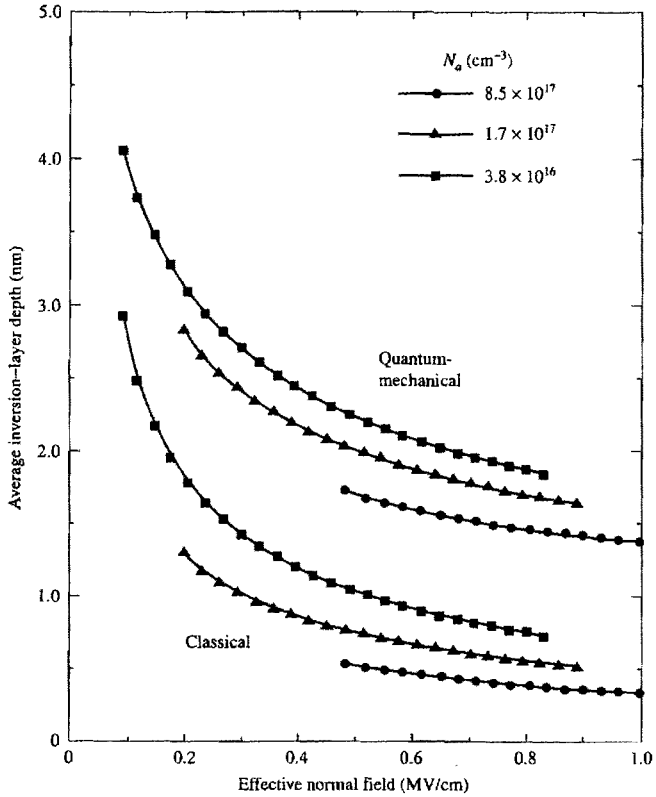


Figure 4.21. Calculated QM and classical inversion-layer depth versus effective normal field for several uniform doping concentrations. (After Ohkura, 1990.)

4.2.5.1 A Simple First-Order Model

To estimate the effect of depletion charge fluctuation on threshold voltage, we consider a small volume $dx\,dy\,dz$ at a point (x, y, z) in the depletion region of a uniformly doped (N_a) MOSFET. The x-axis is in the depth direction, the y-axis in the length direction, and the z-axis in the width direction. The average number of dopant atoms in this small volume is $N_a\,dx\,dy\,dz$. The actual number fluctuates around this value with a standard deviation of $\sigma_{dN} = (N_a\,dx\,dy\,dz)^{1/2}$. This fluctuation can be thought of as a small delta function of nonuniform doping (either positive or negative) at (x, y, z) superimposed on a uniformly doped background N_a. Here we focus on the linearly extrapolated threshold voltage V_{on}, as defined in Fig. 3.18. When there is a slight local nonuniformity of doping in either the channel-width or the channel-length direction, the first-order influence on the linear threshold voltage is through its effect on the total depletion charge integrated over the entire channel area (Nguyen, 1984). The effect of the above doping fluctuation on the

linear threshold voltage is then equivalent to that of a uniform (in W and L directions) delta-function implant of dose (number of ions per unit area) ΔD and depth x, where $\langle(\Delta D)^2\rangle^{1/2} = \sigma_{dN}/WL = (N_a\,dx\,dy\,dz)^{1/2}/WL$. The threshold-voltage shift is obtained by substituting $D_I = \Delta D$ and $x_c = x$ in Eq. (4.39) and retaining only the first-order terms in ΔD:

$$\Delta V_{on} = \frac{q\Delta D}{C_{ox}}\left(1 - x\sqrt{\frac{qN_a}{2\epsilon_{si}(2\psi_B)}}\right) = \frac{q\Delta D}{C_{ox}}\left(1 - \frac{x}{W_{dm}}\right). \tag{4.56}$$

The last expression is quite general and is applicable to a nonuniformly doped background as well. It has its roots in Eq. (4.26). The mean square deviation (variance) of threshold voltage due to the depletion charge fluctuation in $dx\,dy\,dz$ is then

$$\langle \Delta V_{on}^2 \rangle_{x,y,z} = \frac{q^2 N_a}{C_{ox}^2 L^2 W^2}\left(1 - \frac{x}{W_{dm}}\right)^2 dx\,dy\,dz. \tag{4.57}$$

Since dopant number fluctuations at various points are completely random and uncorrelated, the total mean square fluctuation of the threshold voltage is obtained by integrating Eq. (4.57) over the entire depletion region:

$$\sigma_{V_{on}}^2 = \frac{q^2 N_a}{C_{ox}^2 L^2 W^2}\int_0^W \int_0^L \int_0^{W_{dm}} \left(1 - \frac{x}{W_{dm}}\right)^2 dx\,dy\,dz. \tag{4.58}$$

It is straightforward to carry out the integration and obtain

$$\sigma_{V_{on}} = \frac{q}{C_{ox}}\sqrt{\frac{N_a W_{dm}}{3LW}}. \tag{4.59}$$

In the above 0.1-μm example, $\sigma_{V_{on}} = 17.5$ mV if $t_{ox} = 35$ Å. This is small compared with the short-channel threshold rolloff in Section 4.2.2.1, but can be significant in minimum-geometry devices, for example, in an SRAM cell.

In the above analysis, it was assumed that the surface potential is uniform in both the length and the width directions of the device. In other words, all the lumpiness due to local fluctuations of the depletion charge is smoothed out and the surface potential depends only on the average (or total) depletion charge of the device. This assumption is not valid in the subthreshold region, where current injection is dominated by the highest potential barrier in the channel rather than by the average value (Nguyen, 1984). In general, the problem needs to be solved by 3-D numerical simulations (Wong and Taur, 1993). The results indicate that in addition to the threshold fluctuations of a similar magnitude to that expected from Eq. (4.59), there is also a negative shift of the average threshold voltage, especially in the subthreshold region. This is believed to be due to the inhomogeneity of surface potential resulting from the microscopic random distribution of discrete dopant atoms in the channel.

4.2.5.2 Discrete Dopant Effects in a Retrograde-Doped Channel

Threshold voltage fluctuations due to discrete dopants are greatly reduced in a retrograde-doped channel. Consider the profile in Fig. 4.12 with $N_s = 0$, i.e., the channel

is undoped within $0 < x < x_s$. The average threshold voltage and the maximum depletion width W_{dm} are given by Eq. (4.41) and Eq. (4.42), respectively. For a small volume of dopants at (x, y, z) where $x_s < x < W_{dm}$, Eq. (4.57) still holds. The x-integral in Eq. (4.58), however, is carried out from x_s to W_{dm}, which results in

$$\sigma_{V_{on}} = \frac{q}{C_{ox}} \sqrt{\frac{N_a W_{dm}}{3LW}} \left(1 - \frac{x_s}{W_{dm}}\right)^{3/2} \tag{4.60}$$

for a retrograde-doped channel. ***In the extreme retrograde or ground-plane limit shown in Fig. 4.15, $x_s = W_{dm}$, and the threshold voltage fluctuation goes to zero.*** This is also clear from Eq. (4.43) , where the threshold voltage is essentially independent of N_a. Of course, the technological challenge is then to control the tolerance of the undoped-layer thickness x_s so that it does not introduce a different kind of threshold voltage variations.

In practice, retrograde channel doping reduces the threshold fluctuations due to discrete dopants, but does not eliminate them. For an optimally designed 25 nm MOSFET with superhalo (Taur *et al.*, 1998), a 3-D Monte-Carlo simulation has shown that the 1σ threshold voltage fluctuation due to discrete, random dopants is $10 \times W^{-1/2}$ mV where W is the device width in microns (Frank *et al.*, 1999). This is tolerable for logic devices with $W/L > 10$, but could be problematic for SRAM cell transistors which have minimum widths and require 6σ guard band for large arrays on a chip.

4.3 MOSFET Channel Length

Channel length is a key parameter in CMOS technology used for performance projection (circuit models), short-channel design, and model–hardware correlation. This section focuses on MOSFET channel length: its definition, extraction, and physical interpretation.

4.3.1 Various Definitions of Channel Length

A number of quantities, e.g., *mask length* (L_{mask}), *gate length* (L_{gate}), *metallurgical channel length* (L_{met}), and *effective channel length* (L_{eff}), have been used to describe the length of a MOSFET. Even though they are all related to each other, their relationships are strongly process-dependent.

Figure 4.22 shows schematically how various channel lengths are defined. L_{mask} is the design length on the polysilicon etch mask. It is reproduced on the wafer as L_{gate} through lithography and etching processes. Depending on the lithography and etching biases, L_{gate} can be either longer or shorter than L_{mask}. There are also process tolerances associated with L_{gate}. For the same L_{mask} design, L_{gate} may vary from chip to chip, wafer to wafer, and run to run. Although L_{gate} is an important parameter for process control and monitoring, there is no simple way of making a large number of measurements of it. Usually, L_{gate} is measured with a scanning electron microscope (SEM) and only sporadically across the wafer. There is also an uncertainty in the precise definition of L_{gate} when the polysilicon etch profile is not vertical, as to whether L_{gate} refers to the top or to the bottom dimension of the gate.

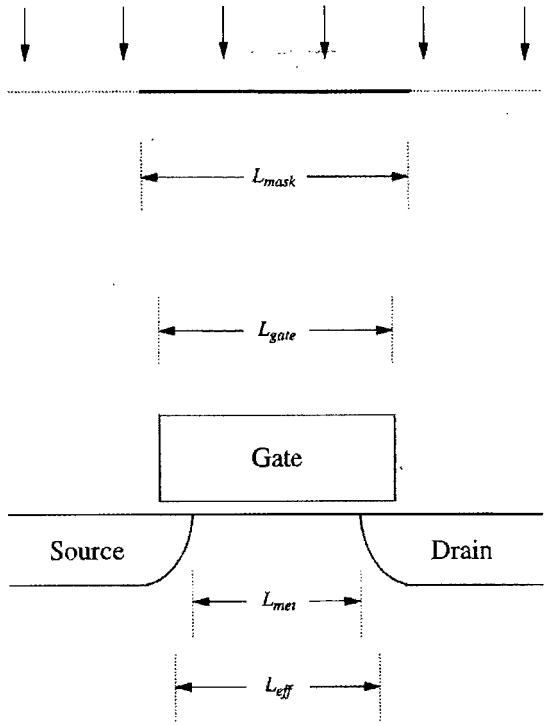


Figure 4.22. Schematic diagram showing the definitions of and relationship among the various notions of channel length. The physical interpretation of L_{eff} is examined in Section 4.3.3.

L_{met} is defined as the distance between the metallurgical junctions of the source and drain diffusions at the silicon surface. In a modern CMOS process, the source and drain regions are self-aligned to the polysilicon gate by performing the source–drain implant after gate patterning (Kerwin *et al.*, 1969). As a result, there is a close correlation between L_{met} and L_{gate}. Usually, L_{met} is shorter than L_{gate} by a certain amount due to the lateral implant straggle and the lateral source–drain diffusion in the process. Accurate physical measurement of L_{met} in actual hardware is very difficult. Normally, L_{met} is used only in 2-D models for short-channel device design. Even for that purpose, difficulties arise in defining L_{met} when dealing with a buried-channel device or a retrograde channel profile with zero surface doping, where there are no metallurgical junctions at the silicon surface.

The parameter L_{eff} is different from all other channel lengths discussed above in that it is defined through some electrical characteristics of the MOSFET device and is not a physical parameter. ***Basically, L_{eff} is a measure of how much gate-controlled current a MOSFET delivers and is therefore most suitable for circuit models.*** L_{eff} also allows for a large number of automated measurements, since it can be extracted from electrically measured terminal currents. The basis of the L_{eff} definition lies in the fact that the channel resistance of a MOSFET in the linear or low-drain bias region is proportional to the

channel length, as indicated by Eq. (3.102) (Dennard *et al.*, 1974). Further details of the definition and the extraction of L_{eff} are given in the next subsection.

For submicron CMOS technologies, it is important to distinguish among the various notions of channel length. The errors can be significant, since lithography and etching bias, junction depletion width, and lateral source–drain diffusions are all becoming an appreciable fraction of the channel length.

4.3.2 Extraction of the Effective Channel Length

As discussed in the last subsection, the effective channel length L_{eff} is defined by its proportionality to the linear or low-drain channel resistance. That is,

$$R_{ch} \equiv \frac{V_{ds}}{I_{ds}} = \frac{L_{eff}}{\mu_{eff} C_{inv} W (V_{gs} - V_{on} - mV_{ds}/2)} \tag{4.61}$$

from Eq. (3.102), where V_{on} is the linearly extrapolated threshold voltage and μ_{eff} is the effective mobility. C_{inv} contains the inversion-layer capacitance effect; μ_{eff} is a weak function of V_{gs}. For different L_{mask}, L_{eff} differs but is assumed to be related to L_{mask} by a constant *channel length bias* ΔL:

$$L_{eff} = L_{mask} - \Delta L. \tag{4.62}$$

All the lithography and etch biases as well as the lateral source–drain implant straggle and diffusion are lumped into ΔL. The assumption that the channel length bias is constant is a reasonable one when the channel length is not too short. However, ΔL can be linewidth-dependent when L_{mask} approaches the resolution limit of the lithography tool used in the process. This issue will be addressed later.

In the simplest scheme of channel-length extraction (Dennard *et al.*, 1974), R_{ch} is measured for a set of devices with different L_{mask}. Based on Eq. (4.61) and Eq. (4.62), a plot of R_{ch} for a given V_{gs} versus L_{mask} should yield a straight line whose intercept with the x-axis gives ΔL and therefore L_{eff}. In practice, however, two issues must be addressed for short-channel devices. The first one is the source–drain series resistance. The second one is the short-channel effect (SCE), which causes V_{on} in Eq.(4.61) to depend on L_{mask}.

4.3.2.1 Channel-Resistance Method

The effect of source–drain resistance is examined using the equivalent circuit in Fig. 4.23. A source resistance R_s and a drain resistance R_d are assumed to connect an *intrinsic MOSFET* to the external terminals where voltages V_{ds} and V_{gs} are applied. The internal voltages are V'_{ds} and V'_{gs} for the intrinsic MOSFET. One can write the following relations:

$$V'_{ds} = V_{ds} - (R_s + R_d) I_{ds} \tag{4.63}$$

and

$$V'_{gs} = V_{gs} - R_s I_{ds}. \tag{4.64}$$

As shown in Fig. 4.23, the intrinsic part of an actual device with parasitic resistance is equivalent to an intrinsic MOSFET with a grounded source, with V'_{gs} and V'_{ds} at the

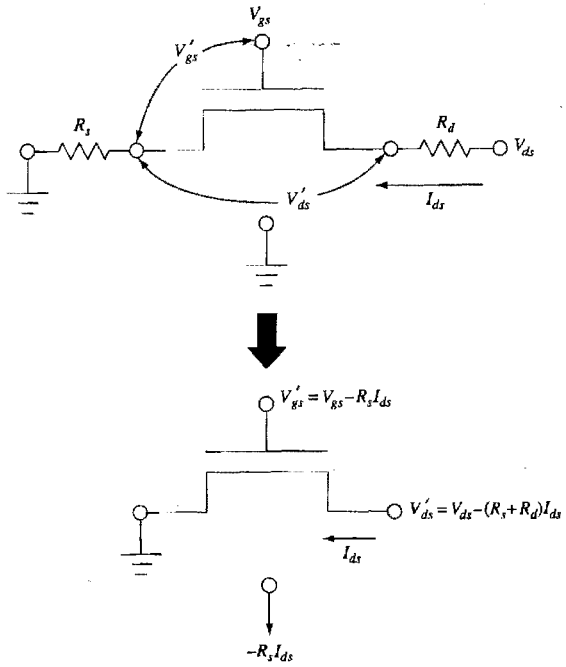


Figure 4.23. Equivalent circuit of MOSFET with source and drain series resistance. The intrinsic part of the top circuit is equivalent to the bottom circuit with redefined terminal voltages.

gate and the drain terminals, and with a reverse bias $-R_s I_{ds}$ on the substrate. Based on Eq. (4.61), but with redefined voltage symbols on the intrinsic nodes, the channel resistance of the intrinsic device is given by

$$R_{ch} \equiv \frac{V'_{ds}}{I_{ds}} = \frac{L_{eff}}{\mu_{eff} C_{inv} W (V'_{gs} - V'_{on} - mV'_{ds}/2)}, \tag{4.65}$$

where V'_{on} is the linear threshold voltage with the reverse bias on the substrate. It is related to the zero-substrate-bias threshold voltage V_{on} by

$$V'_{on} = V_{on} + (m - 1) R_s I_{ds}, \tag{4.66}$$

where $m - 1$ is the substrate sensitivity [Eq. (4.36)]. In a normal CMOS process, the source and drain regions are symmetrical, and therefore $R_s = R_d = R_{sd}/2$, where R_{sd} is the total source–drain parasitic resistance. Using Eqs. (4.62)–(4.66), one can write the externally measured total device resistance as

$$R_{tot} \equiv \frac{V_{ds}}{I_{ds}} = R_{sd} + R_{ch} = R_{sd} + \frac{L_{mask} - \Delta L}{\mu_{eff} C_{inv} W (V_{gs} - V_{on} - mV_{ds}/2)}. \tag{4.67}$$

Here all the internal voltages have been replaced by the voltages at the external terminals, since $V'_{gs} - V'_{on} - mV'_{ds}/2 = V_{gs} - V_{on} - mV_{ds}/2$ from Eqs. (4.63), (4.64), and (4.66). Note that V_{on} is defined in terms of the intrinsic device, i.e., the threshold that would be obtained from linear extrapolation if there were no parasitic resistances.

For a set of devices with different L_{mask} but the same W, the parameters R_{sd}, ΔL, and C_{inv} are the same within process tolerances. It is also assumed that μ_{eff} does not change with channel length. The linear threshold voltage V_{on}, however, does depend on channel length because of short-channel effects. When comparing R_{tot} of devices with different L_{mask}, therefore, it is important to measure V_{on} for each device and adjust V_{gs} so that the gate overdrive $V_{gs} - V_{on}$ is the same from device to device. ***A plot of R_{tot} (at small V_{ds}) versus L_{mask} for a given $V_{gs} - V_{on}$ will then yield a straight line that passes through the point (ΔL, R_{sd})***. An example is shown in Fig. 4.24. The slope of the line depends on the

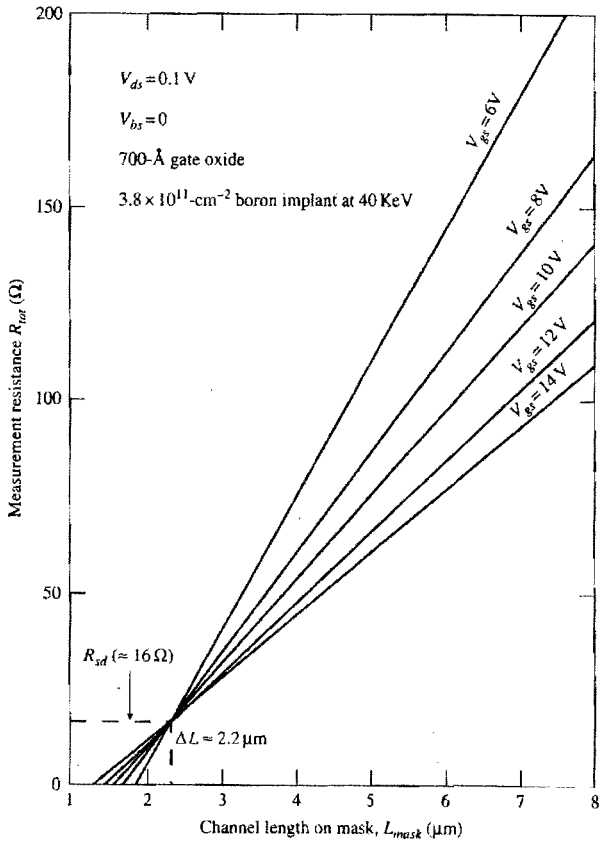


Figure 4.24. Measured R_{tot} at a low drain voltage versus L_{mask} for several different values of $V_{gs} - V_{on}$. The common intercept determines both ΔL and R_{sd}. (After Chern *et al.*, 1980.)

specific value of the gate overdrive. ΔL and R_{sd} are determined by the common intercept of several lines, each for a different $V_{gs} - V_{on}$ (Chern *et al.*, 1980).

4.3.2.2 Shift-and-Ratio Method

Despite the simplicity of the channel-resistance method described above, two main issues remain. First, it is not always straightforward to find the intrinsic V_{on} of short-channel devices. The presence of R_{sd} adds considerable difficulty in the usual linear extrapolation of V_{on} from the measured I_{ds}–V_{gs} curve (Sun *et al.*, 1986). Typically, one tends to underestimate V_{on} in short-channel devices, as the degradation of I_{ds} by R_{sd} is more severe at higher currents. This introduces errors in channel-length extraction. The problem is further aggravated by a strong dependence of mobility on gate voltage, for example, in low-temperature and/or 0.1-μm MOSFETs. The second problem with the resistance method is that the R_{tot}-versus-L_{mask} lines for different gate overdrives may not intersect at a common point. Significant errors may result if only a limited number of $V_{gs} - V_{on}$ are investigated.

An improved channel-length extraction algorithm, called the *shift-and-ratio* (S&R) method, is able to circumvent the above problems (Taur *et al.*, 1992). This method is based on the same channel-resistance concept described above. It starts with a generalization of Eq. (4.67) to the form

$$R^i_{tot}(V_{gs}) = R_{sd} + L^i_{eff} f(V_{gs} - V^i_{on}), \tag{4.68}$$

where f is a general function of gate overdrive common to all the measured devices. The superscript i denotes the ith device, with an unknown effective channel length $L^i_{eff} = L^i_{mask} - \Delta L$ and linear threshold voltage V^i_{on}. The key assumption behind Eq. (4.68) is that the effective mobility μ_{eff} is a common function of $V_{gs} - V_{on}$ for all the measured devices. This is a reasonable assumption in view of Eq. (3.54) and Eq. (3.55).

The task is to calculate R_{sd}, L^i_{eff}, and V^i_{on} in Eq. (4.68) from the measured data on $R^i_{tot}(V_{gs})$. The S&R algorithm simplifies the procedure by differentiating Eq. (4.68) with respect to V_{gs}. Since the parasitic resistance R_{sd} is either independent or a weak function of V_{gs}, its derivative can be neglected:

$$S^i(V_{gs}) \equiv \frac{dR^i_{tot}}{dV_{gs}} = L^i_{eff} \frac{df(V_{gs} - V^i_{on})}{dV_{gs}}. \tag{4.69}$$

Here df/dV_{gs} is also a general function of gate overdrive common to all the devices measured. An important benefit of working with the derivatives is that R_{sd} drops completely out of the picture, so it does not matter if R_{sd} varies from device to device as long as it is constant. An algorithm has been developed in the S&R method to take the ratio of $S^i(V_{gs})$ between two different devices – typically one long and one short, by shifting one along V_{gs} with respect to the other. The ratio becomes nearly constant, i.e., independent of V_{gs}, when the shift equals the difference between V^i_{on} of the two devices. L^i_{eff} is then determined from the ratio of the S functions at such a shift (Taur *et al.*, 1992).

4.3.3 Physical Meaning of Effective Channel Length

This subsection examines the physical meaning of L_{eff} extracted from electrically measured terminal currents. The effective channel length is defined through the linear channel resistance by Eq. (4.61). This equation is derived for long-channel devices and is not strictly valid for short-channel devices. By its definition, L_{eff} represents a measure of the effective gate-controlled resistance of the device and is not associated with any fixed physical quantity. When the channel profile is reasonably uniform and the source–drain doping is not too graded, L_{eff} is approximately equal to L_{met} (Laux, 1984). In general, however, one cannot take $L_{eff} = L_{met}$ for granted. The more graded (laterally) the source and drain profiles are, the longer L_{eff} is over L_{met}. This can be understood in terms of the spatial dependence of channel sheet resistivity discussed below.

4.3.3.1 Sheet Resistivity in Short-Channel Devices

Equation (4.61) implicitly assumes that the sheet resistivity, ρ_{ch} given by Eq. (3.103), is spatially uniform in both the MOSFET width and length directions. If the device is wide enough, ρ_{ch} can be considered uniform in that direction. However, the variation of ρ_{ch} in the length direction cannot be ignored in a short-channel device. From Eq. (3.8),

$$I_{ds} = -\mu_{eff} W Q_i(y) \frac{dV}{dy}, \tag{4.70}$$

where $V(y)$ is the quasi-Fermi level at a point y along the channel length direction. I_{ds} is a constant independent of y as required by current continuity. One can define a laterally varying sheet resistivity as

$$\rho_{ch}(y) = \frac{dV/dy}{I_{ds}/W} = \frac{1}{-\mu_{eff} Q_i(y)}. \tag{4.71}$$

Note that $Q_i < 0$ for nMOSFETs. This expression is valid as long as the current flow is largely parallel to the y-direction and the equipotential contours are perpendicular to the silicon surface. The total resistance is given by

$$\frac{V_{ds}}{I_{ds}} = \frac{1}{W} \int_{S-D} \rho_{ch}(y) dy, \tag{4.72}$$

where the integration is carried out from the heavily doped source region to the heavily doped drain region.

Figure 4.25 plots $\rho_{ch}(y)$ calculated from a 2-D device simulator versus distance at different gate voltages for both an abrupt and a graded source–drain (Taur *et al.*, 1995b). The area under each curve gives the total source-to-drain resistance at that gate voltage. In Fig. 4.25(a) for an infinitely abrupt (laterally) source-drain junction, the sheet resistivity is modulated by the gate voltage inside the (metallurgical) channel and independent of the gate voltage outside the (metallurgical) channel. However, in contrast to a long-channel device, $\rho_{ch}(y)$ is highly nonuniform, with a peak near the middle of the channel and decreasing toward the edges. This is due to SCEs from the source–drain fields, which help lower the potential barrier near the junctions and raise the local inversion

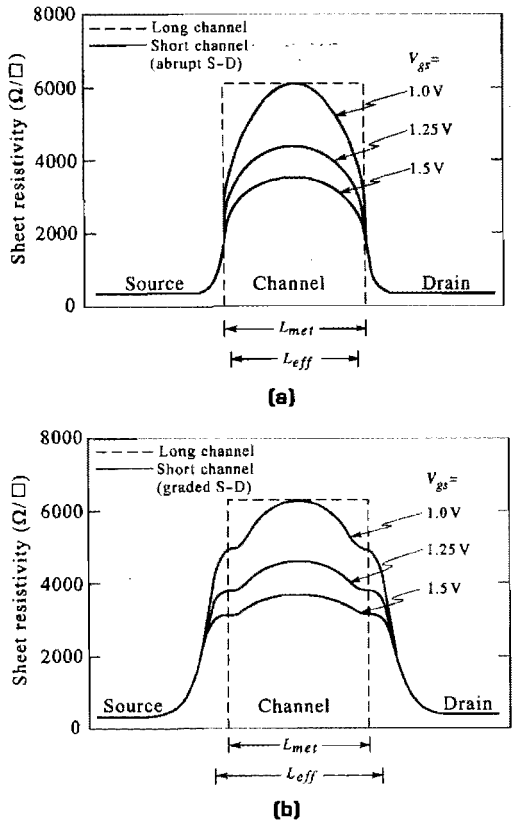


Figure 4.25. Simulated channel sheet resistivity at three different gate voltages versus distance from source to drain of an $L_{met} = 0.10$-μm MOSFET. The curves in (a) are for an infinitely abrupt (laterally) source–drain which yields $L_{eff} = 0.091$ μm. The curves in (b) are for a graded (lateral straggle $\sigma_L = 165$ Å) source–drain which yields $L_{eff} = 0.124$ μm. In both cases, the dashed lines represent the ideal, uniform-sheet resistivity of a scaled long-channel device. (After Taur *et al.*, 1995b.)

charge density there (Wordeman *et al.*, 1985). This effect is more pronounced at low gate voltages near threshold. The resulting L_{eff} extracted by the S&R method is slightly shorter than L_{met}.

Figure 4.25(b) shows similar plots for the same L_{met}, but with a finite lateral source–drain gradient. $\rho_{ch}(y)$ again is nonuniform inside the channel, being modulated by the gate voltage. In this case, however, a nonnegligible portion of the sheet resistivity outside the metallurgical channel is also gate-voltage-dependent. This is because of accumulation (Section 2.3.1) or gate modulation of the series resistance associated with the finite source–drain doping gradient. Since, according to the L_{eff} definition in Eq. (4.68), any part of the sheet resistivity that is gate-voltage dependent contributes to the effective

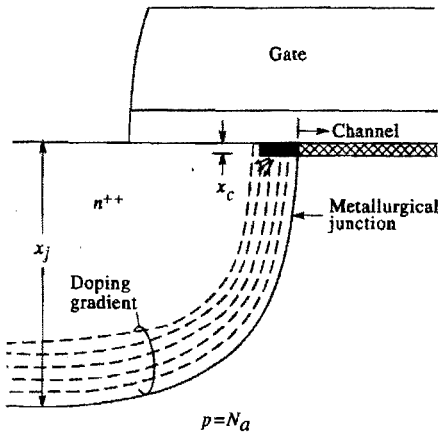


Figure 4.26. Schematic diagram showing doping distribution and current flow pattern near the end of the channel and the beginning of the source or drain. The dashed lines are contours of constant donor concentration, i.e., constant resistivity. The dark region represents the accumulation layer. (After Ng and Lynch, 1986.)

channel length, the extracted L_{eff} is substantially longer than L_{met}. At the same time, the extracted R_{sd}, which represents the constant part of the resistance in Eq. (4.68), only accounts for a portion of the series resistance outside the metallurgical channel.

4.3.3.2 Gate-Modulated Accumulation-Layer Resistance

Because of the finite lateral gradient of source–drain doping in practical devices, current injection from the surface inversion layer into the bulk source–drain region does not occur immediately at the metallurgical junction. When the gate voltage is high enough to turn on the MOSFET channel, an n^+ surface accumulation layer is also formed in the gate-to-source or -drain overlap region, as shown schematically in Fig. 4.26 (Ng and Lynch, 1986). Near the metallurgical junction and away from the surface, the donor concentration (also compensated by the p-type background) is low and the conductivity of the accumulation layer is higher than that of the bulk source–drain. As a result, current flow stays in the accumulation layer near the surface. This continues until the source–drain doping becomes high enough that the bulk conductance exceeds that of the accumulation layer. The point or region of current injection into the bulk depends on the lateral source–drain doping gradient. The more graded the profile is, the farther away the injection point is from the metallurgical junction.

The sheet resistivity of the accumulation layer can be estimated by applying Eq. (2.195) to the gate-to-source–drain overlap region:

$$V_{gs} = V_{fb} + \psi_s - \frac{Q_{ac}}{C_{ox}}, \tag{4.73}$$

where $Q_{ac} < 0$ is the accumulation charge (electrons) per unit area induced by the gate field, ψ_s is the band bending at the surface with respect to the bulk n-type region, and V_{fb} is the flat-band voltage largely determined by the work-function difference between the gate electrode and the n-type silicon. For an n^+-polysilicon-gated nMOSFET, $V_{fb} = -E_g/2q + \psi_B$, where ψ_B is given by Eq. (2.48) in terms of the local n-type doping concentration. The band bending in accumulation is approximately given by the distance between the n-type Fermi level and the conduction-band edge, i.e., $\psi_s \approx E_g/2q - \psi_B$. Therefore, V_{fb} and ψ_s in Eq. (4.73) nearly cancel each other and one obtains $V_{gs} \approx -Q_{ac}/C_{ox}$. The sheet resistivity of the accumulation layer is then

$$\rho_{ac} = \frac{1}{\mu_{ac}|Q_{ac}|} = \frac{1}{\mu_{ac} C_{ox} V_{gs}}, \tag{4.74}$$

where μ_{ac} is the average electron mobility in the accumulation layer (Sun and Plummer, 1980).

4.3.3.3 Interpretation of L_{eff} in Terms of Current Injection Points

The dependence of ρ_{ac} on V_{gs} in Eq. (4.74) is too similar to that of ρ_{ch} in Eq. (3.103) to allow separation of the accumulation-layer resistance from the channel resistance. ***The region where the current flows predominantly in the accumulation layer is therefore considered as a part of L_{eff}.*** The physical interpretation of L_{eff} in terms of injection points where the sheet resistivity of bulk source–drain equals that of the accumulation layer is consistent with 2-D device simulation results (Taur *et al.*, 1995b). For more graded (laterally) source–drain profiles, the injection points hence L_{eff} can be gate voltage dependent. At low gate overdrives, the injection point is closer to the metallurgical junction edge. As the gate voltage increases, the injection point moves out toward the more heavily doped source–drain region, resulting in a longer L_{eff}.

4.3.3.4 Implications for Short-Channel Effects

The fact that L_{eff} can be much longer than L_{met} has significant implications for the short-channel V_t rolloff curves. Figure 4.27 shows the low-drain threshold voltage rolloff versus L_{eff} for several different source–drain doping gradients. The abrupt doping profile has the best short-channel effect. As the lateral straggle σ_L increases, the short-channel effect becomes progressively worse. This can be understood from the above interpretation of L_{eff}. Current injection from the surface layer takes place at a certain source–drain doping concentration, e.g., 10^{19} cm^{-3} for nMOSFETs. For a given L_{eff}, the distance between the points where the doping concentration falls to 10^{19} cm^{-3} is fixed. The portion of the source–drain doping below 10^{19} cm^{-3} penetrates into the L_{eff} region from both ends. The more graded the source–drain profile is, the deeper such an n-type doping tail penetrates into the channel and compensates or reverses the p-type doping inside the channel. This is detrimental to the short-channel effect, as the edge regions become more easily depleted and inverted by the gate field (opposite to the halo effect). ***It is therefore very important to reduce the width of the (laterally) graded source–drain region as the channel length is scaled down.***

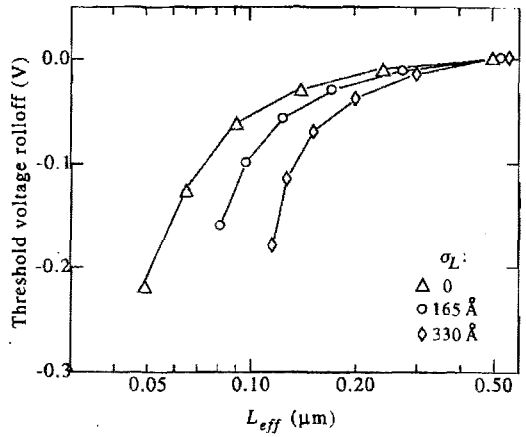


Figure 4.27. Simulated short-channel threshold rolloff versus L_{eff} for three different lateral source–drain doping gradients. On each curve, the points are for L_{met} = 0.05, 0.07, 0.10, 0.15, 0.25, and 0.50 μm. (After Taur *et al.*, 1995b.)

4.3.4 Extraction of Channel Length by *C–V* Measurements

In an entirely different approach, another type of channel length has been extracted from the measured *C–V* data of a series of MOSFETs with different L_{mask} (Sheu and Ko, 1984). Capacitance measurements in general are more difficult to perform, as they require specially designed test sites. It is by no means straightforward to interpret the capacitively measured channel length and apply it to circuit models for current calculations.

The capacitive extraction of channel length is based on the fact that when a MOSFET is turned on, the intrinsic gate-to-channel capacitance is proportional to the channel length:

$$C_{gc} = C_{inv} W L_{cap}. \tag{4.75}$$

Here L_{cap} is the capacitively defined channel length, which may or may not be the same as L_{eff} or L_{met}. The gate-to-channel capacitance is usually measured in a split *C–V* setup that separates the majority-carrier response from the minority-carrier response, as shown in the inset of Fig. 4.28. The total measured capacitance consists of both the intrinsic gate-to-channel capacitance and a parasitic overlap capacitance from the gate to source–drain which is independent of channel length:

$$C_{tot} = C_{gc} + 2C_{ov} = C_{inv} W L_{cap} + 2C_{ov}. \tag{4.76}$$

Here C_{ov} is the overlap capacitance per gate edge (see Fig. 5.19). Typical examples of C_{tot}–V_{gs} curves are shown in Fig. 4.28 (Guo *et al.*, 1994). Using a large-area MOS capacitor, one can easily calibrate C_{inv}, taking all the polysilicon depletion and inversion-layer quantum effects into account. To find out L_{cap}, it is critical to determine what $2C_{ov}$ to subtract from the measured C_{tot}. In principle, $2C_{ov}$ in Eq. (4.76) is the parasitic

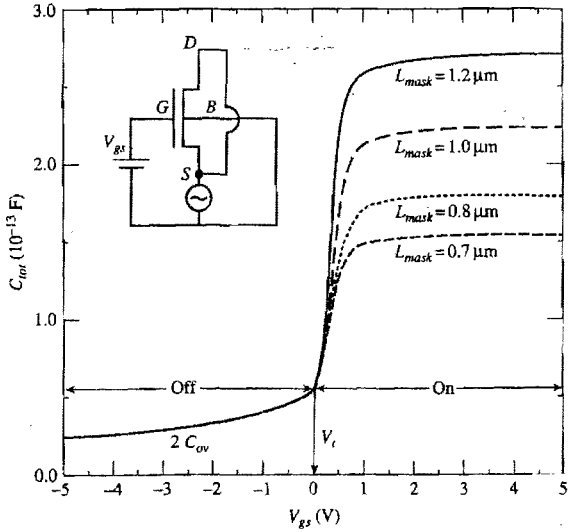


Figure 4.28. Example of measured capacitance from gate to source–drain versus gate voltage for MOSFETs of different mask lengths. The inset shows the split *C–V* measurement setup. (After Guo *et al.*, 1994.)

capacitance at a gate voltage when the MOSFET is on and C_{gc} is given by Eq. (4.75). In practice, $2C_{ov}$ cannot be separated from C_{gc}, since, unlike channel resistance, channel capacitance does not vary significantly with gate voltage once the device is turned on. What is usually done is to take $2C_{ov}$ as the measured capacitance when the MOSFET is off. However, from Fig. 4.28 it is clear that $2C_{ov}$ varies with the gate voltage (Oh *et al.*, 1990). There is no guarantee that $2C_{ov}$ in the off state is the same as $2C_{ov}$ in the on state. If $2C_{ov}$ is taken as the capacitance right below the threshold voltage, it will contain an unwanted inner-fringe term that is absent when the conducting channel is formed. If $2C_{ov}$ is taken at a negative gate voltage where the substrate is accumulated to eliminate the inner-fringe component, the lightly doped source–drain in the direct overlap region will be depleted (Sheu and Ko, 1984). Any such error in $2C_{ov}$ translates into a large error in L_{cap} when dealing with short-channel devices having small intrinsic capacitances.

A better interpretation of the capacitively extracted channel length is in terms of the gate length, L_{gate} (Fig. 4.22), since as the gate voltage varies, the same amount of charge per unit area is induced at the silicon surface whether it is in the inversion channel or in the source–drain overlap region under the gate. In other words, as far as the capacitance is concerned, the direct overlap length should be lumped into the channel length. This also circumvents the problem with the inner-fringe component mentioned above. One still needs to estimate the outer fringe capacitance and subtract it from the measured capacitance. But the outer fringe is smaller and can be estimated reasonably accurately using a simple formula (Section 5.2.2).

Exercises

4.1 Apply constant-field scaling rules to the long-channel currents [Eq. (3.23) for the linear region and Eq. (3.28) for the saturation region], and show that they behave as indicated in Table 4.1.

4.2 Apply constant-field scaling rules to the subthreshold current, Eq. (3.40), and show that instead of decreasing with scaling ($1/\kappa$), it actually increases with scaling (note that $V_{gs} < V_t$ in subthreshold). What if the temperature is also scaled down by the same factor ($T \rightarrow T/\kappa$)?

4.3 Apply constant-field scaling rules to the saturation current from the $n = 1$ velocity saturation model [Eq. (3.79)] and the fully saturation-velocity limited current [Eq. (3.81)], and show that they behave as indicated in Table 4.1.

4.4 Apply generalized scaling rules to the saturation current from the $n = 1$ velocity saturation model [Eq. (3.79)], and show that it behaves as indicated in Table 4.3 (between the two limits).

4.5 Consider an n-channel MOSFET with n^+ polysilicon gate (neglect poly depletion effect). The gate oxide is 7 nm thick, and the p-type body (or substrate) has a retrograde doping as shown in Fig. 4.12 with $N_s = 0$. Take $2\psi_B = 1$ V.

(a) Choose the values of x_s and N_a such that the maximum depletion width is $W_{dm} = 0.1$ μm and the threshold voltage (at $2\psi_B$) is $V_t = 0.3$ V.

(b) Following (a), what is the body effect coefficient, m, and the inverse slope of log subthreshold current versus gate voltage (long-channel device)?

(c) Following (a), how short a channel length can the device be scaled to before short-channel effect becomes severe?

4.6 *Nonuniform V_t in the width direction.* A MOSFET is nonuniformly doped in the width direction. Part of the width (W_1) has a linear threshold voltage V_{on1} (defined in Fig. 3.18). The other part of the width (W_2) has a linear threshold voltage V_{on2}. Show that as far as the linear region characteristics are concerned, this device is equivalent to a uniform MOSFET of width $W_1 + W_2$ with a linear threshold voltage $V_{on} = (W_1 V_{on1} + W_2 V_{on2})/(W_1 + W_2)$. Ignore any fringing fields that may exist near the boundary between the two regions.

4.7 *Nonuniform V_t in the length direction.* A MOSFET is nonuniformly doped in the length direction. Part of the length (L_1) has a linear threshold voltage V_{on1}. The other part of the length (L_2) has a linear threshold voltage V_{on2}. Assume $V_{on1} \approx V_{on2}$, and consider only the first-order terms of $V_{on1} - V_{on2}$. Show that as far as the linear region characteristics are concerned, this device is equivalent to a uniform MOSFET of length $L_1 + L_2$ with a linear threshold voltage $V_{on} = (L_1 V_{on1} + L_2 V_{on2})/(L_1 + L_2)$. Ignore any fringing fields that may exist near the boundary between the two regions.

4.8 In the top equivalent circuit of Fig. 4.23, the source–drain current can be considered either as a function of the internal voltages: $I_{ds}(V'_{gs}, V'_{ds})$, or as a function of the external voltages: I_{ds} (V_{gs}, V_{ds}). The internal voltages are related to the external voltages by Eqs. (4.63) and (4.64). Show that the transconductance of the intrinsic MOSFET can be expressed as

$$g'_m \equiv \left(\frac{\partial I_{ds}}{\partial V'_{gs}}\right)_{V'_{ds}} = \frac{g_m}{1 - g_m R_s - g_{ds}(R_s + R_d)},$$

where

$$g_m \equiv \left(\frac{\partial I_{ds}}{\partial V_{gs}}\right)_{V_{ds}}$$

is the extrinsic transconductance, and

$$g_{ds} \equiv \left(\frac{\partial I_{ds}}{\partial V_{ds}}\right)_{V_{gs}}$$

is the extrinsic output conductance.

4.9 Show that in the subthreshold region and when the drain bias is low, Eq. (3.12) leads to Eq. (4.51):

$$Q_i = \frac{kTn_i^2}{\mathscr{E}_s N_a} e^{q\psi_s/kT},$$

where ψ_s is the surface potential and $\mathscr{E}_s$ is the surface electric field. This equation is more general than Eq. (3.36) since it is valid for nonuniform (vertically) dopings with N_a being the p-type concentration at the edge of the depletion layer. (Note that the factor N_a merely reflects the fact in Fig. 2.32 that the band bending ψ_s is defined with respect to the bands of the neutral bulk region of doping N_a.)

4.10 In a short-channel device or in a nonuniformly doped (laterally) MOSFET, ψ_s may vary along the channel length direction from the source to drain. Generalize the expression in Exercise 4.9 and show that

$$\frac{V_{ds}}{I_{ds}} = \frac{1}{\mu_{eff} W} \int_0^L \frac{dy}{Q_i(y)} = \frac{N_a}{\mu_{eff} W k T n_i^2} \int_0^L \mathscr{E}_s(y) e^{-q\psi_s(y)/kT} dy$$

for the subthreshold region at low drain biases. Since $\mathscr{E}_s(y) \approx [V_{gs} - V_{fb} - \psi_s(y)]/3t_{ox}$ is not a strong function of ψ_s, the exponential factor dominates. This implies that the subthreshold current is controlled by the point of highest barrier (lowest ψ_s) in the channel. It also implies that the *channel length* factor entering the subthreshold current expression is different from the *effective channel length* defined by the linear region characteristics, Eq. (4.61).

4.11 Consider a uniformly doped nMOSFET of $N_a = 10^{18}$ cm^{-3} biased at the threshold condition. Calculate the first three quantum mechanical energy levels for inversion electrons in the lower valley with an effective mass of $0.92m_0$ where m_0 is the free electron mass. Express the answers in eV.

4.12 For an nMOSFET with $t_{ox} = 10$ nm and a uniform p-type doping of 10^{17} cm^{-3}, the gate is n^+ polysilicon doped to 10^{20} cm^{-3}. Estimate the depletion layer width in the polysilicon gate at a gate voltage of 3 V.

5 CMOS Performance Factors

The performance of a CMOS VLSI chip is measured by its integration density, switching speed, and power dissipation. CMOS circuits have the unique characteristic of practically zero standby power, which enables higher integration levels and makes them the technology of choice for most VLSI applications. This chapter examines the various factors that determine the switching speed of basic CMOS circuit elements.

5.1 Basic CMOS Circuit Elements

In a modern CMOS VLSI chip, the most important function components are CMOS static gates. In gate array circuits, CMOS static gates are used almost exclusively. In microprocessors and supporting circuits of memory chips, most of the control interface logic is implemented using CMOS static gates. Static logic gates are the most widely used CMOS circuit because of their simplicity and noise immunity. This section describes basic static CMOS circuit elements and their switching characteristics.

Circuit symbols for nMOSFETs and pMOSFETs are defined in Fig. 5.1. A MOSFET is a four-terminal device, although usually only three are shown. Unless specified, the body (p-substrate) terminal of an nMOSFET is connected to the ground (lowest voltage), while the body terminal (n-well) of a pMOSFET is connected to the power supply V_{dd} (highest voltage).

5.1.1 CMOS Inverters

The most basic element of digital static CMOS circuits is a CMOS inverter. A CMOS inverter is a combination of an nMOSFET and a pMOSFET, as shown in Fig. 5.2 (Burns, 1964). The source terminal of the nMOSFET is connected to the ground, while the source of the pMOSFET is connected to V_{dd}. The gates of the two MOSFETs are tied together as the input node. The two drains are tied together as the output node. In such an arrangement, the complementary nature of n- and pMOSFETs allows one and only one transistor to be conducting in one of the two stable states. For example, when the input voltage is high or when $V_{in} = V_{dd}$, the gate-to-source voltage of the nMOSFET equals V_{dd}, which turns it on. At the same time, the gate-to-source voltage of the pMOSFET is zero, so the pMOSFET is off. The output node is then pulled down to the ground potential by currents through the conducting nMOSFET, which is referred to as the *pull-down* transistor. On

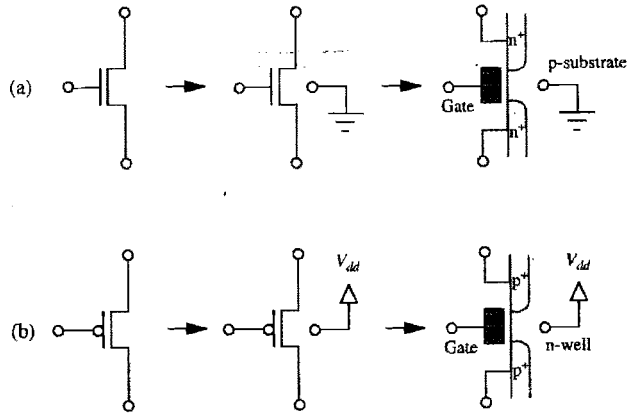


Figure 5.1. Circuit symbols and voltage terminals of (a) nMOSFET and (b) pMOSFET.

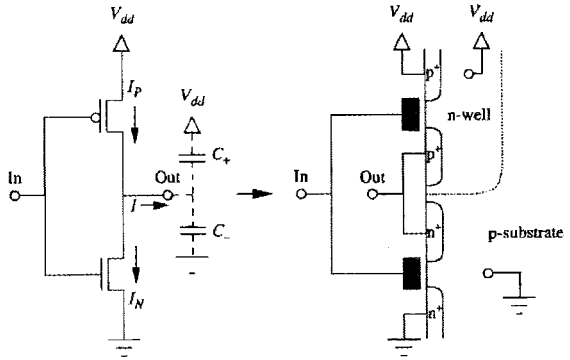


Figure 5.2. Circuit diagram and schematic cross section of a CMOS inverter.

the other hand, when the input voltage is low or when $V_{in} = 0$, the nMOSFET is off, since its gate-to-source voltage is zero. The gate-to-source voltage of the pMOSFET, however, is $-V_{dd}$, which turns it on (a negative gate voltage turns on a pMOSFET). The output node is now pulled up to V_{dd} by the conducting pMOSFET, which is referred to as the *pull-up* transistor. Since the output voltage is always opposite to the input voltage (V_{out} is high when V_{in} is low and vice versa), this circuit is called an inverter. Notice that ***since only one of the transistors is on in the steady state, there is no static current or static power dissipation***. Power dissipation occurs only during switching transients when a charging or discharging current is flowing through the circuit.

5.1.1.1 CMOS Inverter Transfer Curve

In a CMOS inverter, both the current through the nMOSFET ($I_N > 0$) and the current through the pMOSFET ($I_P > 0$) are functions of the input voltage to the gates, V_{in},

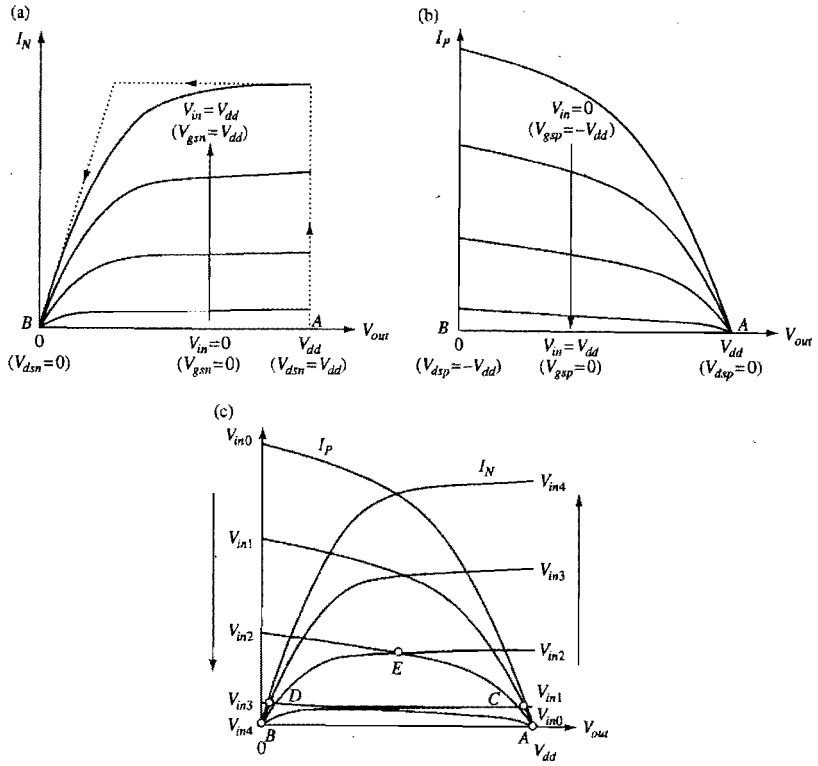


Figure 5.3. (a) nMOSFET current I_N and (b) pMOSFET current I_P in a CMOS inverter versus output node (drain) voltage for a series of input node (gate) voltages from 0 to V_{dd}. Both plots are superimposed in (c) to find the steady state points of operation (circles) given by the intersections where $I_P = I_N$ under the same V_{in} and V_{out}. The curves are labeled by the input voltages: $0 = V_{in0} < V_{in1} < V_{in2} < V_{in3} < V_{in4} = V_{dd}$, with the corresponding intercepts: A, C, E, D, B. The dotted lines in (a) depict an approximate bias point trajectory of an nMOSFET pull-down transition from A to B following an abrupt switching of V_{in} from 0 to V_{dd}. It is used later in Section 5.1.1.3 for discussion of the switching delay.

and the output node voltage, V_{out}. A typical example is shown in Fig. 5.3 where I_N and I_P are plotted versus V_{out} with V_{in} as a parameter. Note that for pMOSFET in Fig. 5.3(b), the drain to source voltage is $V_{dsp} = V_{out} - V_{dd}$, and the gate to source voltage is $V_{gsp} = V_{in} - V_{dd}$. Both are negative or zero in normal operations. Also note that I_P enters saturation softer than I_N because holes have a more gradual velocity-field relationship than electrons (Section 3.2.2.1). The net current flowing out of the inverter is given by $I = I_P - I_N$. The output node voltage increases or decreases depending on whether $I > 0$ or $I < 0$. The directions of the currents are depicted in Fig. 5.2.

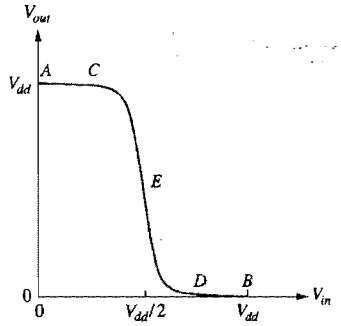


Figure 5.4. V_{out} versus V_{in} curve (transfer curve) of a CMOS inverter. Points labeled A, C, E, D, B correspond to the steady state points of operation (circles) indicated in Fig. 5.3(c).

In the steady state, $I = 0$, i.e., $I_P = I_N$. There are two points of operation where both I_P and I_N are zero: point A where $V_{in} = 0$ and $V_{out} = V_{dd}$, and point B where $V_{in} = V_{dd}$ and $V_{out} = 0$. For other values of V_{in} in between, the corresponding V_{out} is obtained from the intersection of two curves, $I_N(V_{in})$ and $I_P(V_{in})$, as shown in Fig. 5.3(c). In this way, one can construct a V_{out} versus V_{in} curve, or a *transfer curve* of the CMOS inverter in Fig. 5.4. For low values of V_{in} such as point C, V_{out} is high and the nMOSFET is biased in saturation while the pMOSFET is biased in the linear region [V_{in1} in Fig. 5.3(c)]. For high V_{in} such as point D, V_{out} is low and the nMOSFET is in the linear region while the pMOSFET is in saturation [V_{in3} in Fig. 5.3(c)]. For point E near $V_{in} = V_{dd}/2$ [V_{in2} in Fig. 5.3(c)], both devices are in saturation. It is in a transition region where V_{out} changes steeply with V_{in}.

In order for the high-to-low transition of the transfer curve to occur close to the midpoint, $V_{in} = V_{dd}/2$, it is desired for I_P and I_N to be nearly symmetrical, as illustrated in the example in Fig. 5.3. This requires the threshold voltages of the n- and pMOSFETs to be symmetrically matched. In addition, since the pMOSFET current per width, $I_p = I_P / W_p$, is inherently lower than that of the nMOSFET, $I_n = I_N / W_n$, the device width ratio in a CMOS inverter should be

$$\frac{W_p}{W_n} = \frac{I_n}{I_p}, \tag{5.1}$$

such that $I_P \approx I_N$. In the long-channel limit, $I_n/I_p \propto \mu_n/\mu_p \approx 4$ from Eq. (3.28) and Figs 3.15 and 3.16, assuming matched channel lengths and threshold voltages. For short-channel devices, however, the ratio is smaller since nMOSFETs are more velocity saturated than pMOSFETs. Typically, the current-per-width ratio I_n/I_p is about 2–2.5 for deep-submicron CMOS technologies; therefore, ***$W_p/W_n = 2$ is a good choice for CMOS inverter design.***

5.1.1.2 CMOS Inverter Noise Margin

Because of the nonlinear saturation characteristics of the MOSFET I_{ds}–V_{ds} curves, the V_{out}–V_{in} curve is also highly nonlinear. The maximum slope of the high-to-low transition

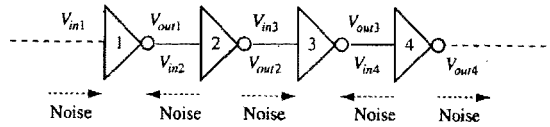


Figure 5.5. A cascade chain of identical CMOS inverters. The noise voltages at the input of each stage are for the discussion of Fig. 5.7.

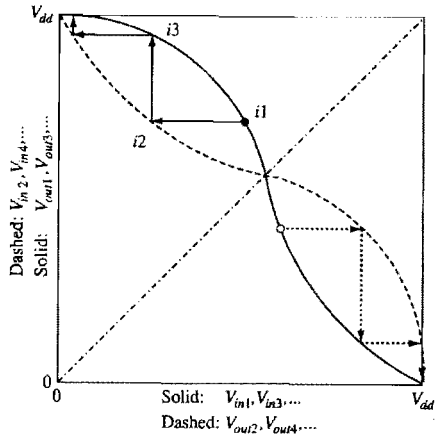


Figure 5.6. The solid transfer curve is for odd numbered inverter stages. The flipped, dashed transfer curve is for even numbered stages. The connected line segments between the curves depict the trajectory of node voltages through successive inverter stages.

of the V_{out}–V_{in} curve, $|dV_{out}/dV_{in}|$, referred to as the maximum voltage gain, is a measure of the g_m/g_{ds} (Exercise 4.8) ratio of the two transistors. From the condition $W_nI_n(V_{gsn}, V_{dsn}) = W_pI_p(V_{gsp}, V_{dsp})$, it can be shown that

$$\frac{dV_{out}}{dV_{in}} = -\frac{W_n g_{mn} + W_p g_{mp}}{W_n g_{dsn} + W_p g_{dsp}}, \tag{5.2}$$

where $g_{mn} \equiv \partial I_n/\partial V_{gsn}$, $g_{mp} \equiv -\partial I_p/\partial V_{gsp}(>0)$, $g_{dsn} \equiv \partial I_n/\partial V_{dsn}$, $g_{dsp} \equiv -\partial I_p/\partial V_{dsp}$ (>0), etc.

A commonly employed scheme to quantify the noise margin of a transfer curve is to consider a chain of identical inverters in cascade as shown in Fig. 5.5. The solid curve in Fig. 5.6 represents the transfer curve of inverters #1, #3, #5, ..., i.e., V_{out1} vs. V_{in1}, V_{out3} vs. V_{in3}, etc. A complementary dashed curve is generated by flipping or mirror imaging the solid curve with respect to the chained line, $V_{in} = V_{out}$. It represents the inverse transfer curve of inverters #2, #4, #6, ..., i.e., V_{in2} vs. V_{out2}, V_{in4} vs. V_{out4}, etc. In this graphical construction, one can visualize a trajectory of alternating horizontal and vertical lines between the two curves as the node voltage makes its transitions through the inverter

stages. Starting with dot $i1$ on the solid curve at coordinates (V_{in1}, V_{out1}), the next point $i2$ is on the dashed curve at coordinates (V_{out2}, V_{in2}). The line between $i1$ and $i2$ is horizontal since $V_{in2} = V_{out1}$. The next point $i3$ is back on the solid curve with coordinates (V_{in3}, V_{out3}), and is connected to $i2$ by a vertical line as $V_{in3} = V_{out2}$, etc. In this example, the node voltage is pushed after each inverter stage closer and closer to the upper left corner corresponding to $V_{out} = V_{dd}$ for subsequent odd stages and $V_{out} = 0$ for subsequent even stages. If the starting point is below the $V_{in} = V_{out}$ intercept such as the circle in Fig. 5.6, it will be pushed in dotted line segments to the lower right corner, i.e., $V_{out} = 0$ for subsequent odd stages and $V_{out} = V_{dd}$ for subsequent even stages. Such a characteristic is called "regenerative" which widens the noise margin as the node voltage is restored to one of the extremes of the binary digital states.

To add noise to the above picture, we consider only two inverter stages with the transfer curves depicted in Fig. 5.7(a). A positive noise voltage at the input to inverter #1 (Fig. 5.5) kicks the starting point from $i1$ to $i1'$ on the solid curve. If there is no noise at the input to inverter #2, the output after two inverter stages will end up at point $i3$ shown. If $i3$ is to the left of $i1$, then there is a net gain of noise margin after the two inverters with noise. On the other hand, if $i3$ is to the right of $i1$, then there is a net loss of noise margin. In that case, the input voltage to the odd-numbered inverters may keep increasing through repeated cycles with noise. Finally it will cross over the $V_{in} = V_{out}$ line and the logic state is lost (flipped). The maximum noise voltage that can be tolerated is then the one that causes $i3$ to fall back on top of $i1$.

We now add a negative going noise voltage of the same magnitude to the input of inverter #2 (Fig. 5.5). Note that for this example, while a positive going noise voltage is worst at input 1, a negative going noise is worst at input 2. As shown in Fig. 5.7(b), the negative noise voltage kicks the input to inverter #2 from $i2$ to $i2'$. The maximum noise magnitude that can be tolerated without eventually losing the logic state is the one that returns exactly to $i1$ after two noisy stages. Therefore, ***the noise margin for a given transfer curve is measured by the size of the maximum square that can fit between itself and its complementary curve*** (Hill, 1968). A different way of arriving at the same result is described in Fig. 9.7 for the noise margin of SRAM cells. It is evident that for given n- and pMOSFETs, a wider noise margin is achieved with the width ratio of Eq. (5.1) so that the high-to-low transition of the transfer curve happens at $V_{dd}/2$.

Since most of the noise interference in a chip environment originates from coupling of voltage transients in the neighboring lines or devices, the noise magnitude is expected to scale with the power supply voltage (except those with other natural origins such as "soft error" due to high-energy particles). Thermal noise has too low a magnitude of concern as long as $V_{dd} >> kT/q$. Therefore, a relevant measure of the noise margin in a CMOS circuit is the normalized V_{NM}/V_{dd}, where V_{NM} is the side of the maximum square in Fig. 5.7(b). Large V_{NM}/V_{dd} (up to 0.5 in principle) is obtained with a highly skewed, symmetric transfer curve, i.e., one that has V_{out} staying high at low to medium V_{in}, then making an abrupt high-to-low transition at $V_{in} = V_{dd}/2$. It can be seen from the construction of the transfer curve in Fig. 5.3(c) that for a given V_{dd}, V_{NM}/V_{dd} improves with a higher threshold voltage, V_t/V_{dd}. In fact, the best noise margin is achieved with subthreshold operation (Frank *et al.*, 2001), although with poor delay performance. As V_{dd} is scaled

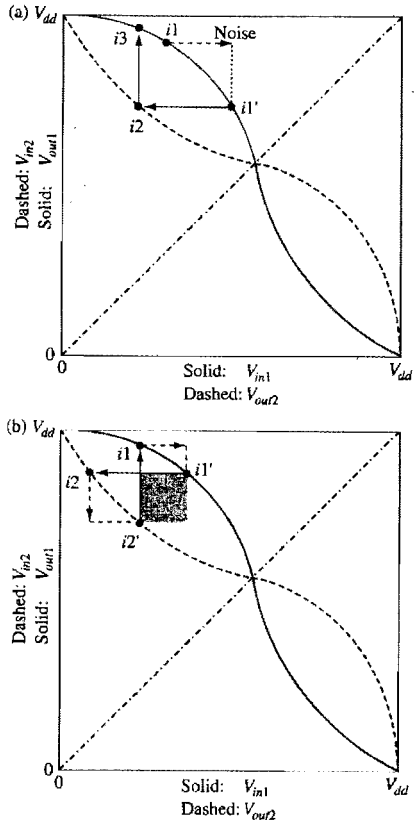


Figure 5.7. (a) Node voltage trajectory with noise added to the input to inverter #1. (b) Node voltage trajectory with positive noise at inverter #1 and negative noise at inverter #2. The shaded area represents the largest square that can be circumscribed in between the two transfer curves. The side of the square, V_{NM}, is a measure of the noise margin.

down, V_{NM}/V_{dd} is not particularly sensitive to V_{dd} until V_{dd} becomes comparable to kT/q. In order to have the nonlinear I–V characteristics necessary for digital circuit function, a minimum V_{dd} of several kT/q, e.g., 100–200 mV, is required (Swanson and Meindl, 1972). At $V_{dd} \sim 1$ V level, the choice of power supply voltage for static CMOS logic circuits is largely based on power and performance considerations discussed in Section 5.3.3, not noise margin.

5.1.1.3 CMOS Inverter Switching Characteristics

We now consider the basic switching characteristics of a CMOS inverter. The simplest input waveform is when the gate voltage makes an abrupt or infinitely sharp transition

from low to high or vice versa. For example, consider the inverter biased at point A in Fig. 5.3(a) when V_{in} makes a step transition from 0 to V_{dd}. Before the transition, the nMOSFET is off and the pMOSFET is on. After the transition, the nMOSFET is on and the pMOSFET is off. The trajectory of V_{out} from point A to point B follows the $V_{in} = V_{dd}$ curve of the nMOSFET as shown in Fig. 5.3(a). If the total capacitance of the output node (including both the output capacitance of the switching inverter and the input capacitance of the next stage or stages it drives) is represented by two capacitors – one (C_-) to the ground and one (C_+) to the V_{dd} rail, as illustrated in Fig. 5.2 – then the pull-down switching characteristics are described by

$$C_- \frac{d(V_{out} - 0)}{dt} + C_+ \frac{d(V_{out} - V_{dd})}{dt} = -I_N(V_{in} = V_{dd}),$$

or

$$(C_- + C_+)\frac{dV_{out}}{dt} = C\frac{dV_{out}}{dt} = -I_N(V_{in} = V_{dd}), \tag{5.3}$$

with the initial condition $V_{out}(t=0) = V_{dd}$. Here $C = C_- + C_+$ includes both the capacitance to ground and the capacitance to V_{dd}. For simplicity, we approximate the I_N ($V_{in} = V_{dd}$) curve by two piecewise continuous lines. In the saturation region ($V_{out} > V_{dsat}$), $I_N = I_{onN}$ is a constant. In the linear region ($V_{out} < V_{dsat}$), $I_N = (I_{onN}/V_{dsat})V_{out}$, like a resistor with a resistance V_{dsat}/I_{onN}. These are shown as dotted lines in Fig. 5.3(a). The solution $V_{out}(t)$ is depicted in Fig. 5.8(a). Right after V_{in} switches from 0 to V_{dd}, V_{out} decreases linearly with time at a rate given by I_{onN}/C until $V_{out} = V_{dsat}$ is reached. Below that, V_{out} decreases exponentially toward zero with a time constant CV_{dsat}/I_{onN}. Similar switching characteristics are shown in Fig. 5.8(b) for pMOSFET pull-up when V_{in} switches abruptly from V_{dd} to 0 [point B to point A in Fig. 5.3(b)]. In this case, V_{out} follows the $V_{in} = 0$ curve of the pMOSFET, and the initial rate of increase with time is given by I_{onP}/C.

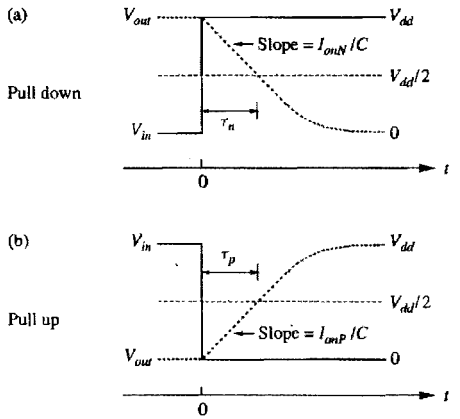


Figure 5.8. Waveforms of the output node voltage (dotted) of a CMOS inverter. (a) Pull-down transition after an abrupt rise of input voltage (solid). (b) Pull-up transition after an abrupt fall of input voltage (solid).

While it takes a significantly longer time for the output voltage to approach zero, it is conventional to define an nMOSFET pull-down delay τ_n as the time it takes for the output node voltage to reach $V_{dd}/2$. From Fig. 5.8(a), it is clear that

$$\tau_n = \frac{CV_{dd}}{2I_{onN}} = \frac{CV_{dd}}{2W_n I_{on,n}}, \tag{5.4}$$

where $I_{on,n} \equiv I_{onN}/W_n$ is the nMOSFET on current per unit width. Similarly, the pMOSFET pull-up delay is

$$\tau_p = \frac{CV_{dd}}{2I_{onP}} = \frac{CV_{dd}}{2W_p I_{on,p}}, \tag{5.5}$$

where $I_{on,p} \equiv I_{onP}/W_p$ is the pMOSFET on current per unit width. ***If a CMOS inverter is designed with a device width ratio given by Eq. (5.1) for a symmetrical transfer curve, it also follows that the pull-up and pull-down delays are equal.*** The less conductive pMOSFET is compensated by having a width wider than that of the nMOSFET. The width ratio for the minimum switching delay, $\tau = (\tau_n + \tau_p)/2$, is generally different from that of Eq. (5.1) (Exercise 5.1) (Hedenstierna and Jeppson, 1987). However, the minimum is rather shallow, and the difference between the switching delay of a symmetric CMOS inverter and the minimum delay is usually no more than 5%.

5.1.1.4 Switching Energy and Power Dissipation

Switching a CMOS inverter or other logic circuit in general takes a certain amount of energy from the power supply. Let us first focus on the capacitor C_- between the output node and ground in Fig. 5.2. During the pull-up transition of a CMOS inverter, the charge on C_- changes from zero to $\Delta Q_- = C_- V_{dd}$. This means that there is an energy of $V_{dd}\Delta Q_- = C_- V_{dd}^2$ flowing out of the power supply in the pull-up transition. Half of this energy, or $C_- V_{dd}^2/2$, is dissipated by the charging current in the pMOSFET resistance. The other half, another $C_- V_{dd}^2/2$, is stored in the capacitor C_-. This energy stays in the capacitor until the next pull-down takes place. Then the charge on C_- drops to zero and the stored energy is dissipated by the discharging current through the nMOSFET resistance. Likewise, for the capacitor C_+ between the output node and V_{dd} in Fig. 5.2, an amount of energy $C_+ V_{dd}^2$ is supplied by the power source during the pull-down transition; half of which is dissipated by the discharging current in the nMOSFET resistance, while the other half ($C_+ V_{dd}^2/2$) is stored in the capacitor C_+. The stored energy is later dissipated in the pMOSFET resistance during the next pull-up transition.

From the above discussion, it is clear that for any capacitor C (either to the ground or to V_{dd}) to be charged or discharged, an energy of $CV_{dd}^2/2$ is dissipated irreversibly. It is often conventional to consider a complete cycle consisting of a pair of transitions, either up–down ($0 \rightarrow V_{dd} \rightarrow 0$) or down–up ($V_{dd} \rightarrow 0 \rightarrow V_{dd}$). In that case, we say an energy of CV_{dd}^2 is dissipated per cycle. (An exception is Miller capacitances between two switching nodes, to be discussed in Section 5.3.4.)

Since dc power dissipation is negligible in CMOS circuits, the only power consumption comes from switching. (Standby power dissipation of low-V_t devices is discussed in Section 5.3.3.) While the peak power dissipation in a CMOS inverter can reach $V_{dd} I_{onN}$ or $V_{dd} I_{onP}$, the average power dissipation depends on how often it switches. In a CMOS processor, the switching of logic gates is controlled by a clock generator of frequency f. If on the average a total equivalent capacitance C is charged and discharged within a clock cycle of period $T = 1/f$, the average power dissipation is

$$P = \frac{CV_{dd}^2}{T} = CV_{dd}^2 f. \tag{5.6}$$

Note here that each up or down transition of a capacitor within the period T contributes half of that capacitance to C. If, for example, a capacitor is switched four times (goes through the up–down cycle twice) within the clock period, its capacitance is counted twice in C. Equation (5.6) will be used in the discussion of power–delay tradeoff in Section 5.3.

The above simplified delay and power analysis assumes abrupt switching of V_{in}. In general, V_{in} is fed from a previous logic stage and has a finite rise or fall time associated with it. The switching trajectory from A to B or from B to A in Fig. 5.3 then becomes much more complicated. Instead of staying on one constant-V_{in} curve, the bias point moves through different curves as V_{in} ramps up or down. Furthermore, both I_N and I_P must be considered during either a pull-up or a pull-down transition, since the other transistor is not switched off completely as one transistor is turned on. This also means that there is a *crossover*, or *short-circuit*, current that flows momentarily between the power-supply terminal and the ground in a switching event, which adds another power dissipation component to Eq. (5.6). One last complication is that the output node capacitances C_- and C_+ are generally voltage-dependent rather than being constant as assumed above. More extensive numerical analysis of the general case will be given in Section 5.3.

5.1.1.5 Quasistatic Assumption

In the above discussion of CMOS switching characteristics, it was implicitly assumed that the device response time, i.e., the time required for charge redistribution, is fast compared with the time scale the terminal voltage is changed. This is called the *quasistatic* assumption. In other words, the device current responds instantaneously to an external voltage change. This assumption is valid if the input rise or fall time is much longer than the carrier transit time across the channel. In general, the carrier transit time can be expressed as

$$t_r = \int_0^L \frac{dy}{v(y)}, \tag{5.7}$$

where $v(y)$ is the carrier velocity at a point y in the channel. Current continuity requires $I = WQ_i(y)v(y)$ be a constant, independent of y. Equation (5.7) then becomes

$$t_r = \frac{W}{I}\int_0^L Q_i(y)dy = \frac{Q_I}{I}, \tag{5.8}$$

where Q_I is the total mobile charge in the device.

For a long-channel MOSFET in saturation, I is given by Eq. (3.28) and Q_I is given by the integration of Eq. (3.59) or the expression above Eq. (3.60). Therefore, the transit

time is of the order of $L^2/\mu_{eff}V_{dd}$. For a completely velocity saturated device, the transit time approaches L/v_{sat}, which is of the order of 10 ps for 1 μm MOSFETs and 1 ps for 0.1 μm MOSFETs. These numbers are at least an order of magnitude shorter than the delay of an unloaded CMOS inverter made in the corresponding technology (Taur *et al.*, 1985; Taur *et al.*, 1993c). This indicates that the switching time is limited by the parasitic capacitances rather than by the time required for charge re-distribution within the transistor itself and thus validates the quasistatic approach.

5.1.2 CMOS NAND and NOR Gates

CMOS inverters described in the last subsection are used to invert a logic signal, to act as a buffer or output driver, or to form a latch (two inverters connected back to back). However, they cannot perform logic computation, since there is only one input voltage. In the static CMOS logic family, the most widely used circuits with multiple inputs are *NAND* and *NOR* gates as shown in Fig. 5.9. In a NAND gate, a number of nMOSFETs are connected in series between the output node and the ground. The same number of pMOSFETs are connected in parallel between V_{dd} and the output node. Each input signal is connected to the gates of a pair of n- and pMOSFETs as in the inverter case. In this configuration, the output node is pulled to ground only if all the nMOSFETs are turned on, i.e., only if all the input voltages are high (V_{dd}). If one of the input signals is low (zero voltage), the low-resistance path between the output node and ground is broken, but one of the pMOSFETs is turned on, which pulls the output node to V_{dd} . On the contrary, the NOR circuit in Fig. 5.9(b) consists of parallel-connected nMOSFETs between the output node and ground, but serially connected pMOSFETs between V_{dd} and the output node. The output voltage is high only if all the input voltages are low, i.e., all the pMOSFETs are on and all the nMOSFETs are off. Otherwise, the output is low.

Due to the complementary nature of n- and pMOSFETs and the serial-versus-parallel connections, there is no direct low-resistance path between V_{dd} and ground except during switching. In other words, just like CMOS inverters, ***there is no static current or standby power dissipation for any combination of inputs in either the CMOS NAND or NOR circuits***. The circuit output resistance is low, however, because of the conducting transistor(s).

In CMOS technology, NAND circuits are much more frequently used than NOR. This is because it is preferable to put the transistors with the higher resistance in parallel and those with the lower resistance in series. Since pMOSFETs have a higher resistance due to the lower hole mobility, they are rarely used in series (stacked). By connecting low-resistance nMOSFETs in series and high-resistance pMOSFETs in parallel, a NAND gate is more balanced in terms of the pull-up and the pull-down operations and achieves better noise immunity as well as a higher overall circuit speed.

5.1.2.1 Two-Input CMOS NAND Gate

As an example, we will examine the transfer curve and the switching characteristics of a two-input NAND gate, also referred to as a two-way NAND, or NAND with a fan-in of two, shown in Fig. 5.10. With the two pMOSFETs connected in parallel between V_{dd} and the output node, the pull-up operation of a two-way NAND is similar to that of an

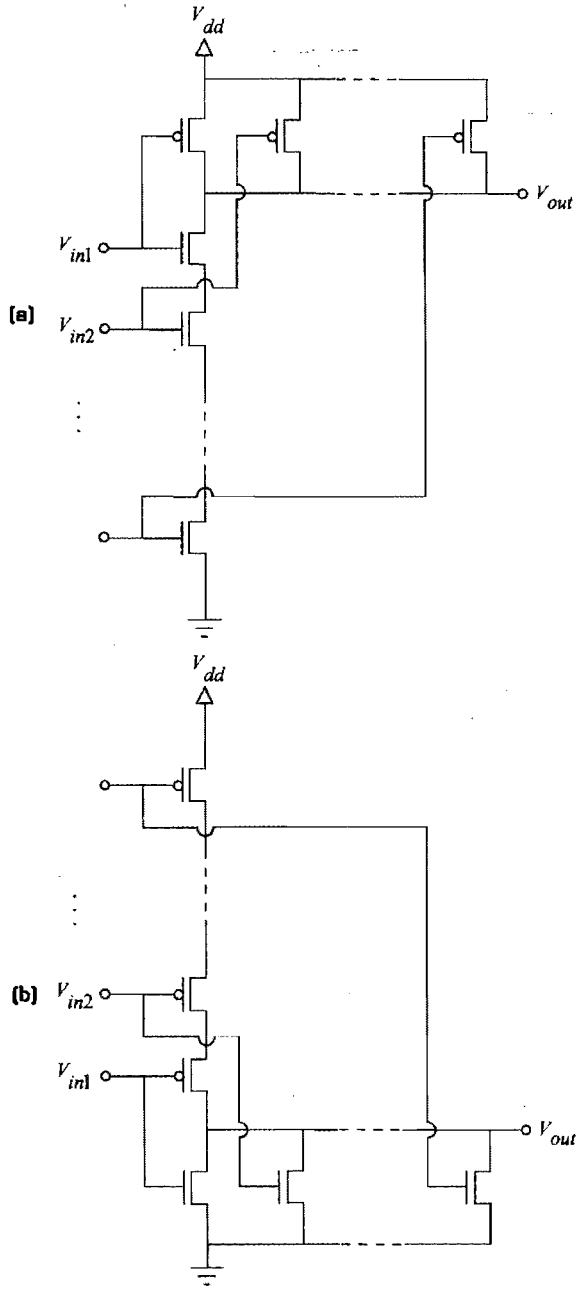


Figure 5.9. Circuit diagram of (a) CMOS NAND and (b) CMOS NOR. Multiple input signals are labeled

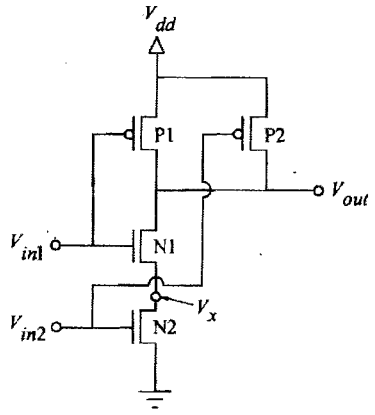


Figure 5.10. Circuit diagram of a two-input CMOS NAND. The transistors are labeled P1, P2 and N1, N2.

inverter. If either one of the transistors is being turned on while the other one is off, the charging current is identical to that of the pMOSFET pull-up in a CMOS inverter discussed in the previous subsection. If both transistors are pulling up, the total charging current is doubled as if the pMOSFET width had been increased by a factor of two. On the other hand, the two nMOSFETs are connected in series (stacked) between the output and ground, and their switching behavior is quite different from that of the inverters. For the bottom transistor N2, its source is connected to the ground and the gate-to-source voltage is simply the input voltage V_{in2}. However, for the top transistor N1, its source is at a voltage V_x (Fig. 5.10) higher than the ground. V_x plays a crucial role in the switching characteristics of N1, since the gate-to-source voltage that determines how far N1 is turned on is given by $V_{in1} - V_x$. Transistor N1 is also subject to the body-bias effect, as a source voltage V_x is analogous to a reverse body (substrate) bias $V_{bs} = -V_x$ in Fig 3.13, which raises the threshold voltage of N1 as described by Eq. (3.44).

There are three possible switching scenarios, each with different characteristics. They are described below.

- *Case A. Bottom switching: Input 2 switches while input 1 stays at V_{dd}.* Initially, even though $V_{in1} = V_{dd}$, but $V_x > V_{dd} - V_t$ so that $V_{gs}(\text{N1}) = V_{dd} - V_x < V_t$ and both N1 and N2 are in subthreshold. The pull-down transition in case A when input 2 rises from 0 to V_{dd} is most similar to the nMOSFET pull-down in an inverter. For low input voltages $V_{in2} \leq V_{dd}/2$, transistor N2 is in saturation. Transistor N1 can be in the linear region or in saturation. In either case, N1 only acts to reduce the drain voltage of N2 with little effect on the current. The transfer curve of V_{out} versus V_{in2} in this case is similar to that of a CMOS inverter, which exhibits symmetrical characteristics if $W_p/W_n = I_n/I_p \approx 2$, as shown in Fig. 5.11. For high input voltages $V_{in2} > V_{dd}/2$, the current is somewhat degraded by the resistance of N1 as transistor N2 moves out of saturation.
- *Case B. Top switching: Input 1 switches while input 2 stays at V_{dd}.* For the pull-down transition in case B, transistor N1 is in saturation while N2 is in linear mode during

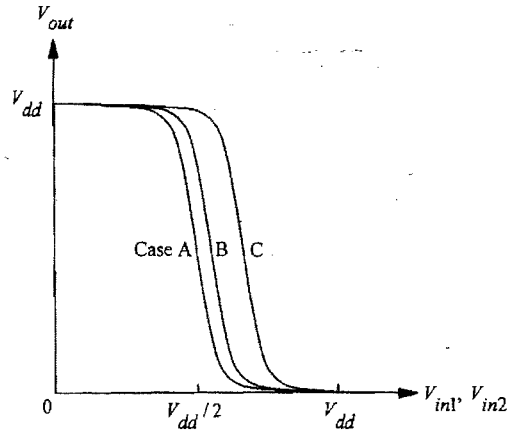


Figure 5.11. Transfer curves of a two-input CMOS NAND for different cases of switching discussed in the text. The device width ratio W_p / W_n is taken to be 2 in this illustration.

most part of the switching cycle. Transistor N2 therefore acts like a series resistance connected in the source terminal of N1. The voltage V_x between the two transistors rises slightly above ground, depending on the current level. This degrades the pull-down current as the gate-to-source voltage of N1 is reduced to $V_{in1} - V_x$ and its threshold voltage is increased by $(m-1)V_x$ due to the body effect. As a result, a slightly higher input voltage V_{in1} is needed to reach the high-to-low transition of the transfer curve in Fig. 5.11. Even though the pull-down current in case B is slightly less than in case A, the switching time in case B is comparable to that in A if the output is not too heavily loaded. This is because of the additional capacitance in case A associated with the top transistor N1 that needs to be discharged from V_{dd} to ground when the bottom device is switching. These factors are further discussed in detail in Section 5.3.5.

- *Case C. Both input 1 and input 2 switch simultaneously.* The worst case for pull-down in a two-input CMOS NAND is case C, in which both inputs rise from 0 to V_{dd}. It can be seen that transistor N2 is always biased in the linear region, while transistor N1 is in the linear region for small values of V_{out} and in saturation for large V_{out}. In this case, the nMOSFET pull-down current is reduced by approximately a factor of two from the inverter case because of the serial connection. The pull-up current, on the other hand, is twice that of the inverter case due to the parallel connection of pMOSFETs. This moves the high-to-low transition in the transfer curve to a V_{in} significantly higher than $V_{dd}/2$, as shown in case C of Fig. 5.11.

5.1.2.2 Noise Margin of NAND Circuits

Because of the spread of transfer curves under different switching conditions, the noise margin of a CMOS NAND gate is inferior to that of a CMOS inverter. In an

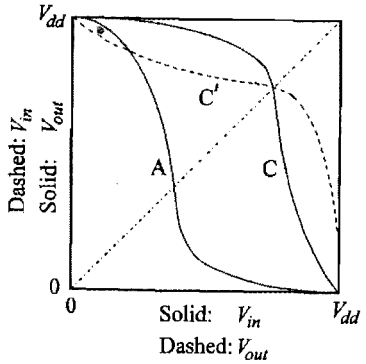


Figure 5.12. An example of worst-case noise margin of NAND circuits. Curve C′ is the mirror image of C with respect to the axis $V_{in} = V_{out}$ (Liu *et al.*, 2006).

exaggerated case shown in Fig. 5.12, curves A and C represent the extremes of all possible transfer curves. The best that can be done with the width selection (W_p/W_n) is such that A and C are symmetric on either side of $V_{dd}/2$. In the worst case scenario, one needs to consider the noise margin of a cascade chain of NAND stages with alternating switching conditions A and C. In other words, the worst case noise margin is given by the size of the smaller square that can be circumscribed between curves A and C′, the flipped counterpart of C. In the example illustrated, the noise margin is severely degraded, but still positive. If the power supply voltage is too low and the number of fan-in too large, there could end up with no intersection between A and C′. That would mean eventual loss of logic state after repeated stages of worst case switching events. Higher threshold voltages are usually beneficial to noise margin, although at the expense of switching speed. ***The minimum power supply voltage for maintaining logic consistency of NAND or NOR circuits is of the order of (5–10)kT/q*** (Frank *et al.*, 2001).

5.1.3 Inverter and NAND Layouts

5.1.3.1 Layout of a Single Device

Both the CMOS circuit density and the delay performance are determined by the layout ground rules of the particular technology. Figure 5.13 shows a typical layout of an isolated MOSFET and its corresponding cross section. Only three major masking levels are shown: active region (isolation), polysilicon gate, and contact hole. To complete a CMOS process, several additional implant blockout masks are needed for doping the channel and the source–drain regions of nMOSFETs and pMOSFETs, respectively (Appendix 1). After the device or the front-end-of-line (FEOL) process, a number of metal levels are laid down in the back-end-of-line (BEOL) process to connect the transistors into various circuits that make up the chip.

In Fig. 5.13, the device length and width are indicated by L and W, respectively. The contact-hole size, represented by c, is limited by lithography. The spacings between the

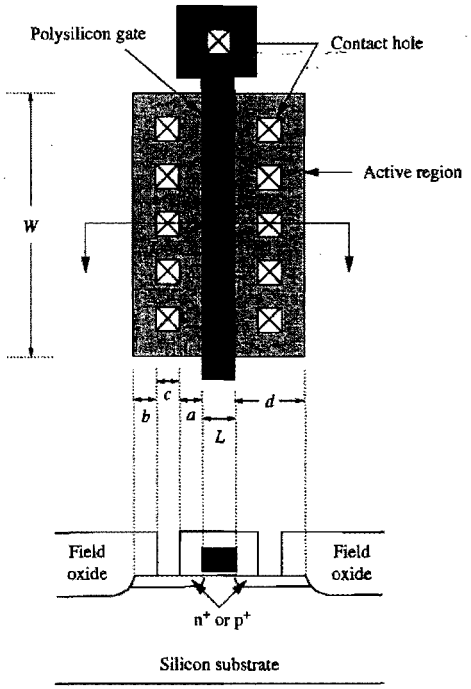


Figure 5.13. Basic layout and corresponding cross-section of a single MOSFET, illustrating several key layout ground rules.

contact and the gate and between the contact and the edge of the active region are represented by a and b, respectively. These minimum distances are required in the ground rules to allow for alignment tolerances between the levels as well as linewidth biases and variations. Added together, a, b, and c determine the distance between the gate and the field isolation, i.e., the width of n^+ or p^+ diffusion, d. As far as CMOS delay is concerned, d should be kept as small as possible, since a larger diffusion area adds more parasitic capacitance to be switched during a transition. In a silicided technology, the diffusion area of a sufficiently wide MOSFET can be somewhat reduced by not extending the contact holes throughout the entire device width. The polysilicon contact area outside the active region is not a critical factor, as the additional capacitance it introduces is negligible because of the thick field oxide (about 50 times thicker than gate oxide) underneath.

5.1.3.2 Layout of a CMOS Inverter

Figure 5.14(a) shows a simple layout of a CMOS inverter with $W_p/W_n \approx 2$. Four metal wires are shown, leading to V_{dd}, ground, input, and output. The pMOSFET

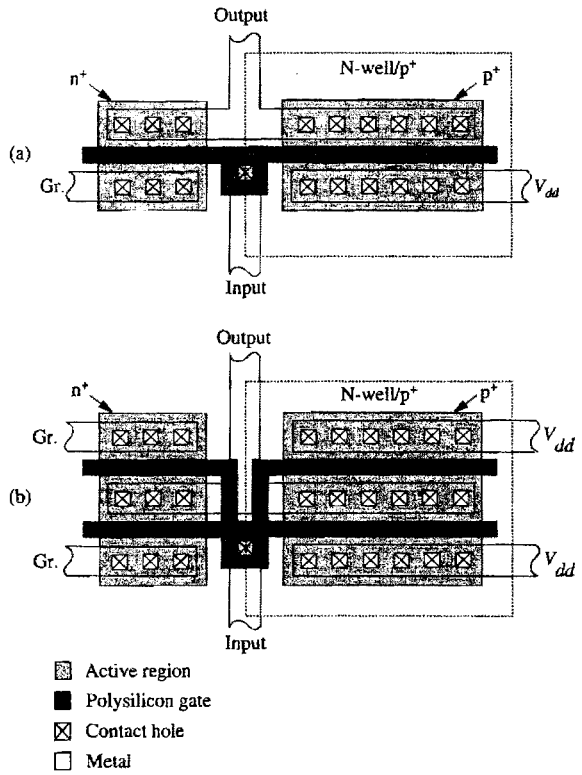


Figure 5.14. Layout of a CMOS inverter with (a) straight gates and (b) folded gates for minimizing the parasitic diffusion capacitance.

receives n-well and p^+ source–drain implants with the use of two block-out masks. The nMOSFET receives a p-type channel implant and an n^+ source–drain implant with block-out masks of the complementary polarity. The intrinsic delay of a CMOS inverter, defined in terms of one stage driving an identical stage (fan-out = 1), is independent of the device width except for some parasitic effects at the ends. This is because both the current and the capacitance (gate and diffusion) are proportional to the device width in the straight-gate layout in Fig. 5.14(a). ***A dramatic reduction in the junction contributions to the parasitic capacitance can be achieved using the folded layout shown in Fig. 5.14(b)***. By sandwiching the drain node between two symmetric source regions with a fork-shaped polysilicon gate, the device width and therefore the current is effectively doubled without increasing the diffusion area. In other words, the junction capacitance per effective device width in layout (b) is about half of that in layout (a), assuming $a + b + c$ is comparable to $2a + c$ in Fig. 5.13. Note that the area

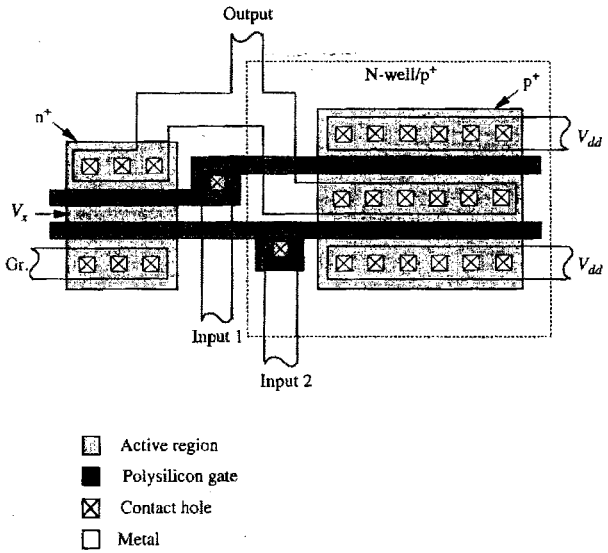


Figure 5.15. Layout example of a two-input CMOS NAND with the equivalent circuit in Fig. 5.10.

of the source regions is of no importance to the delay, since the source voltage is not being switched.

5.1.3.3 Layout of a Two-Input CMOS NAND

A typical layout for a two-input CMOS NAND is shown in Fig. 5.15. The two parallel-connected pMOSFETs are arranged as in the folded inverter, with the switching node sandwiched between the two input gates. This again minimizes junction capacitance. The two nMOSFETs are connected in series via a V_x-node between the input gates. Since no contact to the V_x-diffusion is necessary, its width can be kept as narrow as the minimum linewidth, i.e., comparable to L, c, etc., so that the capacitance associated with it is relatively small.

5.2 Parasitic Elements

From the previous section, it is clear that for a given supply voltage, the CMOS delay is mainly determined by the device current and the capacitance of the switching node. In addition to the intrinsic current and capacitance discussed in Chapter 3, however, any parasitic resistances and capacitances that reduce the current drive or increase the node capacitance can also affect the CMOS delay. This section examines such parasitic elements as source–drain resistance, junction capacitance, overlap capacitance, gate resistance, and interconnect RC components.

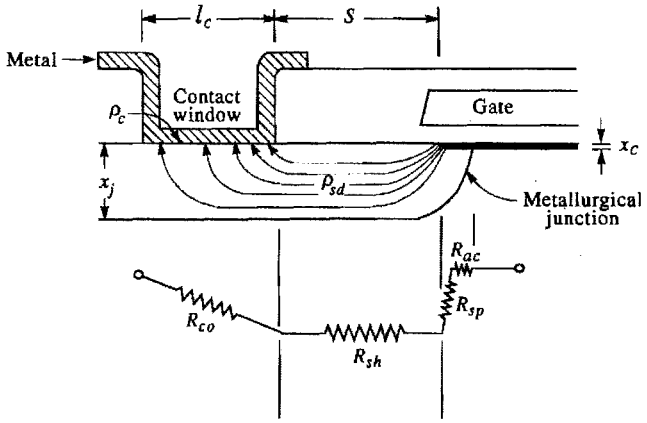


Figure 5.16. A schematic cross section showing the pattern of current flow from a MOSFET channel through the source or drain region to the metal contact. The diagram identifies various contributions to the series resistance. The device width in the z-direction is assumed to be W. (After Ng and Lynch, 1986.)

5.2.1 Source–Drain Resistance

It was discussed in Section 3.2.4 that source–drain series resistance degrades the current of a short-channel MOSFET whose intrinsic resistance is low. Resistance on the source side is particularly troublesome, as it degrades the gate drive as well.

A schematic diagram of the current-flow pattern in the source or drain region of a MOSFET is shown in Fig. 5.16 (Ng and Lynch, 1986). The total source or drain resistance can be divided into several parts: R_{ac} is the accumulation-layer resistance in the gate–source (or –drain) overlap region where the current mainly stays at the surface; R_{sp} is associated with current spreading from the surface layer into a uniform pattern across the depth of the source–drain; R_{sh} is the sheet resistance of the source–drain region where the current flows uniformly; and R_{co} is the contact resistance (including the spreading resistance in silicon under the contact) in the region where the current flows into a metal line. Once the current flows into an aluminum line, there is very little additional resistance, since the resistivity of aluminum is very low, $\rho_{Al} \approx 3 \times 10^{-6}\Omega$-cm. In VLSI interconnects, the aluminum thickness is typically 0.5–1.0 μm. From Eq. (2.34), the sheet resistivity is on the order of 0.05 Ω/□. This is negligible compared with the channel sheet resistivity, $\rho_{ch} \approx 2000$–$7000\ \Omega/\square$, except when a long, thin wire is connected to a wide MOSFET. Figure 5.16 shows only the series resistance on one side of the device. The total source–drain series resistance per device is, of course, twice that shown in Fig. 5.16, assuming that the source and drain are symmetrical. Below we examine the various components of the source–drain resistance.

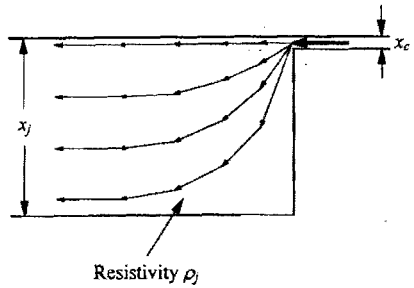


Figure 5.17. Schematic diagram showing the resistance component associated with the injection region where the current spreads from a thin surface layer into a uniformly doped source or drain region. (After Baccarani and Sai-Halasz, 1983.)

5.2.1.1 Accumulation-Layer Resistance and Spreading Resistance

The accumulation-layer resistance R_{ac} depends on the gate voltage. Since it is not easily separable from the active channel resistance, R_{ac} is considered as a part of L_{eff} as discussed in Section 4.3.3.

Next we consider the spreading resistance component, R_{sp}. An analytical expression has been derived for R_{sp} assuming an idealized case shown in Fig. 5.17 where the current spreading takes place in a uniformly doped medium with resistivity ρ_j (Baccarani and Sai-Halasz, 1983):

$$R_{sp} = \frac{2\rho_j}{\pi W} \ln\left(0.75 \frac{x_j}{x_c}\right). \tag{5.9}$$

Here W is the device width, and x_j and x_c are the junction depth and the inversion (or accumulation) layer thickness, respectively. For typical values of $x_j/x_c \approx 40$, we have $R_{sp} \approx 2\rho_j/W$. In practice, however, it is difficult to apply Eq. (5.9), since current spreading usually takes place in a region where the local resistivity is highly nonuniform due to the lateral source–drain doping gradient. In general, a 2-D numerical simulation is needed to evaluate R_{ac} and R_{sp}. Qualitatively, current injection from the surface into the bulk takes place such that the sum of the resistances, $R_{ac}+R_{sp}$, is a minimum (Ng and Lynch, 1986). For an abrupt source–drain profile, the injection point is close to the metallurgical end of the channel; both R_{ac} and R_{sp} are low, and $L_{eff} \approx L_{met}$. For a graded profile, the injection point moves away from the metallurgical junction toward the gate edge, resulting in higher $R_{ac}+R_{sp}$ and $L_{eff} > L_{met}$ (Section 4.3.3).

5.2.1.2 Sheet Resistance

Next, we examine R_{sh} and R_{co}. In Fig. 5.16, the sheet resistance of the source–drain diffusion region is simply,

$$R_{sh} = \rho_{sd} \frac{S}{W}, \tag{5.10}$$

where W is the device width, S is the spacing between the gate edge and the contact edge, and ρ_{sd} is the sheet resistivity of the source–drain diffusion, typically of the order of 50–500 Ω/□. Since $\rho_{sd} \ll \rho_{ch}$ of the device, this term is usually negligible if S is kept to a minimum limited by the overlay tolerance between the contact and the gate lithography levels. In a nonsilicided technology, $S = a$ in Fig. 5.13, provided that most of the device width dimension is covered by contacts.

5.2.1.3 Contact Resistance

Based on a transmission-line model (Berger, 1972), the contact resistance can be expressed as

$$R_{co} = \frac{\sqrt{\rho_{sd}\rho_c}}{W} \coth\left(l_c\sqrt{\frac{\rho_{sd}}{\rho_c}}\right), \tag{5.11}$$

where l_c is the width of the contact window (Fig. 5.16), and ρ_c is the interfacial *contact resistivity* (in Ω-cm^2) of the ohmic contact between the metal and silicon. R_{co} includes the resistance of the current crowding region in silicon underneath the contact. In a nonsilicided technology, $l_c = c$ in Fig. 5.13. Equation (5.11) has two limiting cases: *short contact* and *long contact*. In the short-contact limit, $l_c \ll (\rho_c/\rho_{sd})^{1/2}$, and

$$R_{co} = \frac{\rho_c}{Wl_c} \tag{5.12}$$

is dominated by the interfacial contact resistance. The current flows more or less uniformly across the entire contact. In the long-contact limit, $l_c \gg (\rho_c/\rho_{sd})^{1/2}$, and

$$R_{co} = \frac{\sqrt{\rho_{sd}\rho_c}}{W}. \tag{5.13}$$

This is independent of the contact width l_c, since most of the current flows into the front edge of the contact. Once in the long-contact regime, there is no advantage increasing the contact width; $(\rho_c/\rho_{sd})^{1/2}$ is referred to as the *transfer length* in some literature.

For ohmic contacts between metal and heavily doped silicon, current conduction is dominated by tunneling or field emission. The contact resistivity ρ_c depends exponentially on the barrier height ϕ_B and the surface doping concentration N_d (Yu, 1970):

$$\rho_c \propto \exp\left(\frac{4\pi\phi_B}{qh}\sqrt{\frac{m^*\varepsilon_{si}}{N_d}}\right), \tag{5.14}$$

where h is Planck's constant and m^* is the electron effective mass. Depending on the doping concentration and contact metallurgy, ρ_c is typically in the range of $10^{-6}-10^{-8}$ Ω-cm^2.

5.2.1.4 Resistance in a Self-Aligned Silicide Technology

Both R_{sh} and R_{co} are greatly reduced in advanced CMOS technologies with self-aligned silicide (Ting *et al.*, 1982). As shown schematically in Fig. 5.18, a highly conductive (≈2–10 Ω/□) silicide film is formed on all the gate and source–drain surfaces

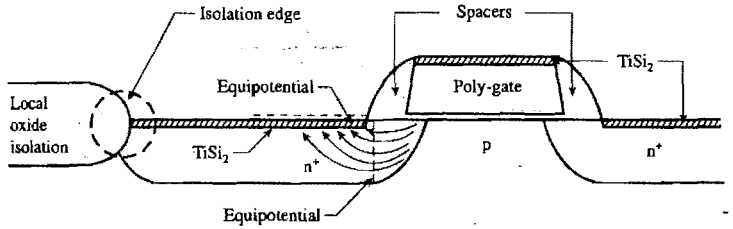


Figure 5.18. Schematic diagram of an n-channel MOSFET fabricated with self-aligned $TiSi_2$, showing the current flow pattern between the channel and the silicide. (After Taur *et al.*, 1987.)

separated by dielectric spacers in a self-aligned process. Since the sheet resistivity of silicide is 1–2 orders of magnitude lower than that of the source–drain, the silicide layer practically shunts all the currents, and the only significant contribution to R_{sh} is from the nonsilicided region under the spacer. This reduces the length S in Eq. (5.10) to 0.1–0.2 μm, which means that $R_{sh}W$ should be no more than 50 Ω-μm. At the same time, R_{co} between the source–drain and silicide is also reduced, since now the contact area is the entire diffusion. In other words, the diffusion width d in Fig. 5.13 becomes the contact length l_c in Eq. (5.11). Current flow in this case is almost always in the long-contact limit, so that Eq. (5.13) applies. However, the parameters ρ_{sd} and ρ_c in Eq. (5.13) should be replaced by ρ'_{sd} and ρ'_c: the sheet resistivity of the source–drain region under the silicide and the contact resistivity between the silicide and silicon. ρ'_{sd} is higher than the nonsilicided sheet resistivity ρ_{sd}, since a surface layer of heavily doped silicon is consumed in the silicidation process (Taur *et al.*, 1987). ρ'_c is also higher than ρ_c if the interface doping concentration becomes lower due to silicon consumption. This is particularly a concern when a thick silicide film is formed over a shallow source–drain junction. As a rule of thumb, no more than a third of the source–drain depth should be consumed in the silicide process.

In a CMOS process, a silicide material such as $TiSi_2$ with a near-midgap work function is needed to obtain approximately equal barrier heights to n^+ and p^+ silicon. The experimentally measured ρ'_c between $TiSi_2$ and n^+ or p^+ silicon is of the order of $10^{-6}-10^{-7}$ Ω-cm^2 (Hui *et al.*, 1985). Based on Eq. (5.13), therefore, R_{co} for a silicided diffusion is in the range of 50–200 Ω-μm (Taur *et al.*, 1987). The minimum contact width l_c (or diffusion width d) required to satisfy the long contact criterion can be estimated from $(\rho'_c/\rho'_{sd})^{1/2}$ to be about 0.25 μm. Contact resistance between silicide and metal is usually negligible, since the interfacial contact resistivity is of the order of $10^{-7}-10^{-8}$ Ω-cm^2 in a properly performed process.

5.2.2 Parasitic Capacitances

A schematic diagram of the MOSFET capacitances is shown in Fig. 5.19. In addition to the intrinsic capacitance C_G discussed in Section 3.1.6, there are also parasitic capacitances: namely, junction capacitance between the source or drain diffusion and the substrate (or n-well in the case of pMOSFETs), and overlap capacitance between the gate and the source or drain region. These capacitances have a significant effect on the CMOS delay.

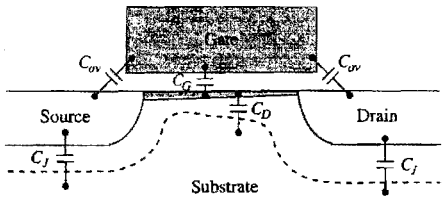


Figure 5.19. Schematic diagram of a MOSFET showing both the intrinsic capacitance C_G and the parasitic capacitances C_J, C_D, C_{ov}. The two C_J's at the source and the drain may have different values depending on the bias voltages.

5.2.2.1 Junction Capacitance

Junction or diffusion capacitance arises from the depletion charge between the source or drain and the oppositely doped substrate. As the source or drain voltage varies, the depletion charge increases or decreases accordingly. Note that when the MOSFET is on, the channel-to-substrate depletion capacitance $C_D = WLC_d$ in Fig. 5.19 can also be considered as a part of the source or drain junction capacitance. It is usually a small contribution, since the channel area of a short-channel device is generally much less than the diffusion area.

From Eq. (2.85), the capacitance per unit area of an abrupt p–n junction is

$$C_j = \frac{\varepsilon_{si}}{W_{dj}} = \sqrt{\frac{\varepsilon_{si} q N_a}{2(\psi_{bi} + V_j)}}, \tag{5.15}$$

where W_{dj} is the depletion-layer width, N_a is the impurity concentration of the lightly doped side, ψ_{bi} is the built-in potential, typically around 0.9 V as shown in Fig. 2.15, and V_j is the reverse bias voltage across the junction. Equation (5.15) indicates that the source or drain junction capacitance is voltage-dependent. At a higher drain voltage, the depletion layer widens and the capacitance decreases. Figure 2.16 plots the depletion-layer width and the capacitance at zero bias versus N_a. Since the junction capacitance increases with N_a, one should avoid doping the substrate (or n-well) regions under the source–drain junctions unnecessarily highly. Too low a doping concentration between the source and drain, however, would cause excessive short-channel effect or lead to punch-through as discussed in Section 3.2.1.

The total diffusion-to-substrate capacitance is simply equal to C_j times the diffusion area in the layout:

$$C_J = WdC_j, \tag{5.16}$$

where W is the device width, and d is the diffusion width in Fig. 5.13. For a noncontacted diffusion, d can be as small as the minimum linewidth of the lithography. The diffusion capacitance of the switching node can be reduced by a factor of two using the folded layout in Fig. 5.14(b). Strictly speaking, there are also perimeter contributions to the diffusion capacitance, since the substrate doping concentration is usually higher at the diffusion boundary due to field implants. The extra contribution can be minimized by careful process design or by using again the folded layout in which the diffusion is bounded by two gates, thus avoiding diffusion-field boundaries except at the ends.

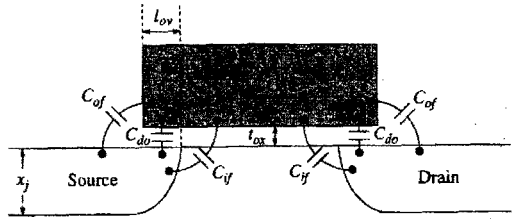


Figure 5.20. Schematic diagram showing the three components of the gate-to-diffusion overlap capacitance.

5.2.2.2 Overlap Capacitance

Another parasitic capacitance in a MOSFET is the gate-to-source or gate-to-drain overlap capacitance. It consists of three components: direct overlap, outer fringe, and inner fringe, as shown schematically in Fig. 5.20. The direct overlap component is simply

$$C_{do} = Wl_{ov}C_{ox} = \frac{\varepsilon_{ox} W l_{ov}}{t_{ox}}, \tag{5.17}$$

where l_{ov} is the length of the source or drain region under the gate. In a typical process, the oxide in the overlap region is somewhat thicker than t_{ox} due to bird's-beak near the gate edge resulting from a reoxidation step (Wong *et al.*, 1989). Therefore, l_{ov} should be interpreted as an equivalent overlap length, rather than an actual physical length.

By solving Laplace's equation analytically with proper boundary conditions, the outer and inner fringe components can be expressed as (Shrivastava and Fitzpatrick, 1982)

$$C_{of} = \frac{2\varepsilon_{ox} W}{\pi} \ln\left(1 + \frac{t_{gate}}{t_{ox}}\right), \tag{5.18}$$

and

$$C_{if} = \frac{2\varepsilon_{si} W}{\pi} \ln\left(1 + \frac{x_j}{2t_{ox}}\right), \tag{5.19}$$

where t_{gate} is the height of the polysilicon gate, and x_j is the depth of the source or drain junction. Equations (5.18) and (5.19) assume ideal shapes of the polysilicon gate and source–drain regions with square corners. In case the source–drain junctions are deeper than the gate depletion depth W_d, x_j in Eq. (5.19) should be replaced by W_d. For typical values of $t_{gate}/t_{ox} \approx 40$ and $x_j/t_{ox} \approx 20$, one obtains $C_{of}/W \approx 2.3\varepsilon_{ox} \approx 0.08$ fF/μm and $C_{if}/W \approx 1.5\varepsilon_{si} \approx 0.16$ fF/μm. Even though the inner fringe component is larger due to the higher dielectric constant of silicon, it is present only when $V_{gs} < V_t$ and the region under the gate is depleted. Once $V_{gs} > V_t$, the inversion layer forms, which effectively shields any electrostatic coupling between the gate and the inner edges of the source or drain junction. Similar shielding of the inner fringe capacitance also takes place when the gate voltage is negative (for nMOSFETs) and the surface is accumulated. Under these

conditions, the overlap capacitance consists of only the direct overlap and the outer fringe components.

From the above numerical estimates, one can write the total overlap capacitance at $V_{gs} = 0$ (silicon is depleted under the gate) as

$$C_{ov}(V_{gs} = 0) = C_{do} + C_{of} + C_{if} \approx \varepsilon_{ox} W \left(\frac{l_{ov}}{t_{ox}} + 7 \right). \tag{5.20}$$

Note that Eq. (5.20) is the maximum overlap capacitance per edge. It assumes perfectly conducting source and drain regions. In reality, because of the lateral source–drain doping gradient at the surface, the overlap capacitance depends on the drain voltage. When the drain voltage increases in an nMOSFET (with the same gate-to-substrate voltage), the overlap capacitance tends to decrease slightly because the reverse bias widens the depletion region at the surface and therefore reduces the effective overlap length (Oh *et al.*, 1990). This is especially the case with LDD (lightly doped drain) MOSFETs.

It has been reported that a minimum length of direct overlap region of the order of $l_{ov} \approx (2\text{–}3)t_{ox}$ is needed to avoid reliability problems arising from hot-carrier injection into the ungated region (Chan *et al.*, 1987b). In other words, such a margin is required to avoid "underlap" of the gate and the source–drain. Combining this requirement with Eq. (5.20), one obtains $C_{ov}/W \approx 10\,\varepsilon_{ox} \approx 0.3$ fF/μm at zero gate voltage, independent of technology generation.

5.2.3 Gate Resistance

In modern CMOS technologies, silicides are formed over polysilicon gates to lower the resistance and provide ohmic contacts to both n^+ and p^+ gates. The sheet resistivity of silicides is of the order of 2–10 Ω/□, which is generally adequate for 0.5-μm CMOS technology and above. For 0.25-μm CMOS technology and below, however, the device delay improves and gate *RC* delays may not be negligible. Compounding the problem is a tendency for silicide resistivity to increase in fine-line structures. This is due to either agglomeration or lack of nucleation sites to initiate the phase transformation in the case of $TiSi_2$. Gate *RC* delay is an ac effect not observable in dc *I–V* curves. It shows up as an additional delay component in ring oscillators, delay chains, and other logic circuits.

Gate *RC* delay can be analyzed with a distributed network shown in Fig. 5.21 for a MOSFET device of width *W* and length *L*. The resistance per unit length *R* is related to the silicide sheet resistivity ρ_g (Ω/□) by

$$R = \frac{\rho_g}{L}. \tag{5.21}$$

The capacitance per unit length, *C*, mainly arises from the inversion charge that must be supplied (or taken away) when the voltage at a particular point along the gate increases (or decreases). To a good approximation, *C* is given by the gate oxide capacitance,

$$C = C_{ox} L = \frac{\varepsilon_{ox} L}{t_{ox}}. \tag{5.22}$$

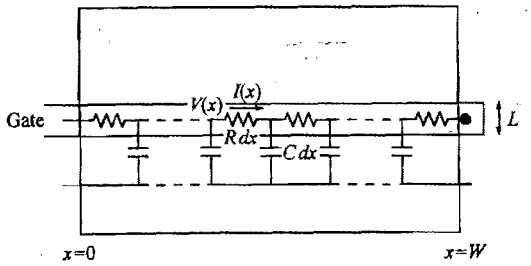


Figure 5.21. A distributed network for gate *RC* delay analysis. The lower rail represents the MOSFET channel, which is connected to the source–drain. The input step voltage is applied at the left.

For a higher accuracy, one should also include the overlap capacitance per unit gate width in *C*.

At any point *x* along the gate, one can write

$$V(x + dx) - V(x) = \frac{\partial V}{\partial x} dx = -I(x) R\, dx, \tag{5.23}$$

and

$$I(x + dx) - I(x) = \frac{\partial I}{\partial x} dx = -C\, dx \frac{\partial V}{\partial t}. \tag{5.24}$$

Eliminating $I(x)$ from the above equations, one obtains

$$\frac{\partial^2 V}{\partial x^2} = RC \frac{\partial V}{\partial t}. \tag{5.25}$$

The differential equation that governs the *RC* delay of a distributed network therefore resembles the diffusion equation with a diffusion coefficient $D = 1/RC$. If a step voltage from 0 to V_{dd} is applied at $x = 0$, the boundary conditions are $V(0, t) = V_{dd}$ and $I(W, t) = 0$, which is analogous to constant-source diffusion into a finite-width medium. The numerical solution for this case is plotted in Fig. 5.22 (Sakurai, 1983). For $t \ll RCW^2/8$, the solution can be approximated by a *complementary error function*,

$$V(x, t) = V_{dd} \operatorname{erfc}\left(\frac{x}{\sqrt{4t/RC}} \right), \tag{5.26}$$

where

$$\operatorname{erfc}(y) \equiv \frac{2}{\sqrt{\pi}} \int_y^\infty e^{-z^2} dz. \tag{5.27}$$

For $t \gg RCW^2/4$, the approximate solution is given by

$$V(x, t) = V_{dd} \left[1 - \frac{4}{\pi} \sin\left(\frac{\pi x}{2W} \right) \exp\left(-\frac{\pi^2 t}{4RCW^2} \right) \right]. \tag{5.28}$$

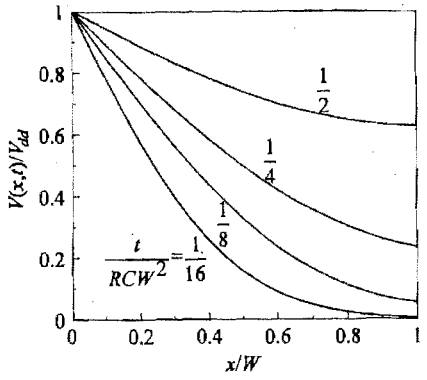


Figure 5.22. Local gate voltage versus distance along the width of the device at different time intervals after an input voltage V_{dd} is applied at $x = 0$.

It can be seen from Fig. 5.22 that the average value of $V(x, t)$ within $0 < x < W$ reaches $V_{dd}/2$ when $t \approx RCW^2/4$. If one takes this value as the effective RC delay τ_g due to the gate resistance and substitutes Eqs. (5.21) and (5.22) for R and C, one obtains

$$\tau_g = \frac{\rho_g C_{ox} W^2}{4}. \tag{5.29}$$

Note that τ_g is independent of device length but is proportional to the square of the device width. Clearly, the gate RC delay has a more significant effect percentage-wise in fast-switching unloaded inverters than in heavily loaded circuits. In order to limit τ_g to less than 1 ps, assuming $\rho_g = 10\,\Omega/\square$ and $t_{ox} = 50$ Å, the device width W must be restricted to 7.6 μm or below. ***Multiple-finger gate layouts with interdigitated source and drain regions should be used when higher-current drives are needed.*** Such types of layouts also offer the benefit of reduced (by 2×) drain junction capacitance, just like the folded layout in Fig. 5.14(b).

It should be pointed out that Eq. (5.29) only serves as an estimate of the gate RC delay for a particular case. In another model approximation, the distributed gate resistance is replaced by a lumped resistance of $(\rho_g W/L)/3$ in front of a zero-resistance gate (Razavi *et al.*, 1994). That means when a step input from 0 to V_{dd} is applied, the gate voltage rises to $1 - e^{-1}$ or 0.63 of the full V_{dd} value in an RC delay time of $\rho_g C_{ox} W^2/3$. This result is more or less consistent with the previously described model, which gives a delay time of $\rho_g C_{ox} W^2/4$ [Eq. (5.29)] for the average gate voltage to reach $V_{dd}/2$. In practice, the gate is driven by a rising (or falling) signal with a finite ramp rate. Also, partial current conduction takes place early in the near end ($x = 0$) of the device, which constitutes the leading edge of signal propagation. Generally speaking, the gate RC delay depends on the drive condition of the previous stage and the capacitive loading of the following stage. Quantitative results should be obtained numerically from appropriate circuit models.

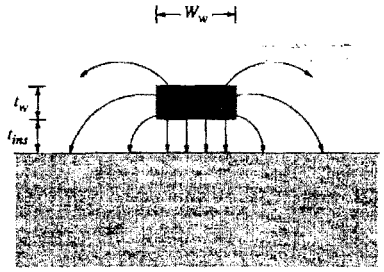


Figure 5.23. Schematic diagram showing electrostatic coupling between an isolated wire and a conducting plane. The straight field lines underneath the wire represent the parallel-plate component of the capacitance. The field lines emerging from the side and the top of the wire make up the fringing-field component of the capacitance. (After Bakoglu, 1990.)

5.2.4 Interconnect *R* and *C*

Unlike other parasitic elements discussed above, interconnect capacitance and resistance have negligible effects on the delay of local circuits such as CMOS inverters or NAND gates discussed in Section 5.1. On a VLSI chip or system level, however, interconnect R and C can play a major role in system performance, especially in standard-cell designs where wire capacitance dominates circuit delay. We shall discuss interconnect capacitance first, followed by interconnect resistance.

5.2.4.1 Interconnect Capacitance

Because of its geometry, the capacitance of an interconnect line cannot be calculated by the parallel-plate capacitance alone. In general, interconnect capacitance has three components: the parallel-plate (or area) component, fringing-field component, and wire-to-wire component. Figure 5.23 shows schematically electric field lines that constitute the parallel-plate and the fringing-field capacitance of an isolated line (Bakoglu, 1990). The total capacitance per unit length, C_w, calculated numerically by solving a 2-D Laplace's equation, is shown in Fig. 5.24 versus the ratio of wire width to insulator thickness, W_w/t_{ins} (Schaper and Amey, 1983). Only when $W_w \gg t_{ins}$ can the total capacitance per unit length be approximated by the parallel-plate component, $\varepsilon_{ins} W_w/t_{ins}$ (the straight chain line in Fig. 5.24). As $W_w/t_{ins} \to 1$, the fringing-field component becomes important and the total capacitance can be much higher than the parallel-plate component. In fact, a minimum capacitance of about 1 pF/cm (for silicon dioxide as the interlevel dielectric) is reached even if $W_w \ll t_{ins}$. This shows that reducing the wire capacitance by increasing the insulator thickness becomes ineffective when the insulator thickness becomes comparable to the width of the wire. Decreasing the wire thickness t_w does not help much either, as is evident in Fig. 5.24.

To improve the packing density in today's VLSI chips, wires of minimum pitch with nearly equal lines and spaces are frequently used. This causes a still higher wiring capacitance due to contributions from the neighboring lines. Figure 5.25 shows the calculated total capacitance as a sum of two components for an array of wires with

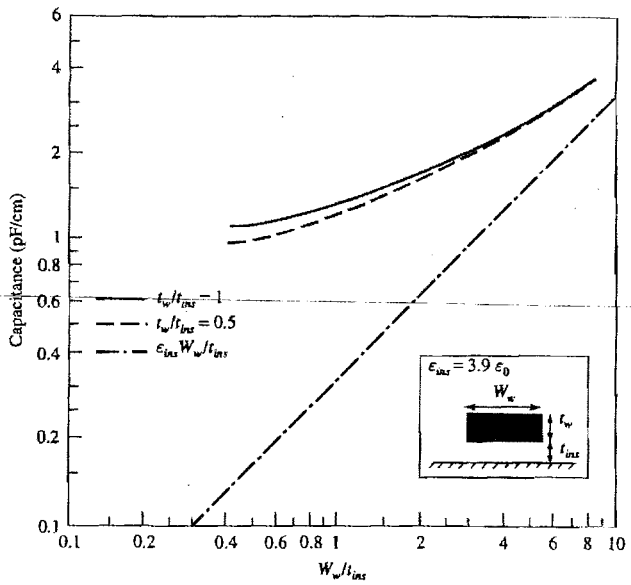


Figure 5.24. Wire capacitance per unit length as a function of width-to-gap ratio, W_w/t_{ins}, for the system in Fig. 5.23. The straight chain line represents the parallel-plate component of the capacitance. The dielectric medium is assumed to be oxide with a dielectric constant of 3.9. (After Schaper and Amey, 1983.)

the thickness of metal lines and the thickness of insulators below (oxide) and above (nitride) them are all assumed to be 1 μm. The capacitances are calculated as a function of the metal line or space dimension. When the metal line and space are much larger than the thicknesses, the capacitance is dominated by the component (parallel-plate plus fringing) to the conducting planes above and below. When the metal pitch is much smaller than the thicknesses, however, wire-to-wire capacitance dominates. ***The total capacitance exhibits a broad minimum value of about 2 pF/cm when the metal line or space dimension is approximately equal to the insulator (and wire) thickness***. This conclusion is more general than the specific dimensions assumed. If all the line, space, metal thickness, and insulator thickness are scaled by the same factor, the result remains unchanged. The number 2 pF/cm can be understood from the capacitance per unit length between two concentric cylinders of radii a and b:

$$C_w = \frac{2\pi\varepsilon_{ins}}{\ln(b/a)}. \tag{5.30}$$

If one takes $\varepsilon_{ins} = \varepsilon_{ox}$ and $b/a = 2$, then $C_w \approx 2\pi\varepsilon_{ox} \approx 2$ pF/cm. If an alternative insulator with a lower dielectric constant than that of oxide is used, C_w will decrease proportionally.

5.2.4.2 Interconnect Scaling

Based on the above discussions, one can easily set a strategy for interconnect scaling

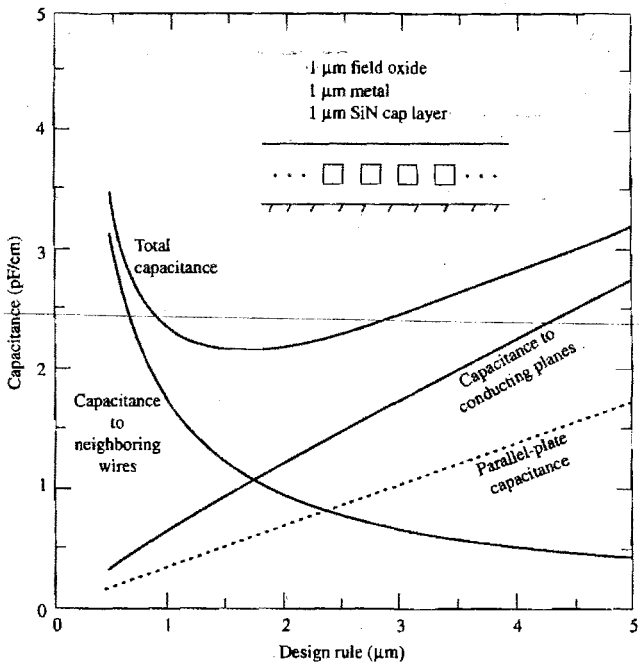


Figure 5.25. Capacitance per unit length as a function of design rules for an array of wires with equal line and space sandwiched between two conducting planes shown in the inset. The total capacitance of each wire is made up of two components: capacitance to the conducting planes, and capacitance to the neighboring wires. Both the metal and insulator thicknesses are held constant as the design rule is varied. The parallel-plate capacitance is also shown (dotted line) for reference. (After Schaper and Amey, 1983.)

schematically in Fig. 5.26 (Dennard *et al.*, 1974). ***All linear dimensions – wire length, width, thickness, spacing, and insulator thickness – are scaled down by the same factor, κ, as the device scaling factor.*** Wire lengths (L_w) are reduced by κ because the linear dimension of the devices and circuits that they connect to is reduced by κ. Both the wire and the insulator thicknesses are scaled down along with the lateral dimension, for otherwise the fringe capacitance and wire-to-wire coupling (crosstalk) would increase disproportionally, as illustrated in Fig. 5.25. Table 5.1 summarizes the rules for interconnect scaling. All material parameters, such as the metal resistivity ρ_w and dielectric constant ε_{ins}, are assumed to remain the same. The wire capacitance then scales down by κ, the same as the device capacitance (Table 4.1), while the wire capacitance per unit length, C_w, remains unchanged (approximately 2 pF/cm for silicon dioxide insulation, as mentioned above). The wire resistance, on the other hand, scales up by κ, in contrast to the device resistance, which does not change with scaling (Table 4.1). The wire resistance per unit length, R_w, then scales up by κ^2, as indicated in Table 5.1. It is also noted that the current density of interconnects increases with κ, which implies that

Table 5.1 Scaling of Local Interconnect Parameters

	Interconnect Parameters	Scaling Factor ($\kappa \geq 1$)
Scaling assumptions	Interconnect dimensions (t_w, L_w, W_w, t_{ins}, W_{sp})	$1/\kappa$
	Resistivity of conductor (ρ_w)	1
	Insulator permittivity (ε_{ins})	1
Derived wire scaling behavior	Wire capacitance per unit length (C_w)	1
	Wire resistance per unit length (R_w)	κ^2
	Wire RC delay (τ_w)	1
	Wire current density ($I/W_w t_w$)	κ

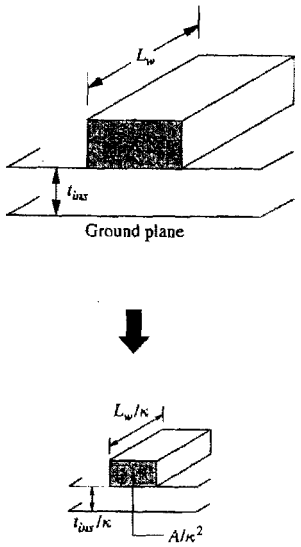


Figure 5.26. Scaling of interconnect lines and insulator thicknesses. (After Dennard, 1986.)

reliability issues such as electromigration may become more serious as the wire dimension is scaled down. In reality, a few material and process advances in metallurgy have taken place over the generations to keep electromigration under control in VLSI technologies.

5.2.4.3 Interconnect Resistance

The interconnect RC delay can be examined using the same distributed RC network model introduced in Section 5.2.3. From Fig. 5.22 or Eq. (5.28), the voltage at the receiving end of an interconnect line rises to $1 - e^{-1} \approx 63\%$ of the source voltage after a delay of $t = RCW^2/2$. If one takes this value as the equivalent RC delay (τ_w) of an interconnect line and substitutes R_w, C_w, L_w for R, C, W, one obtains

$$\tau_w = \frac{1}{2} R_w C_w L_w^2. \tag{5.31}$$

Using $R_w = \rho_w / W_w t_w$ and Eq. (5.30) for C_w with $\ln(b/a) \approx 1$, one can express Eq. (5.31) as

$$\tau_w \approx \pi \varepsilon_{ins} \rho_w \frac{L_w^2}{W_w t_w}, \tag{5.32}$$

where W_w and t_w are the wire width and thickness, respectively. One of the key conclusions of interconnect scaling is that the wire RC delay τ_w does not change as the device dimension and intrinsic delay are scaled down. Eventually, this will impose a limit on VLSI performance. Fortunately, for conventional aluminum metallurgy with silicon dioxide insulation, $\rho_w \approx 3 \times 10^{-6}\ \Omega$-cm and

$$\tau_w \approx (3 \times 10^{-18}\text{s}) \frac{L_w^2}{W_w t_w}. \tag{5.33}$$

It is easy to see that the RC delay of local wires is negligible as long as $L_w^2/W_w t_w < 3 \times 10^5$. For example, a 0.25 μm × 0.25 μm wire 100 μm long has an RC delay of 0.5 ps, which is quite negligible even when compared with the intrinsic delay (≈ 20 ps) of a 0.1-μm CMOS inverter (Taur *et al.*, 1993c). Therefore, ***a local circuit macro can be scaled down with all W_w, t_w, and L_w reduced by the same factor without running into serious RC problems.***

5.2.4.4 *RC* Delay of Global Interconnects

Based on the above discussion, the RC delay of local wires will not limit the circuit speed even though it cannot be reduced through scaling. The RC delay of global wires, on the other hand, is an entirely different matter. Unlike local wires, the length of global wires, on the order of the chip dimension, does not scale down, since the chip size actually increases slightly for advanced technologies with better yield and defect density to accommodate a much larger number of circuit counts. Even if we assume the chip size does not change, the RC delay of global wires scales up by κ^2 from Eq. (5.33). It is clear that one quickly runs into trouble if the cross-sectional area of global wires is scaled down the same way as the local wires. For example, in a 0.25-μm CMOS technology, $L_w^2/W_w t_w \sim 10^8$–10^9 and $\tau_w \sim 1$ ns, severely degrading the system performance. The use of copper wires instead of aluminum would reduce the numerical factor in Eq. (5.33) by a factor of about 1.5 and provide some relief.

A number of solutions have been proposed to deal with the problem. The most obvious one is to minimize the number of cross-chip global interconnects in the critical paths through custom layout/design and use of sophisticated design tools. One can also use repeaters to reduce the dependence of RC delay on wire length from a quadratic one to a linear one (Bakoglu, 1990). A more fundamental solution is to increase or not to scale the cross-sectional area of global wires. However, just increasing the width and thickness of global wires is not enough, since the wire

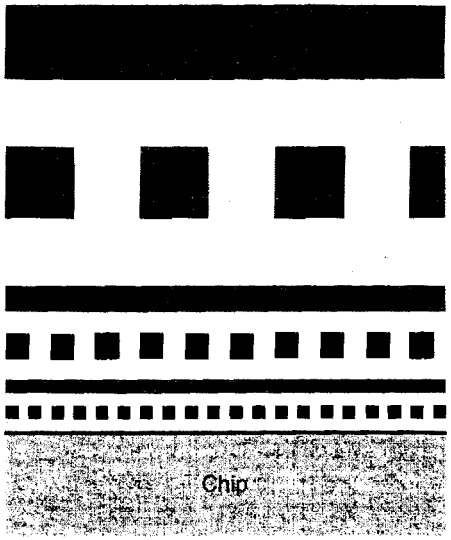


Figure 5.27. Schematic cross section of a wiring hierarchy that addresses both the density and the global *RC* delay in a high-performance CMOS processor. (After Sai-Halasz, 1995.)

capacitance will then increase significantly, which degrades both performance and power. The intermetal dielectric thickness must be increased in proportion to keep the wire capacitance per unit length constant. Of course, there is a technology price to pay in building such low-*RC* global wires. It also means more levels of interconnects, since one still needs several levels of thin, dense local wires to make the chip wirable.

The best strategy for interconnect scaling is then to scale down the size and spacing of lower levels in step with device scaling for local wiring, and to use unscaled or even scaled-up levels on top for global wiring, as shown schematically in Fig. 5.27 (Sai-Halasz, 1995). Unscaled wires allow the global *RC* delay to remain essentially unchanged, as seen from Eq. (5.33). Scaled-up (together with the insulator thickness) wires allow the global *RC* delay to scale down together with the device delay. This is even more necessary if the chip size increases with every generation. Ultimately, the scaled-up global wires will approach the transmission-line limit when the inductive effect becomes more important than the resistive effect. This happens when the signal rise time is shorter than the time of flight over the length of the line. Signal propagation is then limited by the speed of electromagnetic waves, $c/(\varepsilon_{ins}/\varepsilon_0)^{1/2}$, instead of by *RC* delay. Here $c=3\times10^{10}$ cm/s is the velocity of light in vacuum. For oxide insulators, $(\varepsilon_{ins}/\varepsilon_0)^{1/2}\approx2$, the time of flight is approximately 70 ps/cm. Figure 5.28 shows the interconnect delay versus wire length L_w calculated from Eq. (5.33) for three different wire cross sections. Note that the *RC* delays vary quadratically with L_w. Below a certain wire length, the delay is limited by the time of flight which varies linearly with L_w. For a longer global wire to reach the speed-of-light limit, a larger wire cross section is needed. The transmission-line situation is more often encountered in packaging wires (Bakoglu, 1990).

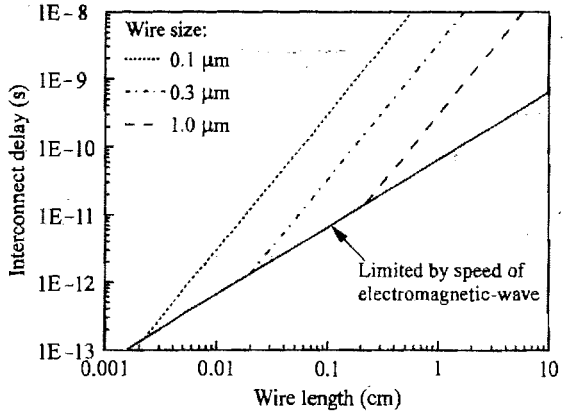


Figure 5.28. *RC* delay versus wire length for three different wire sizes (assuming square wire cross sections). Wires become limited by electromagnetic-wave propagation when the *RC* delay equals the time of flight, $(\varepsilon_{ins}/\varepsilon_0)^{1/2}L_w/c$, over the line length L_w. An oxide insulator is assumed here.

5.3 Sensitivity of CMOS Delay to Device Parameters

This section focuses on the performance factors of basic CMOS circuit elements and their sensitivities to both the intrinsic device parameters and the parasitic resistances and capacitances. Using 1.5 V, 0.1 μm CMOS devices as an example, we first define the propagation delay of an inverter chain and discuss the loading effect due to fan-out and wiring capacitances. Three performance factors – the switching resistance R_{sw}, input capacitance C_{in}, and output capacitance C_{out} – are introduced in terms of a delay equation, followed by several subsections detailing their sensitivity to various device parameters. The last subsection deals with the performance factors of two-way NAND circuits.

5.3.1 Propagation Delay and Delay Equation

In this subsection, we define the propagation delay and the delay equation of a static CMOS gate. While CMOS inverters are used as an example to build the basic framework, most of the formulation and performance factors are equally applicable to other NAND and NOR circuits that perform more general logic functions.

5.3.1.1 Propagation Delay of a CMOS Inverter Chain

The basic switching characteristics of a CMOS inverter with a step input waveform have been briefly touched upon in Section 5.1.1. In a practical logic circuit, a CMOS inverter is driven by the output from a previous stage whose waveform has a finite rise or fall time associated with it. One way to characterize the switching delay or the performance of an inverter is to construct a cascaded chain of identical inverters as shown in Fig. 5.29, and consider the propagation delay of a logic signal going through them. Load capacitors

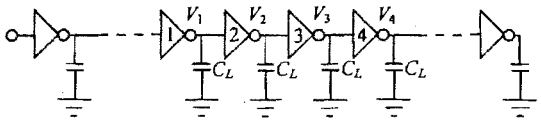


Figure 5.29. A linear chain of CMOS inverters (fan-out = 1). Each triangular symbol represents a CMOS inverter consisting of an nMOSFET and a pMOSFET as shown in Fig. 5.2. Power-supply connections are not shown.

can be added to the output node of each inverter to simulate the wiring capacitance it may drive in addition to the next inverter.

For a given CMOS technology, the propagation delay is experimentally determined by constructing a *ring oscillator* with a large, odd number of CMOS inverters connected head to tail and measuring the oscillating frequency of the signal at any given point in the ring when the power-supply voltage is applied. The sustained oscillation is a result of propagation of alternating logic states ($0 \rightarrow V_{dd} \rightarrow 0 \rightarrow \ldots$) around a ring with an odd number of stages. The period of the oscillation is given by $n(\tau_n + \tau_p)$, where n is the number of stages (an odd number) and τ_n, τ_p are inverter delays per stage for rising and falling inputs, respectively. In other words, in one period the logic signal propagates around the ring twice.

Because of the complexity of the current expressions for short-channel MOSFETs and the voltage dependence of both intrinsic and extrinsic capacitances, a circuit model such as BSIM in SPICE is needed to solve the propagation delay numerically (Cai *et al.*, 2000). In order to gain insight into how the voltage and current waveforms look during a switching event, we consider the example of a 0.1 μm CMOS inverter with the device parameters listed in Table 5.2 (http://www.eas.asu.edu/~ptm/). All lithography dimensions and contact borders, e.g., a, b, and c in Fig. 5.13, are assumed to be 0.15 μm (nonfolded). The power-supply voltage is 1.5 V (Taur *et al.*, 1993c).

The propagation delay is evaluated by introducing a step voltage signal at the input of the linear inverter chain in Fig. 5.29. After a few stages, the signal waveform has become a *standardized signal*, i.e., one that has stabilized and remains a constant shape independent of the number of stages of propagation. There are also a few stages following the ones of interest for maintaining the same capacitive loading of each stage.

For any stage with input voltage V_{in} and output voltage V_{out} (see Fig. 5.2),

$$C\frac{dV_{out}}{dt} = I_P(V_{in}, V_{out}) - I_N(V_{in}, V_{out}), \tag{5.34}$$

where C lumps all the capacitances connected to the output node. (Capacitance components to a node with time-varying voltage are discussed in Section 5.3.4.3.) If $V_{in}(t)$ is known, then $V_{out}(t)$ can be solved from the above differential equation. Numerically, given V_{in} and V_{out} at any time instant t, I_P and I_N can be evaluated, and the next V_{out} is given by

$$V_{out}(t + \Delta t) = V_{out}(t) + \frac{I_P - I_N}{C}\Delta t. \tag{5.35}$$

A $V_{out}(t)$ curve can be generated by repeating these steps.

Table 5.2 0.1 μm CMOS parameters for circuit modeling (25 °C)

Assumed	Power supply voltage, V_{dd} (V)	1.5	
	Channel length, L (μm)	0.1	
	Lithography ground rules, a,b,c (μm)	0.15	
	Gate oxide thickness, t_{ox} (nm)	3.6	
	Linearly extrapolated threshold voltage, V_{on} (V)	±0.4	
		nMOSFET	pMOSFET
	Source and drain series resistance, R_{sd} (Ω-μm)	200	500
	Saturation velocity, v_{sat} (cm/s)	10^7	10^7
	Substrate/well doping concentration, N_a, N_d (cm^{-3})	10^{18}	10^{18}
	Gate to source or drain (per edge) overlap capacitance, C_{ov} (fF/μm)	0.3	0.3
	Drain induced barrier lowering, ΔV_t between $V_{ds} = 0$ and $V_{ds} = V_{dd}$, (V)	0.05	0.11
	Body-effect coefficient, m	1.3	1.3
	Device width, W_n, W_p (μm)	1	2
Computed	Intrinsic channel capacitance per unit width, $\varepsilon_{ox}L/t_{ox}$ (fF/μm)	0.96	0.96
	Source and drain junction capacitance (@ $V_j = 0$, based on N_a, N_d), C_j (fF/μm^2)	2.8	2.8
	Effective mobility (@ $V_{gs} = V_{dd}/2$), μ_{eff} (cm^2/V-s)	375	85
	On current (@ $V_{gs} = V_{ds} = V_{dd}$), I_{on} (mA/μm)	0.56	0.25
	Off current (@$V_{gs} = 0$ and $V_{ds} = V_{dd}$), I_{off} (nA/μm)	0.1	0.5

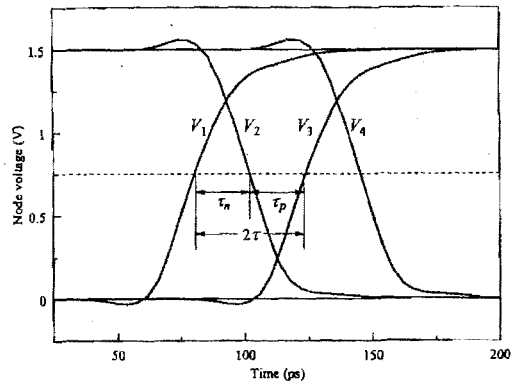


Figure 5.30. Successive voltage waveforms of the CMOS inverter chain in Fig. 5.29 ($C_L = 0$). The delay is measured by their intersections with the $V_{dd}/2$ (dashed) line as shown.

Figure 5.30 shows an example of the waveforms at four successive stages, V_1, V_2, V_3, V_4, for the unloaded case, $C_L = 0$. As V_1 rises, the nMOSFET of inverter 2 is turned on, which pulls V_2 to ground. The fall of V_2 then turns on the pMOSFET of inverter 3, which causes V_3 to rise, and so on. If one draws a straight line through the midpoints of all the waveforms at $V = V_{dd}/2$, one can define the pull-down propagation delay τ_n as the time interval between V_1 and V_2 along that line. Similarly, the pull-up propagation delay τ_p is the time interval between V_2 and V_3 along $V = V_{dd}/2$. The definitions here are consistent

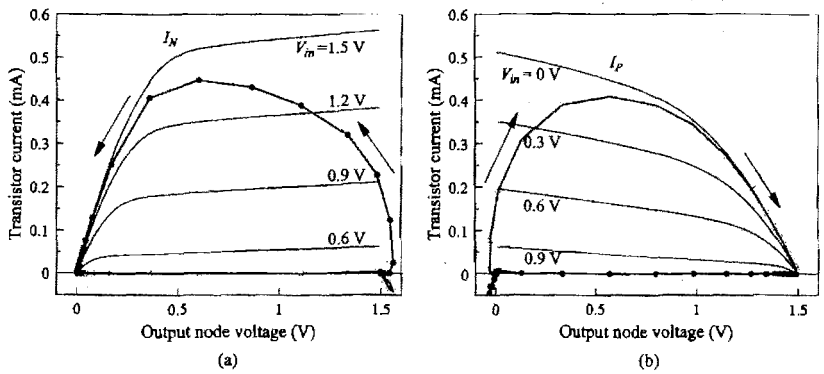


Figure 5.31. Bias-point switching trajectories of n- and pMOSFETs for (a) pull-down transition of node V_2 in inverter 2, and (b) pull-up transition of node V_3 in inverter 3. Solid dots are for I_N and open circles for I_P in both (a) and (b). The bias points are plotted in equal time intervals, which for the unloaded case ($C_L = 0$) are 5 ps apart.

with those in Fig. 5.8 for step inputs. A better-defined quantity is the CMOS propagation delay, $\tau = (\tau_n + \tau_p)/2$, which is one-half the time delay between the parallel waveforms V_1 and V_3 or V_2 and V_4. The time τ is also the delay measured experimentally from CMOS ring oscillators. It is equal to the oscillation period divided by twice the number of stages as stated before. In this specific example, the device width ratio is chosen to be $W_p/W_n = 2$ so that the pull-down delay equals the pull-up delay, i.e., $\tau_n = \tau_p = \tau$. In general, τ_n and τ_p may be different from each other, and the CMOS delay may be dominated by either the nMOSFET or the pMOSFET.

5.3.1.2 Bias-Point Trajectories in a Switching Event

As the logic signal arrives at the input gate of an inverter, a transient current flows in either the nMOSFET or the pMOSFET of that inverter until the output node completes its high-to-low or low-to-high transition. It is instructive to examine the bias-point trajectories through a family of I_{ds}–V_{ds} curves during a pull-down or pull-up switching event. Figure 5.31(a) plots the trajectories of nMOSFET (solid dots) and pMOSFET (open circles) currents of inverter 2 in Fig. 5.29 versus the output node voltage V_2, as V_2 is pulled down from V_{dd} to ground. The points are plotted in constant 5-ps time intervals over a background of nMOSFET $I_{ds} - V_{ds}$ curves (I_N). The pMOSFET current is very low throughout the transition, indicating negligible power dissipation from the crossover currents between the power-supply terminal and the ground. The output node, initially at V_{dd}, is discharged by the nMOSFET current, which reaches its highest value midway during switching when $V_2 \approx V_{dd}/2$ and $V_{in} \approx 0.9V_{dd}$. This point is also where the voltage waveform V_2 in Fig. 5.30 exhibits the maximum downward slope. The peak current is typically 80–90% of the maximum on current at $V_{gs} = V_{ds} = V_{dd}$. The exact percentage depends on the detailed device parameters such as mobility, velocity saturation, threshold voltage, and series resistance. Likewise, Fig. 5.31(b) shows the

bias-point trajectories of both transistors in inverter 3 when the output node V_3 is pulled up from ground to V_{dd}. In this case, the nMOSFET current (solid dots) is negligible, while the pMOSFET current (open circles) reaches its peak value when $V_3 \approx V_{dd}/2$, as in the pull-down case. The two bias trajectories are basically similar to each other and are insensitive to loading conditions. At larger C_L, the delay time between the points in the trajectory increases, but the shape of the curve remains essentially the same.

The delay per stage, as defined in Fig. 5.30, is the time duration between $V_{in} = V_{dd}/2$ and $V_{out} = V_{dd}/2$. From Fig. 5.30, it is clear that when $V_{in} = V_{dd}/2$, V_{out} is just about to start switching from its prior steady-state value. During either the pull-down delay τ_n or the pull-up delay τ_p, V_{out} changes by $\approx V_{dd}/2$. In the pull-down transition, for example, I_P is negligible. Equation (5.34) can be integrated to yield

$$\tau_n = -\int_{V_{dd}}^{V_{dd}/2} \frac{C dV_{out}}{I_N} = \frac{CV_{dd}/2}{\langle I_N \rangle}, \tag{5.36}$$

where $1/\langle I_N \rangle$ is defined as the average reciprocal of the nMOSFET current between $V_{in} = V_{dd}/2$ and $V_{out} = V_{dd}/2$. In general, the capacitance C may have some weak dependence on V_{in} and V_{out}. Here, one can re-define C and $\langle I_N \rangle$ to absorb that effect. Likewise,

$$\tau_p = \int_0^{V_{dd}/2} \frac{C dV_{out}}{I_P} = \frac{CV_{dd}/2}{\langle I_P \rangle} \tag{5.37}$$

for pMOSFET pull-up when I_N is negligible. In the step input case discussed in Section 5.1.1.3, the input rise or fall time is zero, and $\langle I_N \rangle$ or $\langle I_p \rangle$ equals the on-current I_{onN} or I_{onP}, respectively. For the propagation delay considered here, $\langle I_N \rangle$ or $\langle I_P \rangle$ is typically about 3/5 of the on-currents, as can be visually estimated from Fig. 5.31 (average current between the point where $V_{gs} = V_{in} = 0.75$ V in the trajectory and the point where $V_{ds} = V_{out} = 0.75$ V). A semi-empirical expression that has been reported to work well is $\langle I_N \rangle = (1/2)\{I_N(V_{in} = V_{dd}/2, V_{out} = V_{dd}) + I_N(V_{in} = V_{dd}, V_{out} = V_{dd}/2)\}$, and vice versa for $\langle I_P \rangle$ (Na *et al.*, 2002).

From Eqs. (5.36) and (5.37), CMOS inverter propagation delay can be written as

$$\tau = \frac{\tau_n + \tau_p}{2} = \frac{CV_{dd}}{4}\left(\frac{1}{\langle I_N \rangle} + \frac{1}{\langle I_P \rangle}\right). \tag{5.38}$$

5.3.1.3 Delay Equation: Switching Resistance, Input and Output Capacitance

The simulations plotted in Fig. 5.30 and Fig. 5.31 are for the unloaded case with fan-out = 1. In general, $C_L \neq 0$, and the output of an inverter may drive more than one stage. In the latter case the fan-out is 2, 3, ..., which means that each inverter in the chain is driving 2, 3, ... stages in parallel. Each receiving stage is assumed to have the same widths as the sending stage. There are also situations where an inverter is driving another stage wider than its own widths. Such cases can be covered mathematically by generalizing the definition of 'fan-out' to include nonintegral numbers, provided that the same n- to p-width ratio is

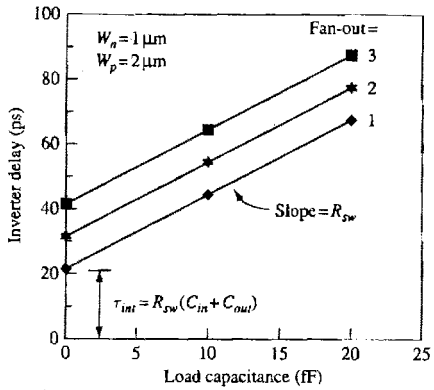


Figure 5.32. Inverter delay τ versus load capacitance C_L for fan-out of 1, 2, and 3.

always maintained. Fan-outs greater than 3 are rarely used in CMOS logic circuits, as they lead to significantly longer delays.

Figure 5.32 plots the inverter delay τ versus the load capacitance C_L for fan-out = 1, 2, and 3, simulated with the device parameters in Table 5.2. Equation (5.35) indicates that the time scale or the delay should scale linearly with the capacitive loading C. This is reflected in Fig. 5.32 that, for each fan-out, the delay increases linearly with C_L with a constant slope independent of fan-out. The intercept with the y-axis, i.e., the delay at $C_L = 0$, in turn increases linearly with the fan-out. These facts can be summarized in a general delay equation (Wordeman, 1989),

$$\tau = R_{sw} \times (C_{out} + \text{FO} \times C_{in} + C_L), \tag{5.39}$$

where FO represents the fan-out. In this way, the *switching resistance* R_{sw} is defined as the slope of the delay-versus-load-capacitance lines in Fig. 5.32, $d\tau/dC_L$. It is a direct indicator of the current drive capability of the logic gate. The *output capacitance* C_{out} represents the equivalent capacitance at the output node of the sending stage, which usually consists of the drain junction capacitance and the drain-to-gate capacitance including the overlap capacitance. C_{out} depends on the layout geometry. The *input capacitance* C_{in} is the equivalent capacitance presented by one-unit (FO = 1) input-gate widths of the receiving stage to the sending stage. C_{in} consists of the gate-to-source, gate-to-drain, and gate-to-substrate capacitances including both the intrinsic and the overlap components. Some of the capacitance components are subject to the Miller effect, discussed later in Section 5.3.4. The minimum unloaded delay at $C_L = 0$, or the intrinsic delay, is given by

$$\tau_{int} = R_{sw}(C_{in} + C_{out}), \tag{5.40}$$

which is 22 ps for the 0.1-μm CMOS inverter shown in Fig. 5.32.

The delay equation (5.39) not only allows the delay to be calculated for any fan-out and loading conditions but also decouples the two important factors that govern

Table 5.3 Extracted Performance Factors of the 0.1-μm CMOS in Table 5.2

nMOSFET switching resistance, $W_n R_{swn}$ (Ω-μm)	2300
pMOSFET switching resistance, $W_p R_{swp}$ (Ω-μm)	4600
Input capacitance, $C_{in}/(W_n + W_p)$ (fF/μm)	1.4
Output capacitance, $C_{out}/(W_n + W_p)$ (fF/μm)	1.7

CMOS performance: current and capacitance. Current drive capability is represented by R_{sw}, which is inversely proportional to the large-signal transconductance I_{on}/V_{dd} appropriate for digital circuits (Solomon, 1982). The switching resistance can be decomposed into R_{swn} and R_{swp} in terms of the pull-down and pull-up delays τ_n and τ_p defined in Fig. 5.30, i.e., $R_{swn} \equiv d\tau_n/dC_L$ and $R_{swp} \equiv d\tau_p/dC_L$. Since $\tau = (\tau_n + \tau_p)/2$, it follows that $R_{sw} = (R_{swn} + R_{swp})/2$. From Eqs. (5.36) and (5.37),

$$R_{swn} = \frac{V_{dd}/2}{\langle I_N \rangle} \tag{5.41}$$

and

$$R_{swp} = \frac{V_{dd}/2}{\langle I_p \rangle}, \tag{5.42}$$

where $\langle I_N \rangle$ and $\langle I_P \rangle$ are about 3/5 of the on-currents at $V_{gs} = V_{ds} = \pm V_{dd}$, as stated before.

The switching resistances extracted from the above specific example are listed in Table 5.3. For the CMOS inverters, W_p/W_n was chosen to be 2 to compensate for the difference between $I_{on,n}$ and $I_{on,p}$, so that $R_{swn} \approx R_{swp} \approx R_{sw}$ and $\tau_n \approx \tau_p \approx \tau$.

Both the input and the output capacitances, C_{in} and C_{out}, in Eq. (5.39) are approximately proportional to $W_n + W_p$, since both nMOSFET and pMOSFET contribute more or less equally per unit width to the node capacitance whether they are being turned on or being turned off. This assumes that all the capacitances per unit width are symmetrical between the n- and p-devices, as is the case in Table 5.2. The specific numbers for the case in Fig. 5.32 are listed in Table 5.3. Note that $(C_{in} + C_{out})/(W_n + W_p)$ is about three times the intrinsic channel capacitance per unit width, 0.96 fF/μm, listed in Table 5.2.

5.3.1.4 CMOS Delay Scaling

It is instructive to reexamine, from the delay-equation point of view, how CMOS performance improves under the rules of constant-field scaling outlined in Section 4.1.1. Let us assume that the first five parameters in Table 5.2 are scaled down by a factor of two, i.e., $V_{dd} = 0.75$ V, $L = 0.05$ μm, $t_{ox} = 1.8$ nm, $V_{on} = \pm 0.2$ V, and a, b, $c = 0.075$ μm (lithography ground rules). If the source and drain series resistances in the scaled CMOS are also reduced by a factor of two, i.e., $R_{sdn} = 100$ Ω-μm and $R_{sdp} = 250$ Ω-μm, the on currents per unit device width will remain essentially unchanged, i.e., $I_{on,n} = 0.56$ mA/μm and $I_{on,p} = 0.25$ mA/μm (both the mobility and the saturation velocity are the same as

before). Since V_{dd} is reduced by a factor of two, both n- and p-switching resistances normalized to unit device width, W_nR_{swn} and W_pR_{swp}, improve by a factor of two. At the same time, all the capacitances per unit width should be kept the same. These include the gate capacitance $\varepsilon_{ox}L/t_{ox}$, the overlap capacitance (0.3 fF/μm), and the junction capacitance. Note that the junction capacitance per unit area, C_j, may go up by a factor of two due to the higher doping needed to control the short-channel effect, but the junction capacitance per unit device width is proportional to $(a + b + c)C_j$ and therefore remains unchanged. Combining all the above factors, one obtains that both $C_{in}/(W_n + W_p)$ and $C_{out}/(W_n + W_p)$ are unchanged and the intrinsic delay given by Eq. (5.40) improves by a factor of two to 11 ps.

In practice, one cannot follow the above ideal scaling for various reasons. The most important one is that the threshold voltage cannot be reduced without a substantial increase in the off current, as discussed extensively in Section 4.2. A more detailed tradeoff among CMOS performance, active power, and standby power will be considered in Section 5.3.3.

5.3.2 Delay Sensitivity to Channel Width, Length, and Gate Oxide Thickness

The next few subsections examine CMOS delay sensitivity to various device parameters, both intrinsic and parasitic, as listed in Table 5.2. To begin with, this subsection discusses the effect of device width, channel length, and gate oxide thickness on CMOS performance.

5.3.2.1 CMOS Delay Sensitivity to pMOSFET/nMOSFET Width Ratio

When the p- to n-device width ratio W_p/W_n is varied in a CMOS inverter, the relative current drive capabilities R_{swn} and R_{swp}, and therefore τ_n and τ_p, also vary. Figure 5.33 plots the intrinsic delay (FO = 1, C_L= 0) of CMOS inverters as a function of the device

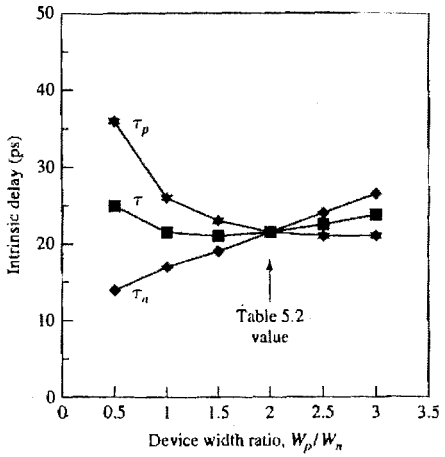


Figure 5.33. Intrinsic CMOS inverter delays τ_n, τ_p, and τ for FO = 1 and C_L = 0 versus p- to n-device width ratio.

width ratio. The rest of the device parameters are the same as in Table 5.2. As W_p/W_n increases, τ_p decreases but τ_n increases. At $W_p/W_n \approx 2$, the pull-up time becomes equal to the pull-down time, which gives the best noise margin, as discussed in Section 5.1.1. The overall delay, $\tau = (\tau_n + \tau_p)/2$, on the other hand, is rather insensitive to the width ratio, showing a shallow minimum at $W_p/W_n \approx 1.5$. The specific example in the last subsection used $W_p/W_n = 2$, so that $\tau_n \approx \tau_p \approx \tau = 22$ ps, which is within 5% of the minimum delay at W_p/W_n = 1.5. It should be noted that only the intrinsic or unloaded delay exhibits a minimum at W_p/W_n= 1.5. The minimum delay for wire-loaded circuits tends to occur at a larger W_p/W_n ratio.

5.3.2.2 Device Width Effect with Respect to Load Capacitance

From the discussions in Section 5.3.1, it is clear that if W_n and W_p are scaled up by the same factor without changing the ratio W_p/W_n, the intrinsic delay remains the same. The switching resistance, R_{sw}= $d\tau/dC_L$, however, is reduced by that same factor. So for a given capacitive load C_L, the delay improves. In fact, it has been argued that for high-performance purposes, one can scale up the device size until the circuit delays are mostly device-limited, i.e., approaching intrinsic delays (Sai-Halasz, 1995). This can be accomplished, if necessary, by increasing the chip size, because the capacitance due to wire loading increases only as the linear dimension of the chip (2 pF/cm in Section 5.2.4), while the effective device width can increase as the area of the chip if one uses corrugated (folded) gate structures. Of course, delays of global interconnects, as well as chip power and cost, will go up as a result.

In practical CMOS circuits, one tries to avoid the situation where a device drives a capacitive load much greater than its own capacitance, as that results in delays much longer than the intrinsic delay. One solution is to insert a *buffer*, or *driver*, between the original sending stage and the load. A driver consists of one or multiple stages of CMOS inverters with progressively wider widths. To illustrate how it works, we consider an inverter with a switching resistance R_{sw}, an input capacitance C_{in}, and an output capacitance C_{out}, driving a load capacitance C_L. Without any buffer, the single-stage delay is

$$\tau = R_{sw}(C_{out} + C_L). \tag{5.43}$$

If $C_L \gg C_{in}$ and C_{out}, the delay may be improved by inserting an inverter with k (>1) times wider widths than the original inverter. Such a buffer stage would present an equivalent FO = k to the sending stage but would have a much improved switching resistance, R_{sw}/k. The overall delay including the delay of the buffer stage would be[1]

$$\begin{aligned}\tau_b &= R_{sw}(C_{out} + kC_{in}) + \frac{R_{sw}}{k}(kC_{out} + C_L) \\ &= R_{sw}\left(2C_{out} + kC_{in} + \frac{C_L}{k}\right).\end{aligned} \tag{5.44}$$

It is easy to see that the best choice of the buffer width is $k = (C_L/C_{in})^{1/2}$, which yields a minimum delay of

[1] Here we apply Eq. (5.39) as an approximation. Strictly speaking, it is not propagation delay without a few repeated stages of identical driving-receiving conditions.

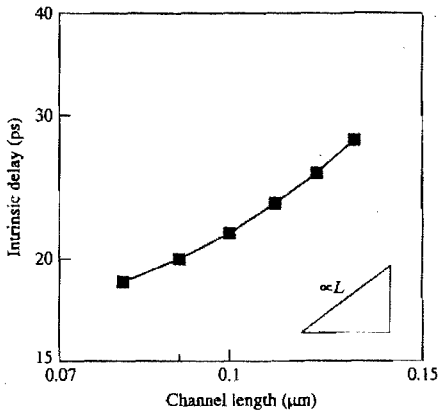


Figure 5.34. Intrinsic CMOS inverter delay versus channel length for the devices listed in Table 5.2. Both n- and pMOSFETs are assumed to have the same channel length.

$$\tau_{b\min} = R_{sw}\left(2C_{out} + 2\sqrt{C_{in}C_L}\right). \tag{5.45}$$

For heavy loads ($C_L \gg C_{in}$, C_{out}), $\tau_{b\min}$ can be substantially shorter than the unbuffered delay τ. To drive even heavier loads, multiple-stage buffers can be designed for best results (see Exercises 5.8, 5.9, 5.10).

5.3.2.3 Sensitivity of Delay to Channel Length

Channel length offers the biggest lever for CMOS performance improvement. At shorter channel lengths, not only does the switching resistance of the driving stage decrease due to higher on-currents, the intrinsic capacitance in the receiving stage is also lower. Figure 5.34 shows the variation of inverter delay with channel length assuming the rest of the device parameters are given by Table 5.2 (with no threshold voltage dependence on channel length). It is observed that the inverter delay improves approximately linearly with channel length at and above the 0.1-μm design point, but sub-linearly below it.

5.3.2.4 Sensitivity of Delay to Gate Oxide Thickness

Switching resistance or current drive capability can also be improved by using a thinner gate oxide. In contrast, however, to shortening the channel length, which helps both the resistance and the capacitance, a thinner oxide leads to a higher gate capacitance. It is shown in Fig. 5.35 that the improvement of intrinsic delay with oxide thickness is not as much as with channel length. Loaded delays improve more as indicated by the switching resistance curve in Fig. 5.35. The R_{sw} dependence on t_{ox} is still sub-linear because mobility decreases in thinner-oxide devices due to the higher vertical field.

It should be pointed out that the above sensitivity study only considers t_{ox} variations at the level of the circuit model, while keeping all other parameters unchanged. In other words, the interdependence between t_{ox} and V_t or L at the process or device level is not

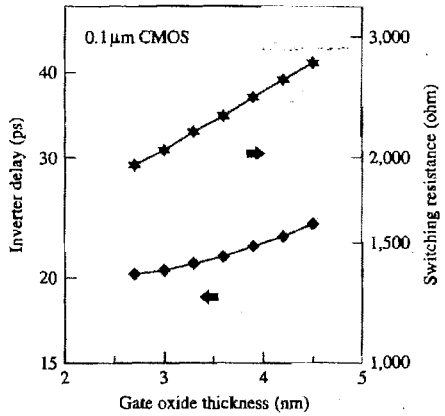


Figure 5.35. Intrinsic delay and switching resistance versus gate oxide thickness for the 0.1-μm CMOS listed in Table 5.2. Both log scales are of the same proportion for comparison.

taken into account. From a device-design point of view, thinner oxides would allow shorter channel lengths and therefore additional performance benefit.

5.3.3 Sensitivity of Delay to Power-Supply Voltage and Threshold Voltage

This subsection addresses the dependence of CMOS delay on power-supply voltage and threshold voltage. The effect is mainly through the switching resistance factor as the large-signal transconductance, I_{on}/V_{dd}, degrades with higher V_t or lower V_{dd}. Both the input and output capacitances are relatively insensitive to V_{dd} and V_t. The effect of threshold voltage on the delay of 0.1-μm CMOS for a given $V_{dd} = 1.5$ V was discussed in Subsection 4.2.1.3 and shown in Fig. 4.2. In that case, the delay for $V_t/V_{dd} < 0.5$ can be fitted to an empirical factor, $1/(0.6 - V_t/V_{dd})$. The dependence of inverter delay on power supply voltage for a fixed threshold voltage (Table 5.2) is shown in Fig. 5.36. The delay increases more rapidly than $1/(0.6 - V_t/V_{dd})$ as the supply voltage is reduced, indicating that while the factor $1/(0.6 - V_t/V_{dd})$ captures the V_t-dependence of the delay, there is additional V_{dd}-dependence. The delays of 2-way NAND gates exhibit a very similar V_{dd}-dependence as the inverter delay. More discussions on 2-way NAND delays can be found in Subsection 5.3.5.

5.3.3.1 Power and Delay Tradeoff

The delay versus supply voltage curve in Fig. 5.36 can be re-plotted as a power versus delay curve with V_{dd} as a parameter in Fig. 5.37. Here the active power is calculated from

$$P_{ac} = (C_{in} + C_{out})V_{dd}^2/(2\tau), \tag{5.46}$$

under the assumption that the inverters are clocked at the highest frequency possible, $f = 1/(2\tau)$, where 2τ is the time it takes to complete a high-to-low-to-high switching cycle

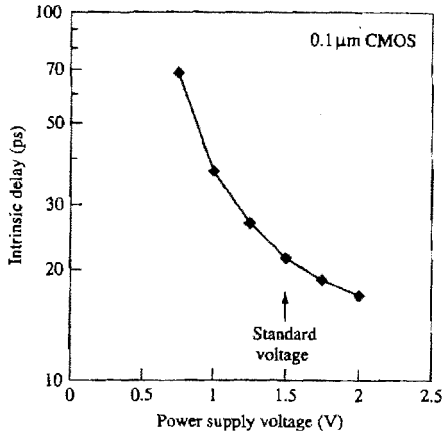


Figure 5.36. CMOS intrinsic delay versus power supply voltage for a constant threshold voltage (Table 5.2).

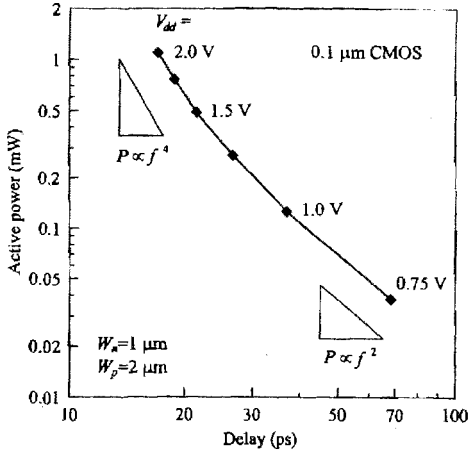


Figure 5.37. CMOS power versus delay by varying the power supply voltage for a constant threshold voltage (Table 5.2).

(Fig. 5.30). Equation. (5.46) accounts for about 90% of the power drained from the power supply source (rail to rail current times V_{dd}). The rest is the cross-over or short-circuit power. For the devices in Table 5.2, the standby power due to subthreshold leakage at room temperature is about 1 nW, negligible during the active switching transient. In Fig. 5.37, lower power-delay product or switching energy is obtained at low supply voltages where $P \propto f^2$. ***For high-performance CMOS operated toward the high end of the supply voltage, premium performance comes at a steep expense of active power*** ($P \propto f^4$).

It is possible to reduce V_{dd} without a severe loss in performance if V_t is reduced as well. Of course, standby power will go up as a result. The tradeoff among performance, active power, and standby power is depicted conceptually in a V_{dd}–V_t design plane in Fig. 4.8. While the standby portion of the total power stays constant with time, the active portion of the total power depends on the circuit activity factor, i.e., how often the circuit switches on average. For high-activity circuits such as clock drivers, active power dominates. In principle, their power can be reduced by operating at low V_{dd} and low V_t while maintaining a similar performance (Cai *et al.*, 2002b). The majority of circuits in a typical VLSI logic chip, however, are of the low-activity type, such as those found in static memories. High-V_t devices are needed in those circuits to limit their collective standby power. High V_{dd} may also be needed for performance. In practice, circuits of different logic swings are rarely mixed in the same chip (except for input from and output to other systems of different voltage level) due to delay and area penalties associated with level translation at their interfaces.

5.3.4 Sensitivity of Delay to Parasitic Resistance and Capacitance

This subsection examines the sensitivity of CMOS delay to parasitic source–drain series resistance, overlap capacitance, and junction capacitance, using the 0.1-μm devices listed in Table 5.2 as an example.

5.3.4.1 Sensitivity of Delay to Series Resistance

The effect of source–drain series resistance on CMOS delay comes through n- and pMOSFET currents and therefore their switching resistances. Figure 5.38 shows the sensitivity of n- and p-switching resistances to the n- and p-series resistances R_{sdn} and R_{sdp}. Since pMOSFETs have a lower current per unit width, they can tolerate a higher series resistance for the same percentage of degradation. For the default values assumed

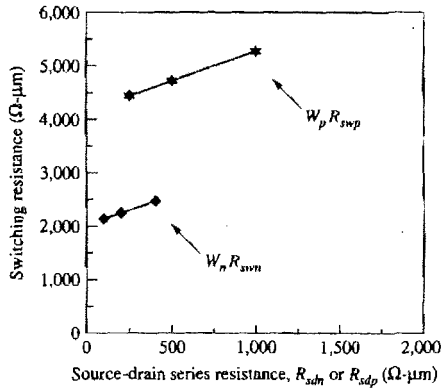


Figure 5.38. Switching resistances versus source–drain series resistance for the 0.1-μm CMOS listed in Table 5.2. $W_n R_{swn}$ is plotted vs. R_{sdn} and $W_p R_{swp}$ is plotted vs. R_{sdp}.

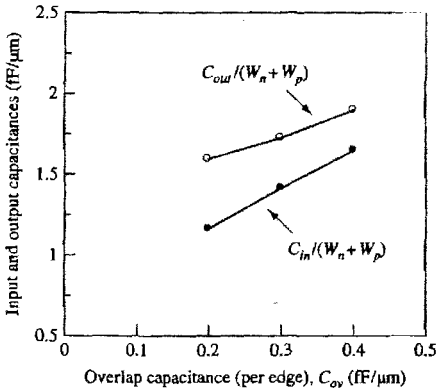


Figure 5.39. Input and output capacitances versus overlap capacitance. Both n- and pMOSFETs are assumed to have the same C_{ov} per edge.

in Table 5.2, $R_{sd} = 200\ \Omega$-μm for n and 500 Ω-μm for p, both devices are degraded by about 10% in terms of their current drive capability. A simple rule of thumb for estimating the performance loss due to series resistance is to add R_{sd} to the intrinsic switching resistance, i.e., $\Delta R_{swn} \approx R_{sdn}$ and $\Delta R_{swp} \approx R_{sdp}$.

5.3.4.2 Sensitivity of Delay to Overlap Capacitance

Gate-to-drain overlap capacitance is a serious performance detractor in lightly loaded CMOS circuits. It not only enters the input capacitance but is also a component of the output capacitance, sometimes further amplified by feedback effects. Figure 5.39 shows both the input and the output capacitances as a function of the overlap capacitance C_{ov} (per edge). The value assumed in Table 5.2 is 0.3 fF/μm, about the lowest C_{ov} that can be achieved in practice, as discussed in Section 5.2.2. Both the gate-to-source and the gate-to-drain capacitances contribute to the input capacitance. Only the gate-to-drain component enters the output capacitance. However, its contribution is nearly doubled from its original value due to the *Miller effect* explained below. It is estimated that overall an overlap capacitance of 0.3 fF/μm accounts for about 35–40% of the intrinsic delay.

5.3.4.3 Miller Effect

The Miller effect arises when the voltages on both sides of a capacitor being charged or discharged vary with time. Figure 5.40 shows an example of three capacitors connected to a node of voltage V being charged. Each capacitor is connected to a different voltage level on the other side. One can express the charging current i as

$$i = C_1\frac{d(V-V_1)}{dt} + C_2\frac{d(V-V_2)}{dt} + C_3\frac{d(V-V_3)}{dt}. \tag{5.47}$$

Since both $V_1 = V_{dd}$ and $V_3 = 0$ are fixed voltages, one obtains

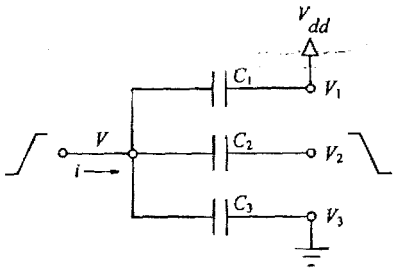


Figure 5.40. A circuit example illustrating the Miller effect.

$$i = C_1\frac{dV}{dt} + C_2\frac{dV}{dt} - C_2\frac{dV_2}{dt} + C_3\frac{dV}{dt}, \tag{5.48}$$

and there is no Miller effect on C_1 and C_3. However, $dV_2/dt \neq 0$, since V_2 varies with time. If V_2 varies with time in a direction opposite to that of V, it will take more time (and charge) to charge up the node voltage V to a certain level than it otherwise would. This happens, for example, between the input gate and the output drain of a CMOS inverter, as can be seen from the waveforms in Fig. 5.30. In particular, if $dV_2/dt = -dV/dt$, Eq. (5.48) becomes

$$i = (C_1 + 2C_2 + C_3)\frac{dV}{dt}. \tag{5.49}$$

In other words, the capacitor C_2 appears to have doubled its capacitance as far as the charging of node voltage V is concerned. From another angle, it takes a net flow of charge of $\Delta Q_2 = 2C_2V_{dd}$ into the capacitor C_2 to switch it from an initial state of $V - V_2 = -V_{dd}$ to a final state of $V - V_2 = V_{dd}$.

Another manifestation of the Miller effect is *feedforward*. For example, when the gate voltage rises in a CMOS inverter, the drain voltage, initially at V_{dd}, will momentarily rise to a value slightly higher than V_{dd} due to the capacitive coupling to the gate. This happens before the nMOSFET current starts to flow and brings the drain voltage down, as can be seen from the initial overshoot of V_2 and V_4 above V_{dd} in Fig. 5.30 and from the I–V trajectory in Fig. 5.31(a). It will take some additional amount of charge to pull the drain node to ground.

5.3.4.4 Sensitivity of Delay to Junction Capacitance

A major part of the output capacitance comes from the junction or drain-to-substrate capacitance. Figure 5.41 shows the input and output capacitances versus junction capacitance by varying the layout. In the folded layout, the junction capacitance is effectively halved as discussed in Section 5.1.3.2. This has a dramatic effect on C_{out}, but not on C_{in}. From Fig. 5.41, it is estimated that the junction capacitance accounts for more than 50% of the output capacitance in the straight-gate layout and that the folded layout improves the intrinsic inverter delay (FO = 1) by about 15%.

Table 5.4 Components of C_{in} and C_{out}

Component	Input Capacitance (%)	Output Capacitance (%)
Intrinsic gate oxide capacitance	49	18
Overlap capacitance	51	26
Junction capacitance (nonfolded)	–	56

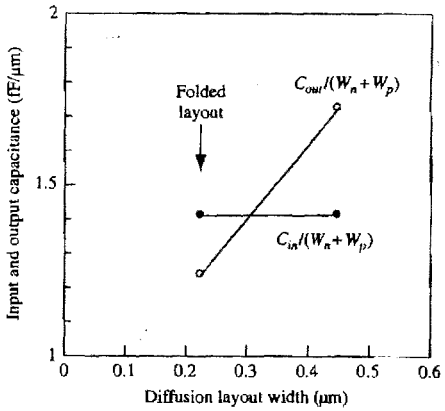


Figure 5.41. Input and output capacitances versus diffusion width d in Fig. 5.13. In the straight-gate layout (default case thus far), $d = a + b + c = 0.45$ μm. In the folded-gate layout (Fig. 5.14), d is effectively cut to half.

It is instructive to break C_{in} and C_{out} for the 0.1-μm CMOS devices listed in Table 5.2 into various components: intrinsic gate capacitance, overlap capacitance, and junction capacitance. This can be done by extrapolating the simulation results in Figs. 5.39, 5.41, and the capacitance components in Fig. 5.34. The results are given in Table 5.4. Note that the values of C_{in} and C_{out} are given in Table 5.3. The unloaded delay is proportional to $C_{in} + C_{out}$, in which only about a third comes from the intrinsic gate oxide capacitance.

5.3.5 Delay of Two-Way NAND and Body Effect

So far we have been using CMOS inverters, i.e., with fan-in of 1, for studying the performance factors. Many of the basic characteristics also apply to more general CMOS circuits. There are, however, a few other factors associated with the multiple fan-in NAND gates in which two or more nMOSFETs are stacked between the output node and the power-supply ground. This subsection examines these factors using a two-way NAND (Fig. 5.10) as an example.

5.3.5.1 Top and Bottom Switching of a Two-Way NAND Gate

The simulation is set up with the layout shown in Fig. 5.15 and with the same 0.1-μm CMOS devices listed in Table 5.2, except the p- to n-device width ratio. Because the

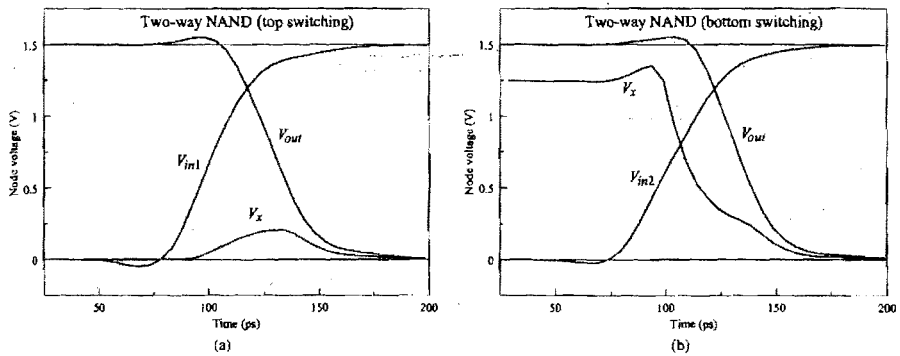


Figure 5.42. Waveforms of V_{in1} (top gate), V_{in2} (bottom gate), V_{out} (drain of the top nMOSFET and both pMOSFETs), and V_x (node between the two stacked nMOSFETs) for (a) top switching and (b) bottom switching in the pull-down event of a 2-way NAND gate. The device parameters are those listed in Table 5.2, except that $W_n/W_p = 1.0/1.5$ μm.

pull-down current in a NAND gate is somewhat lower than that in an inverter due to stacking of nMOSFETs, both the transfer curves and the up and down delays are better matched with a W_p/W_n ratio of 1.5 instead of 2. In this configuration, the two parallel pMOSFETs are naturally folded. The nMOSFETs are nonfolded. The width of the diffusion region (V_x-node) between the two stacked nMOSFETs is assumed to be the minimum lithography dimension, 0.15 μm in this case. To construct a linear chain of two-way NAND gates, one must distinguish between the two cases: *top switching* and *bottom switching*, as was first outlined in Section 5.1.2. Referring to Fig. 5.10, top switching means that transistors N1 and P1 are driven by a logic transition propagated through input 1. Input 2 is tied to V_{dd} in this case, so that N2 is always on and P2 is always off. On the other hand, in bottom switching transistors N2 and P2 are driven by a logic signal from the output of the previous stage through input 2, while input 1 is tied to V_{dd}. These two switching modes have somewhat different delay characteristics as discussed below.

It is instructive to examine the switching waveforms of various node voltages in a two way NAND. Figure 5.42 plots the input, output, and V_x-node voltages versus time during an nMOSFET pull-down event. In the top-switching case in Fig. 5.42(a), the V_x-node voltage starts at zero, rises momentarily to a peak about 15% of V_{dd}, then falls back to zero together with V_{out}. The rise of V_x is a result of the discharging current when the top transistor is turned on. In the bottom-switching case in Fig. 5.42(b), the V_x-node voltage starts at a high value, but quite a bit lower than V_{dd}. Even though its gate is tied to V_{dd}, the top transistor is initially biased in the subthreshold region since $V_{dd} - V_x < V_t + (m-1)V_x$ (the factor m comes from the body effect; see the discussion in Section 5.1.2). The exact starting value of V_x depends on a detailed matching of the subthreshold currents in the top and bottom nMOSFETs. When the bottom nMOSFET is turned on, the V_x-node is pulled down to ground, followed by V_{out}. One can easily figure out the bias point of each transistor, e.g., in the linear or saturation region, from the values of V_{in}, V_x, and $V_{out} - V_x$ at any given instant.

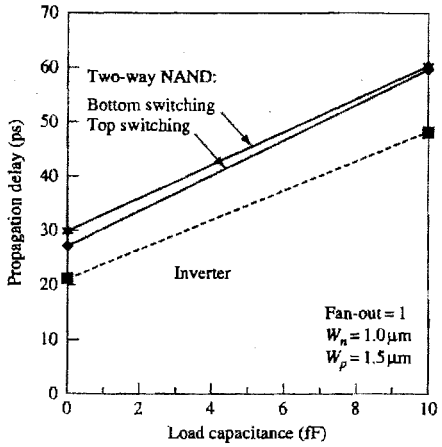


Figure 5.43. Propagation delay versus load capacitance. The two solid lines are for the top switching and the bottom switching cases of a 2-way NAND gate. The dashed line shows the delay of a CMOS inverter of the same device widths for comparison.

Figure 5.43 plots the propagation delay of a two-way NAND gate (solid lines), as described above, versus the load capacitance C_L. The dashed line shows the delay of an inverter of the same widths for comparison. A delay equation of the same form as Eq. (5.39) also applies to the two-way NAND, but with different values of R_{sw}, C_{in}, and C_{out}. The intrinsic delay ($C_L = 0$) of the two-way NAND is about 34% (1.34 ×) longer than that of the inverter, for the following reasons. First, let us consider the capacitances. The input capacitance of a two-way NAND stage is essentially the same as that of an inverter. However, a two-way NAND has a higher output capacitance. In the top-switching case, there is an additional gate-to-drain overlap capacitance C_{ov} (no Miller effect) on the pMOSFET side of the two-way NAND layout in Fig. 5.15, compared with the inverter layout in Fig. 5.14(a). In the bottom-switching case, the output capacitance is further increased by additional components on the nMOSFET side. These include the gate capacitance of N1, some overlap capacitance associated with the gate of N1, and a small junction capacitance of the V_x-node. In addition to the higher capacitances, the switching resistances, i.e., the slopes of the lines in Fig. 5.43, of the two-way NAND are also higher than that of the inverter. This primarily stems from the stacking of the two nMOSFETs between the output node and the ground such that when one nMOSFET is switching, the other acts like a series resistance, which degrades the current. In terms of switching resistances, top switching is worse than bottom switching, since the series resistance in the former case is placed between the source and the ground, which results in additional loss of gate drive. This is evident in Fig. 5.43. For the intrinsic delays in this example, bottom switching is worse than top switching because the extra capacitance outweighs the slight difference in the switching resistance. Under heavy loading conditions, however, top switching is the worst case, in which the switching resistance is about 21% (1.21×) worse than that of the inverter in Fig. 5.43.

The degradation of switching resistance in NAND circuits with fan-in > 1 can be roughly estimated using the following simple model. In the pull-down operation of a two-way NAND, the nonswitching nMOSFET has its gate voltage fixed at V_{dd} and acts like a series resistor to the switching transistor. Since it operates mainly in the linear region during a switching event (see the discussion in Section 5.1.2.), its effective resistance can be approximated by V_{dsat}/I_{onN}, where V_{dsat} and I_{onN} are the saturation voltage and current at $V_{gs} = V_{dd}$. This increases the nMOSFET switching resistance by roughly the same amount, i.e., $\Delta R_{swn} = V_{dsat}/I_{onN}$, based on the discussion following Fig. 5.38. Using $R_{sw} = (R_{swn} + \Delta R_{swn} + R_{swp})/2$ with R_{swn} and R_{swp} for inverters given by Eqs. (5.41) and (5.42), one can write the switching resistance of a two-way NAND gate as

$$R_{sw}(\mathrm{FI} = 2) = \frac{V_{dd}}{4\langle I_N\rangle} + \frac{V_{dsat}}{2I_{onN}} + \frac{V_{dd}}{4\langle I_p\rangle}. \tag{5.50}$$

For the above 0.1-μm CMOS example, $\langle I_N\rangle \approx \langle I_P\rangle \approx (3/5)I_{onN}$, and $V_{dsat} \approx (1/3)V_{dd}$ from Fig. 5.31(a). Substitution of these numbers in Eq. (5.50) yields a switching resistance about 1.2 times that of the inverter, consistent with the numerical results stated before. Equation (5.50) can be generalized to a higher number of fan-ins by inserting a multiplying factor of (FI − 1) in front of the V_{dsat} term. Since R_{sw} degrades rapidly with the number of fan-ins, fan-ins higher than 3 are rarely used in CMOS circuits.

5.3.5.2 Delay Sensitivity to Body Effect

The delays in Fig. 5.43 are computed based on the set of device parameters in Table 5.2, in which the body-effect coefficient is m = 1.3. With all other device parameters being equal, the delay increases with m for two reasons. ***First, the device saturation current decreases with increasing m due to body effect at the drain.*** This can be seen from the saturation current expression, Eq. (3.79), for the $n = 1$ case discussed in Subsection 3.2.2.2. Other values of n lead to qualitatively similar results. The dependence of the saturation current on m is stronger for less velocity saturated devices, e.g., pMOSFETs. The fully velocity saturated current, Eq. (3.81), is independent of m. ***The second factor is more applicable to stacked nMOSFETs in NAND gates: when the source potential is higher than the body potential, as in transistor N1 of Fig. 5.10, the threshold voltage increases because of body effect and the current decreases.*** The source-to-body potential of N1 is given by the V_x voltage shown in Fig. 5.42(a) and (b). To a lesser degree, this effect also occurs in a CMOS inverter due to the presence of series resistances at the n- and p-source terminals.

5.4 Performance Factors of Advanced CMOS Devices

In the last section, we discussed the sensitivity of CMOS delay in digital circuits to various device parameters in a standard bulk CMOS technology. In this section, we start with the performance factors of MOSFETs in RF circuits. Then we examine the effect of transport parameters, e.g., mobilities and saturation velocities, on CMOS performance. Mobility enhancements are possible in many advanced CMOS devices and structures,

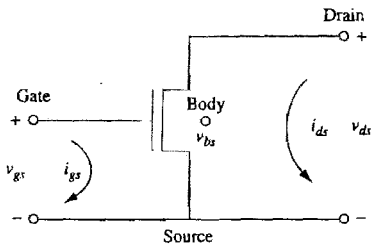


Figure 5.44. MOSFET used as a small-signal amplifier in the common-source configuration.

including SiGe MOSFETs. The performance factors of low-temperature CMOS are addressed at the end.

5.4.1 MOSFETS in RF Circuits

While CMOS devices are predominantly used in digital circuits, they can also perform as small-signal amplifiers in RF circuits. For this purpose, nMOSFETs are used almost exclusively because of their superior performance over pMOSFETs. In this section, a small-signal equivalent circuit is introduced for MOSFET transistors in the frequency domain. The two-port matrix representation is then described, leading to the intrinsic performance factors as an RF amplifier.

5.4.1.1 Small-Signal Equivalent Circuit

Figure 5.44 shows the small-signal, two-port schematic of a MOSFET used in the common-source configuration. The dc bias circuits are not shown. The ac input goes into the gate-source port and the ac output is extracted at the drain–source port. The convention here is that the upper case symbols denote full quantities and lower case symbols small signal quantities, e.g., $v_{gs} = \delta V_{gs}$, $i_{ds} = \delta I_{ds}$, etc. All i_{gs}, i_{ds}, v_{gs}, v_{ds} are complex numbers (phasors) with the common time dependence $e^{j\omega t}$, where ω is the angular frequency of the small signal. In this representation, if the full time-dependent expression for the voltage across the input port is $|v_{gs}|\cos(\omega t + \alpha)$, i.e., the real part of $|v_{gs}|e^{j(\omega t+\alpha)}$ or $\mathrm{Re}\left(|v_{gs}|e^{j(\omega t+\alpha)}\right)$, then $v_{gs} = |v_{gs}|e^{j(\omega t+\alpha)}$. Likewise, $v_{ds} = |v_{ds}|e^{j(\omega t+\beta)}$, etc.

In general, the drain current is a function of V_{gs}, V_{ds}, and V_{bs}, i.e., $I_{ds}(V_{gs}, V_{ds}, V_{bs})$. Its small-signal increment has three components:

$$i_{ds} = \frac{\partial I_{ds}}{\partial V_{gs}} v_{gs} + \frac{\partial I_{ds}}{\partial V_{ds}} v_{ds} + \frac{\partial I_{ds}}{\partial V_{bs}} v_{bs} = g_m v_{gs} + g_{ds} v_{ds} + g_{mb} v_{bs}, \tag{5.51}$$

where $g_m \equiv \left(\partial I_{ds}/\partial V_{gs}\right)\big|_{V_{ds}, V_{bs}}$ is the intrinsic transconductance, $g_{ds} \equiv \left(\partial I_{ds}/\partial V_{ds}\right)\big|_{V_{gs}, V_{bs}}$ is the intrinsic output conductance, and $g_{mb} \equiv \left(\partial I_{ds}/\partial V_{bs}\right)\big|_{V_{gs}, V_{ds}}$ is the body (back gate) transconductance.

In addition to the conductive components above, there are also capacitive components from the gate to source, gate to drain, gate to body, body to source, and body to drain. They give rise to displacement currents, e.g., $i_{gs} = C_{gs} dv_{gs}/dt$, that are 90° out of phase

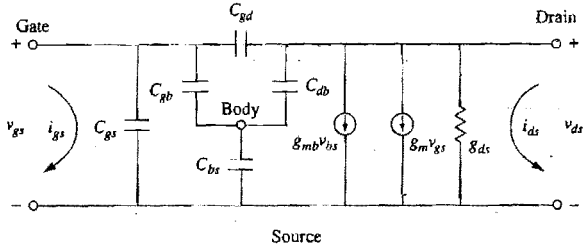


Figure 5.45. Small-signal equivalent circuit of an intrinsic MOSFET in the frequency domain.

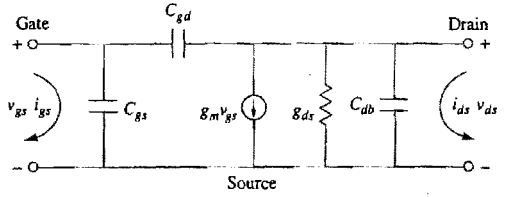


Figure 5.46. Simplified small-signal equivalent circuit of an intrinsic MOSFET.

with the sinusoidal voltage. For i_{ds}, capacitive components like $C_{gd}d(v_{ds} - v_{gs})/dt$ and $C_{db}d(v_{ds} - v_{bs})/dt$ need to be added to the conductive components in Eq. (5.51). The complete small-signal RF equivalent circuit of an intrinsic MOSFET is shown in Fig. 5.45, where Eq. (5.51) is represented by two voltage-dependent current sources and a resistor. Note that the effect of body bias on the threshold voltage is represented by the current generator $g_{mb}v_{bs}$. The conductive components across the source-to-body and the drain-to-body junctions are omitted, assuming they are either unbiased or reverse biased.

If it is further assumed that the body is either tied to the source or biased at a constant dc voltage with respect to the source, the above general circuit can be simplified to the one in Fig. 5.46. Note that C_{gb}, which is small once the MOSFET is turned on, is lumped into C_{gs}.

5.4.1.2 Unity-Current-Gain Frequency of an Intrinsic MOSFET

In the small-signal analysis of a two-port network in the frequency domain, an admittance matrix is often used to describe the linear relationship between the terminal currents and voltages. These relations stem from Kirchhoff's current and voltage laws applied to the equivalent circuit. Based on Fig. 5.46, the intrinsic admittance matrix of a MOSFET can be written as:

$$\begin{bmatrix} i_{gs} \\ i_{ds} \end{bmatrix} = \begin{bmatrix} j\omega(C_{gs} + C_{gd}) & -j\omega C_{gd} \\ g_m - j\omega C_{gd} & g_{ds} + j\omega C_{db} + j\omega C_{gd} \end{bmatrix} \begin{bmatrix} v_{gs} \\ v_{ds} \end{bmatrix}. \tag{5.52}$$

The intrinsic current gain β is obtained from Eq. (5.52), with the assumption that the output port is short-circuited, i.e., $v_{ds} = 0$:

$$\beta = \left.\frac{i_{ds}}{i_{gs}}\right|_{v_{ds}=0} = \frac{g_m - j\omega C_{gd}}{j\omega(C_{gs} + C_{gd})}. \tag{5.53}$$

The magnitude of the current gain is

$$|\beta| = \frac{\sqrt{{g_m}^2 + (\omega C_{gd})^2}}{\omega(C_{gs} + C_{gd})}. \tag{5.54}$$

The intrinsic unity-current-gain frequency is then obtained by setting $|\beta| = 1$ in the above equation and solving for $f_T = \omega/2\pi$:

$$f_T = \frac{g_m}{2\pi\sqrt{{C_{gs}}^2 + 2C_{gs}C_{gd}}}. \tag{5.55}$$

In expressions commonly found in the literature, $({C_{gs}}^2 + 2C_{gs}C_{gd})^{1/2}$ in the denominator is approximated to $C_{gs} + C_{gd}$ under the assumption that $C_{gs} >> C_{gd}$ in saturation. For the 0.1-μm nMOSFET example in Table 5.2, $g_m \approx 600$ mS/mm and $f_T \approx 80$ GHz.

The general condition for power gain in a two-port network is discussed in Appendix 13. The unity-power-gain frequency or the maximum oscillation frequency, f_{max}, can be solved from the condition, Eq. (A13.8). For an intrinsic MOSFET, the power gain condition is always met since $\mathrm{Re}(Y_{11}) = 0$ for the matrix in Eq. (5.52). It is mathematically tedious to deal with an extrinsic MOSFET with parasitic resistances. The unity-current-gain and unity-power-gain frequencies, f_T and f_{max}, of an extrinsic MOSFET are solved in Appendix 14. A simplified expression for f_{max} often found in the literature is Eq. (A14.15),

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_g C_{gd}}}, \tag{5.56}$$

where R_g is the parasitic gate resistance.

The f_T and f_{max} figures of a modern nMOSFET with sub-50-nm channel lengths can be in the range of 200 GHz, rivaling those of modern bipolar transistors. However, as an RF amplifier, the voltage gain of a MOSFET is inferior to that of a bipolar transistor due to the transconductance and output characteristics. Using Eq. (5.52) with an open circuit at the output, i.e., $i_{ds} = 0$, one can find the voltage gain at low frequencies as

$$\left|\frac{v_{ds}}{v_{gs}}\right|(\omega \to 0) = \frac{g_m}{g_{ds}}. \tag{5.57}$$

This has the same form as the maximum slope of an inverter transfer curve, $|dV_{out}/dV_{in}|$, discussed in Section 5.1.1.2. High g_m and f_T figures are obtained with short-channel devices which also have high g_{ds} due to drain-induced-barrier-lowering effects. In the 0.1-μm nMOSFET example in Fig. 5.31(a), $g_m/g_{ds} \approx 17$, significantly lower than the typical voltage gain of bipolar transistors discussed in Section 8.5.1.

5.4.2 Effect of Transport Parameters on CMOS Performance

When CMOS devices were scaled to 0.1-μm channel length around the turn of the millennium, technologists began to develop strained silicon MOSFETs that have mobilities higher than those "universal" values discussed in Section 3.1.5. The strain is either process induced, such as by depositing stressful nitride films on a silicon substrate, or produced by epitaxial alloy growth, such as SiGe (Kesan *et al.*, 1991), with a lattice constant mismatched to that of silicon. Theoretically, the hole mobility increases in silicon under either tensile or compressive strain due to breaking of the valence band degeneracy and reduction of the conductivity mass (Fischetti and Laux, 1996). The electron mobility is also enhanced in silicon under tensile strain because of increased electron populations in the two lower energy valleys with a lower conductivity mass. Reduction of the effective mass may also benefit the source injection velocity in the ballistic model [Eq. (3.99)]. Tensile strain can be created by growing an epitaxial silicon layer on a relaxed SiGe film whose lattice constant is slightly larger than that of bulk silicon. Alloy scattering in SiGe, however, has a negative effect on both the electron and the hole mobilities (Fischetti and Laux, 1996).

The benefit of increased mobilities on CMOS delay can be investigated using the same circuit model as before. The base device case is that of a 0.1-μm CMOS with the parameters listed in Table 5.2. The performance gain due to higher mobilities comes in through the switching resistance factor and is therefore independent of fan-out and wire loading conditions. Similar improvement factors are also found in the delay of 2-way NAND circuits. For given carrier densities and fields, MOSFET current is determined by three transport related parameters: mobility, saturation velocity, and series resistance. If both the mobility and the saturation velocity improve by a factor $\kappa > 1$, and if the series resistance decreases by $1/\kappa$, the current improves by κ, or equivalently, the switching resistance decreases by $1/\kappa$. Empirically, one may write

$$R_{sw} \propto {\mu_{eff}}^{-a} {v_{sat}}^{-b} {R_{sd}}^{c}, \tag{5.58}$$

where $a + b + c = 1$, for small changes of μ_{eff}, v_{sat}, and R_{sd} with respect to their Table 5.2 values. Here, changing each parameter means changing both the n- and p-device corresponding parameters by the same factor. Figure 5.47(a) shows the simulated variation of R_{sw} with the above parameters in a log-log scale. It is observed that $a \approx 0.61$, $b \approx 0.28$, and $c \approx 0.11$ for the 0.1-μm CMOS example considered here. ***Relatively speaking, mobility is the most important parameter for CMOS performance. Even at 0.1 μm length, MOSFETs are not as velocity saturated as one might expect.*** This is mainly because of the universal mobility behavior, i.e., mobility degradation with vertical fields (Section 3.1.5). As MOSFETs are scaled down, the voltage cannot be scaled as much as the device dimension because of subthreshold non-scaling. Lateral fields in the source–drain direction go up as a result. This leads to higher vertical fields as well which are necessary to keep the 2-D short-channel effects in check. The net result is the decrease of mobility as device lengths become shorter, as discussed in Section 4.1.3.2. Therefore, MOSFETs do not necessarily become more velocity saturated as they are scaled down.

Figure 5.47(b) further breaks out the sensitivity of switching resistance to electron and hole mobilities separately. Not surprisingly, higher hole mobility is more advantageous

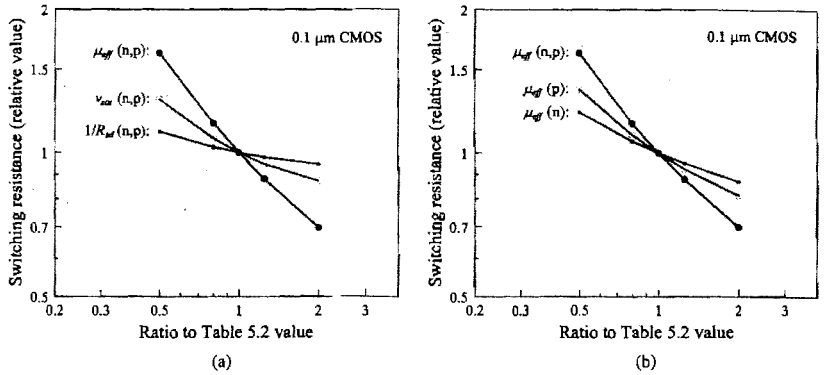


Figure 5.47. (a) Sensitivity of switching resistance to transport parameters: μ_{eff}, v_{sat}, and $1/R_{sd}$. Each curve depicts relative change of R_{sw} with respect to relative change of the specific transport parameter for both n- and pMOSFETs while others are kept constant. (b) Breakdown of the mobility dependence into the electron and hole factors separately.

than higher electron mobility because pMOSFETs are not as velocity saturated as nMOSFETs. Quantitatively, the mobility exponent a in Eq. (5.58) can be decomposed into $a = a_n + a_p$ such that

$$R_{sw} \propto \mu_n^{-a_n} \mu_p^{-a_p} v_{sat}^{-b} R_{sd}^{c}, \tag{5.59}$$

where $a_n = 0.24$ and $a_p = 0.37$ in this case.

5.4.3. Low-Temperature CMOS

The performance advantage of low-temperature operation of MOSFETs has been recognized for some time (Gaensslen *et al.*, 1977; Sun *et al.*, 1987). ***The benefit is mainly derived from two aspects of the MOSFET characteristics at low temperature: higher carrier mobilities and steeper subthreshold slope.*** Field-dependent electron and hole mobilities at 300 and 77 K are shown in Figs 3.15 and 3.16. In this temperature range, the electron mobility improves by a factor of 2–5, depending on the magnitude of the vertical field. This is because of the much reduced electron–phonon scattering at low temperatures. Similarly, hole mobility also improves from 300 to 77 K, although by a more moderate factor of 1.7–4. The improvement factors of both electron and hole mobilities decrease at higher vertical fields where surface roughness scattering, which is largely insensitive to temperature, becomes important. In addition to the mobilities, the saturation velocities of carriers in bulk silicon also improve slightly at low temperatures. There are no extensive experimental data on the saturation velocities in a MOSFET channel as a function of temperature and field. In general, it is expected that v_{sat} improves by some 10–30% from 300 to 77 K (Taur *et al.*, 1993a).

Another important aspect of the MOSFET characteristics at low temperatures is that the subthreshold current slope steepens by a factor proportional to the absolute temperature

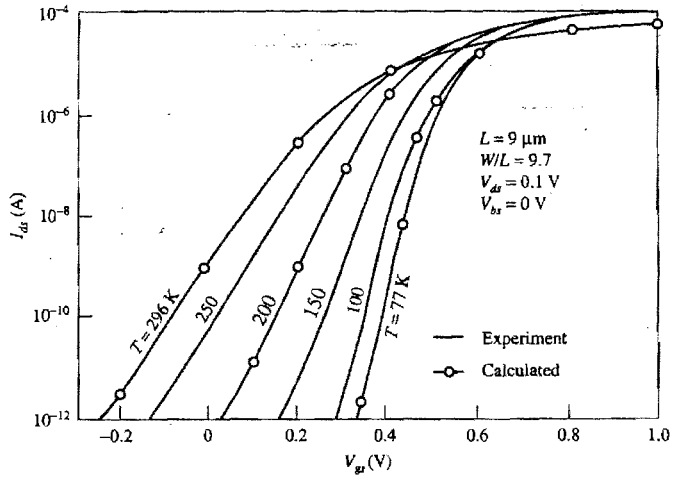


Figure 5.48. Subthreshold I–V characteristics of nMOSFETs as a function of temperature. The gate oxide is 200 Å thick. (After Gaensslen *et al.*, 1977).

(Section 3.1.3), making it much easier to turn off a MOSFET than at room temperature. An example is shown in Fig. 5.48 (Gaensslen *et al.*, 1977). This allows the threshold voltage V_t, and therefore the power-supply voltage V_{dd}, to scale down further below their permissible values at room temperature. For example, a subthreshold slope of 25 mV/decade at 80 K (Taur *et al.*, 1993b) would allow a V_t of 0.1–0.2 V and a V_{dd} of 0.4–0.8 V, provided that the threshold voltage tolerances from short-channel effects can be tightened as well through the use of optimized channel doping profiles (Taur *et al.*, 1997).

To estimate the performance gain of CMOS circuits at low temperatures, we consider the example of 0.1-μm CMOS and evaluate the intrinsic inverter delay as a function of temperature. At each temperature, the electron and the hole mobilities are adjusted according to the published data, e.g., in Figs 3.15 and 3.16. A slight temperature dependence of the saturation velocities is also included in the model. Threshold voltages are adjusted following various strategies described below. In Fig. 5.49, the relative performance factor, defined as inversely proportional to the inverter delay, is plotted versus temperature. Since the capacitances to the first order are independent of temperature, the performance factor mainly reflects the reciprocal of the switching resistance in Eq. (5.39) and should be applicable to various static CMOS circuits with different fan-out and loading conditions.

Three different scenarios are considered in Fig. 5.49, depending on the assumption about the threshold voltage. In each case, the performance factor is normalized to the value at 100 °C, which is the temperature specified for most of the IC products. In the *same-hardware* case, the magnitude of threshold voltage increases toward lower temperatures, governed by the ≈ −0.8-mV/°C coefficient discussed in Section 3.1.4. The

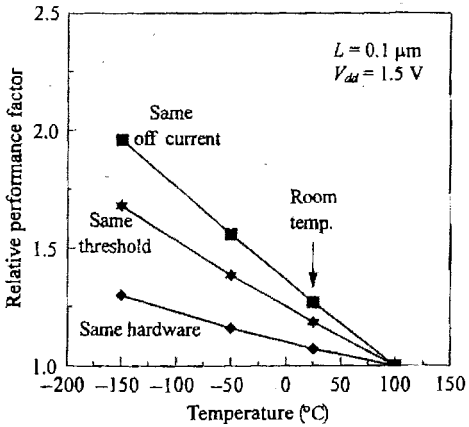


Figure 5.49. Relative performance factor of 0.1-μm CMOS as a function of temperature. Threshold voltages are adjusted differently with temperature in each of the three scenarios as described in the text. All the performance factors are normalized to the value at 100 °C, where $V_{on} = \pm 0.33$ V.

linearly extrapolated threshold voltage at 100 °C is $V_{on} = \pm 0.33$ V, based on the $V_{on} = \pm 0.4$-V figure at room temperature. The threshold behavior versus temperature is also evident in Fig. 5.48. The rise of threshold voltage offsets some of the performance gained from the higher mobilities such that a lesser net improvement is obtained at low temperatures, as shown by the bottom curve in Fig. 5.49. In the *same-threshold* case, the threshold voltages are held constant at $V_{on} = \pm 0.33$V as the temperature is varied. The middle curve in Fig. 5.49 therefore represents the performance gained from the higher mobilities and, to a lesser extent, from the slightly higher saturation velocities. ***To gain the most performance out of low-temperature CMOS, one should turn the threshold-voltage trend around and take advantage of the steeper subthreshold slope***. This is represented by the *same-off-current* case in Fig. 5.49, in which the threshold voltages are adjusted to lower values as temperature decreases such that the off current is maintained at the same level as the product specification at 100 °C (e.g., 50-nA/μm worst case). In principle, this can be accomplished using the retrograde channel doping concept outlined in Section 4.2.3 without degrading the short-channel effect. Up to a factor of two in performance gain can be achieved at −150 °C, as indicated by the top curve in Fig. 5.49. The threshold voltages at that temperature are adjusted to $V_{on} = \pm 0.18$ V, which leaves plenty of gate overdrive for $V_{dd} = 1.5$ V. It may be desirable at this point to trade performance for lower power by operating the CMOS devices at a lower supply voltage, e.g., at 0.8 V. In fact, with the steep field dependence of the low-temperature mobilities (Figs 3.15 and 3.16), a lower voltage allows the devices to operate in a regime of significantly higher mobilities. Following the design principles outlined in Section 5.3.3, one should be able to achieve a 4× power reduction with only a slight loss in performance. This is particularly worthwhile because it is rather expensive to cool a high-power chip to low temperatures.

In addition to the device improvement depicted in Fig. 5.49, which amounts to about 0.3%/°C in the best case, the conductivity of metal interconnects (either aluminum or copper) also improves at low temperatures. Depending on the material purity, the improvement factor lies in the range of 0.3–0.6%/°C. In other words, interconnect *RC* delays will improve by at least as much as the devices. This means that the performance factors projected in Fig. 5.49 at the device level should translate directly to the chip level without extensive design modifications. Apart from packaging issues and system costs, one key challenge for low-temperature CMOS is to be able to tighten the short-channel threshold tolerances through the use of optimized channel doping profiles while following the low-V_t strategy in Fig. 5.49 for best performance.

Exercises

5.1 Consider the CMOS switching delay, $\tau = (\tau_n + \tau_p)/2$, where τ_n and τ_p are given by Eqs. (5.4) and (5.5). If the inverter is driving another stage with the same n- to p-width ratio and if both the n- and p-devices have the same capacitance per unit width, the load capacitance C is proportional to $W_n + W_p$. Show that the minimum delay τ occurs for a width ratio of $W_p/W_n = (I_{on,n}/I_{on,p})^{1/2}$, which is different from Eq. (5.1) for best noise margin where $\tau_n = \tau_p$.

5.2 For an *RC* circuit with a capacitor C connected in series with a resistor R and a switchable voltage source, solve for the waveform of the voltage across the capacitor, $V(t)$, when the voltage source is abruptly switched from 0 to V_{dd} with the initial condition $V(t = 0) = 0$. Show that when the equilibrium condition is established, an energy of $CV_{dd}^2/2$ has been dissipated in the resistor R and the same amount of energy is stored in C. Since the energy dissipated and the energy stored are independent of R, the same results hold even if $R = 0$. What happens if the voltage source is now switched off from V_{dd} to 0 with the initial condition $V(t = 0) = V_{dd}$?

5.3 The carrier transit time is defined as $\tau_{tr} \equiv Q/I$, where Q is the total inversion charge and I is the total conduction current of the device. For a MOSFET device biased in the linear region (low drain voltage), use Eq. (3.23) and the inversion charge expression above Eq. (3.58) to derive an expression for τ_{tr}. Similarly, use Eq. (3.28) and the expression above Eq. (3.60) to derive τ_{tr} for a long-channel MOSFET biased in saturation.

5.4 Use Eq. (3.79) and the inversion-charge expression in Exercise 3.10 to find the carrier transit time τ_{tr} for a short-channel MOSFET biased in saturation. What is the limiting value of τ_{tr} when the device becomes fully velocity-saturated as $L \to 0$?

5.5 A similar distributed network to the one in Fig. 5.21 can be used to formulate the *transmission-line model* of contact resistance in a planar geometry (Berger, 1972). Here we consider the current flow from a thin resistive film (diffusion with a sheet resistivity ρ_{sd}) into a ground plane (metal) with an interfacial contact resistivity ρ_c between them (Fig. 5.16). Thus, in Fig. 5.21, $R\,dx$ corresponds to $(\rho_{sd}/W)dx$, and $C\,dx$ is replaced by a shunt conductance $G\,dx$, which corresponds to $(W/\rho_c)dx$. Show

that both the current and voltage along the current flow direction satisfy the following differential equation:

$$\frac{d^2f}{dx^2} = RGf = \frac{\rho_{sd}}{\rho_c} f,$$

where $f(x) = V(x)$ or $I(x)$ defined in Fig. 5.21.

5.6 Following the above transmission-line model, with the boundary condition $I(x = l_c) = 0$ where $x = 0$ is the leading edge and $x = l_c$ is the far end of the contact window (Fig. 5.16), solve for $V(x)$ and $I(x)$ within a multiplying factor and show that the total contact resistance, $R_{co} = V(x = 0)/I(x = 0)$, is given by Eq.(5.11).

5.7 The insertion of a buffer stage (Section 5.3.2) between the inverter and the load is beneficial only if the load capacitance is higher than a certain value. Find, in terms of C_{in} and C_{out}, the minimum load capacitance C_L above which the single-stage buffered delay given by Eq. (5.45) is shorter than the unbuffered delay given by Eq. (5.43).

5.8 Generalize Eq. (5.44) for one-stage buffered delay to n stages: if the width ratios of the successive buffer stages are $k_1, k_2, k_3, \ldots, k_n$ (all >1), show that the n-stage buffered delay is

$$\tau_b(n) = R_{sw}\left[(n+1)C_{out} + (k_1 + k_2 + \cdots + k_n)C_{in} + \frac{C_L}{k_1 k_2 \cdots k_n}\right].$$

5.9 Following the previous exercise, show that for a given n, the n-stage buffered delay is a minimum,

$$\tau_{b\,\min}(n) = R_{sw}\left[(n+1)C_{out} + (n+1)C_{in}\left(\frac{C_L}{C_{in}}\right)^{1/(n+1)}\right],$$

when $k_1 = k_2 = \cdots = k_n = (C_L/C_{in})^{1/(n+1)}$. Here $\tau_{b\,\min}(n)$, as expected, is reduced to Eq. (5.45) if $n = 1$.

5.10 If one plots the minimum n-stage buffered delay from the previous exercise versus n, it will first decrease and then increase with n. In other words, depending on the ratios of C_L/C_{in} and C_{out}/C_{in}, there is an optimum number of buffer stages for which the overall delay is the shortest. Show that this optimum n is given by the closest integer to

$$n = \frac{\ln(C_L/C_{in})}{\ln k} - 1,$$

where k is a solution of

$$k(\ln k - 1) = \frac{C_{out}}{C_{in}}.$$

For typical C_{out}/C_{in} ratios not too different from unity, k is in the range of 3–5. Note that k also gives the optimum width ratio between the successive buffer stages, i.e., $k_1 = k_2 = \cdots = k_n = k$. Also show that the minimum buffered delay is given by

$$\tau_{b\,\min} \approx kR_{sw}C_{in}\ln(C_L/C_{in}),$$

which only increases logarithmically with load capacitance.

5.11 Consider a chain of CMOS inverters with power supply V_{dd}. The propagation delay between the waveforms can be expressed by Eq. (5.39) with FO = 1. What is the power dissipation while the signal is propagating down the chain? If the device widths are increased or decreased by a factor of k (>1 or <1) to kW_n, kW_p while C_L remains constant, how would the delay and power vary with k?

6 Bipolar Devices

Although most microelectronics products are now made of CMOS transistors, bipolar transistors remain important in microelectronics because of their superior characteristics for analog circuit applications. There are two types of bipolar devices: the n–p–n type which has a p-type base and n-type emitter and collector, and the p–n–p type which has an n-type base and p-type emitter and collector. Commonly used bipolar devices are either *lateral transistors*, where the active device regions are arranged horizontally adjacent to one another and the active currents flow laterally, or *vertical transistors*, where the active device regions are arranged vertically one on top of another and the active currents flow vertically. ***Practically all bipolar transistors used in modern VLSI applications are of the vertical n–p–n type.***

For simplicity, only vertical n–p–n bipolar transistors will be considered explicitly here. The equations derived for vertical transistors apply to horizontal transistors as well, provided that the device parameter values are adjusted accordingly. Also, the equations for an n–p–n transistor can be extended to a p–n–p transistor simply by reversing the voltage and dopant polarities and using the appropriate device parameter values.

6.1 n–p–n Transistors

Figure 6.1(a) shows a one-dimension representation of a vertical n–p–n transistor. The transistor consists of an n^+-type emitter region and an n-type collector region, with a p-type base region sandwiched in between. The collector sits on an n^+-type subcollector region. Figure 6.1(b) shows a cross-sectional schematic of the transistor. The n^+ subcollector is brought to the top surface for electrical contact by a vertical n^+-type *reach-through* region.

The starting substrate material for fabricating a vertical n–p–n transistor is usually a p-type silicon wafer. The subcollector is formed in the substrate, usually by ion implantation and diffusion. Then the n-type collector is formed on top of the subcollector by an epitaxial growth process. An n^+-type vertical reach-through region is formed for electrical connection to the subcollector. After that, the p-type base is formed in the epitaxial layer by ion implantation. Alternatively, the p-type base can be formed by growing a thin epitaxial silicon layer *in situ* doped with boron. This epitaxial base silicon layer may contain Ge and/or C if desired. Then the heavily doped n-type emitter is formed by ion implantation and diffusion, or by depositing a heavily doped n-type polysilicon

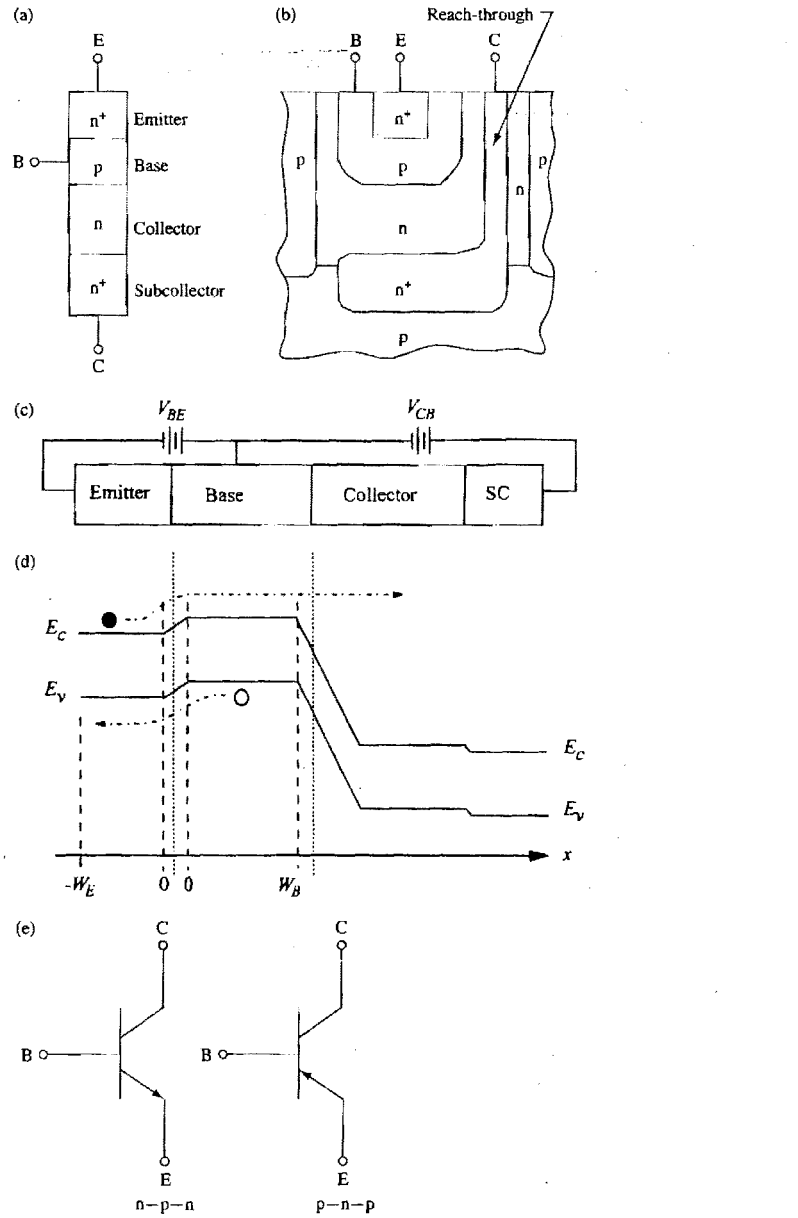


Figure 6.1. (a) One-dimensional representation of an n–p–n transistor, (b) its cross-sectional schematic, (c) schematic illustrating the applied voltages in normal operation, (d) schematics illustrating the energy-band diagram, carrier flows, and locations of the boundaries of the emitter and base quasineutral regions, and (e) circuit symbols for an n–p–n transistor and a p–n–p transistor

layer on top of the base region. Adjacent transistors are isolated from one another by p-type pockets, as illustrated in Fig. 6.1(b), or by oxide-filled trenches. The process for fabricating a typical advanced vertical n–p–n bipolar transistor having an implanted base region is outlined in Appendix 2.

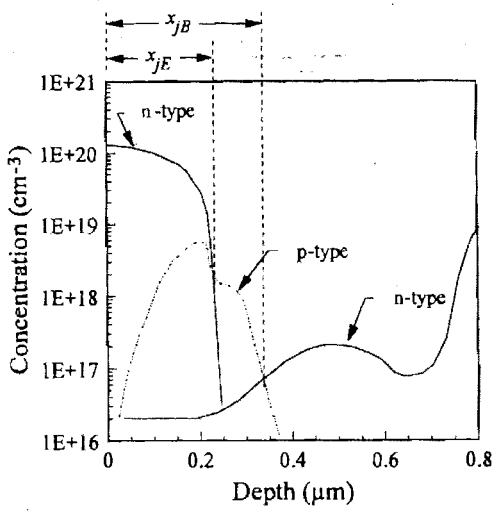
Figure 6.1(c) shows the bias condition for an n–p–n transistor in normal operation. The emitter–base diode is forward biased with a voltage V_{BE}, and the base–collector diode is reverse biased with a voltage V_{CB}. The corresponding energy-band diagram is shown schematically in Fig. 6.1(d). The forward-biased emitter–base diode causes electrons to flow from the emitter into the base and holes to flow from the base into the emitter. Those electrons not recombined in the base layer arrive at the collector and give rise to a *collector current*. The holes injected into the emitter recombine either inside the emitter or at the emitter contact. This flow of holes gives rise to a *base current*. (The operation of a bipolar transistor having both the emitter–base and collector–base diodes forward biased will be discussed in Section 9.1.3 in the context of bipolar inverter circuits and memory cells.)

Also illustrated in Fig. 6.1(d) are the coordinates which we will follow in describing the flow of electrons and holes. Thus, electrons flow in the x-direction, i.e., $J_n(x)$ is negative, and holes flow in the $-x$ direction, i.e., $J_p(x)$ is also negative. The physical junction of the emitter–base diode is assumed to be located at "$x=0$". However, to accommodate the finite thickness of the depletion layer of the emitter–base diode, the mathematical origin ($x=0$) for the quasineutral emitter region is shifted to the left of the physical junction, as illustrated in Fig. 6.1(d). Similarly, the mathematical origin ($x=0$) for the quasineutral base region is shifted to the right of the physical junction. The emitter contact is located at $x=-W_E$, and the quasineutral base region ends at $x=W_B$.

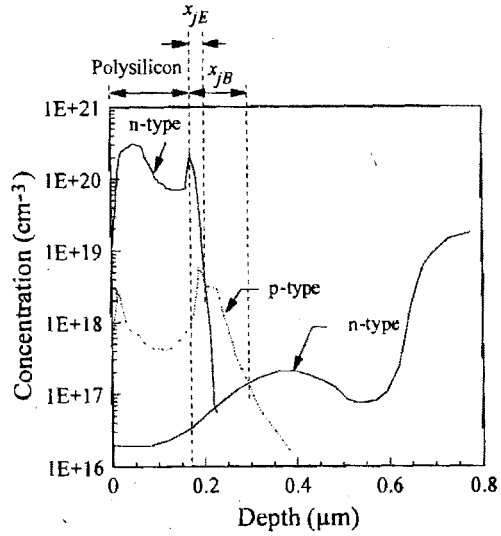
It should be noted that, due to the finite thickness of a junction depletion layer, the widths of the quasineutral p- and n-regions of a diode are always smaller than their corresponding physical widths. Unfortunately, in the literature as well as here, the same symbol is often used to denote both the physical width and the quasineutral width. For example, W_B is used to denote the base width. Sometimes W_B refers to the physical base width, and sometimes it refers to the quasineutral base width. The important point to remember is that ***all the carrier-transport equations for p–n diodes and for bipolar transistors refer to the quasineutral widths.***

In the literature, several different circuit symbols have been used for a bipolar transistor. In this book, we adopt the symbols illustrated in Fig. 6.1(e). The arrow indicates the direction of positive current flow in the emitter. For instance, in the n–p–n transistor, the emitter current is due primarily to electrons flowing from the emitter region towards the base region. Hence, the direction of positive current flow is from the base towards the emitter terminal. Similarly, in the p–n–p transistor, the emitter current is due primarily to holes flowing from the emitter region towards the base region, thus giving rise to a positive current flow from the emitter terminal towards the base.

Figure 6.2(a) illustrates the vertical doping profile of an n–p–n transistor with a diffused, or implanted and then diffused, emitter. The emitter junction depth x_{jE} is typically 0.2 μm or larger (Ning and Isaac, 1980). The base junction depth is x_{jB}, and the physical base width is equal to $x_{jB} - x_{jE}$. Figure 6.2(b) illustrates the vertical doping



(a)



(b)

Figure 6.2. Vertical doping profiles of typical n–p–n transistors: (a) with implanted and/or diffused emitter, and (b) with polysilicon emitter.

profile of an n–p–n transistor with a polysilicon emitter. The polysilicon layer is typically about 0.2 μm thick, with an n^+ diffusion into the single-crystal region of only about 30 nm (Nakamura and Nishizawa, 1995). That is, x_{jE} is only about 30 nm.

The base widths of most modern bipolar transistors are typically 0.1 μm or less. While one of the goals in bipolar transistor design is to achieve a base width as small as possible, there are tradeoffs in thin-base designs, as well as difficulties in fabricating thin-base devices. Suffice it to say that the base of a polysilicon-emitter transistor can be made much thinner than that of a diffused-emitter transistor. Details of the doping profiles of the base and collector regions are determined by the desired device dc and ac characteristics and will be discussed in Chapter 7.

6.1.1 Basic Operation of a Bipolar Transistor

As illustrated in Fig. 6.1(a), a bipolar transistor physically consists of two p–n diodes connected back to back. The basic operation of a bipolar transistor, therefore, can be described by the operation of two back-to-back diodes. To turn on an n–p–n transistor, the emitter–base diode is forward biased, resulting in holes being injected from the base into the emitter, and electrons being injected from the emitter into the base. In normal operation, the base–collector diode is reverse biased so that there is no forward current flow in the base–collector diode. (In some circuits, e.g., in simple bipolar inverters and bipolar memory cells, a bipolar transistor may operate having both the emitter–base and collector–base diodes forward biased. Operation of such circuits is discussed in Section 9.1.3.) The bias condition and the energy-band diagram of an n–p–n transistor in normal operation are illustrated in Figs 6.1(c) and 6.1(d).

As described earlier, as the electrons injected from the emitter into the base reach the collector, they give rise to a collector current. The holes injected from the base into the emitter give rise to a base current. One basic objective in bipolar transistor design is to achieve a collector current significantly larger than the base current. The *current gain* of a bipolar transistor is defined as the ratio of its collector current to its base current.

To first order, the behavior of a bipolar transistor is determined by the characteristics of the forward-biased emitter–base diode, since the collector usually acts only as a sink for the carriers injected from the emitter into the base. The emitter–base diode behaves like a thin-base diode. Thus, qualitatively, the current–voltage characteristics of a thin-base diode discussed in Section 2.2.4 can be applied to describe the current–voltage characteristics of a bipolar transistor.

6.1.2 Modifying the Simple Diode Theory for Describing Bipolar Transistors

In order to extend the simple diode theory discussed in Section 2.2 to describe the behavior of a bipolar transistor quantitatively, three important effects ignored in it must be included. These are the effects of finite electric field in a quasineutral region, heavy doping, and nonuniform energy bandgap. These effects are discussed below.

6.1.2.1 Electric Field in a Quasineutral Region with a Uniform Energy Bandgap

In Section 2.2.4, the current–voltage characteristics of a p–n diode were derived for the case of zero electric field in the p- and n-type quasineutral regions. As will be shown below, the zero-field approximation is valid only where the majority-carrier current is zero and concentration is uniform. For bipolar transistors, as shown in Fig. 6.2(a) and (b), the doping profiles are rather nonuniform. A nonuniform doping profile means that the majority-carrier concentration is also nonuniform. Furthermore, at large emitter–base forward biases, to maintain quasineutrality the high concentration of injected minority carriers can cause significant nonuniformity in the majority-carrier concentration as well. Therefore, the effect of nonuniform majority-carrier concentration in a quasineutral region cannot be ignored in determining the current–voltage characteristics of a bipolar transistor.

For a p-type region, Eq. (2.66) gives

$$\phi_p = \psi_i + \frac{kT}{q}\ln\left(\frac{p_p}{n_i}\right), \tag{6.1}$$

where ϕ_p is the hole quasi-Fermi potential and ψ_i is the intrinsic potential. (Note that p_p is equal to N_a only for the case of low electron injection, i.e., only at low currents.) The electric field is given by Eq. (2.41), namely

$$\begin{aligned}\mathscr{E} \equiv -\frac{d\psi_i}{dx} &= \frac{kT}{q}\frac{1}{p_p}\frac{dp_p}{dx} - \frac{d\phi_p}{dx} \\ &= \frac{kT}{q}\frac{1}{p_p}\frac{dp_p}{dx} + \frac{J_p}{qp_p\mu_p},\end{aligned} \tag{6.2}$$

where we have used Eq. (2.64), which relates $d\phi_p/dx$ to J_p. In Eq. (6.2), the intrinsic-carrier concentration is assumed to be independent of x. The dependence of energy bandgap on x will be discussed later in connection with heavy-doping effects.

Let us apply Eq. (6.2) to the intrinsic-base region of an n–p–n transistor with a typical current gain of 100. At a typical but high collector current density of 1 mA/μm^2, the base current density is 0.01 mA/μm^2, i.e., $J_p = 0.01$ mA/μm^2 in the base layer. As can be seen from Fig. 6.2, the base doping concentration is typically on the order of 10^{18} cm^{-3}, and the corresponding hole mobility is about 150 cm^2/V-s (Fig. 2.8). That is, $p_p \approx 10^{18}$ cm^{-3} and $\mu_p \approx 150$ cm^2/V-s, and $J_p/qp_p\mu_p \approx 40$ V/cm, which is a negligibly small electric field in normal device operation. Therefore, for a p-type region Eq. (6.2) gives

$$\mathscr{E}(\text{p-region}) \approx \frac{kT}{q}\frac{1}{p_p}\frac{dp_p}{dx}. \tag{6.3}$$

Similarly, for an n-type region,

$$\mathscr{E}(\text{n-region}) \approx -\frac{kT}{q}\frac{1}{n_n}\frac{dn_n}{dx}. \tag{6.4}$$

Equations (6.3) and (6.4) show that ***the electric field is negligible in a region of uniform majority-carrier concentration.***

To include the effect of finite electric field, the current-density equations (2.54) and (2.55), which include both the drift and the diffusion components, should be used. These are repeated here:

$$J_n(x) = qn\mu_n\mathscr{E} + qD_n\frac{dn}{dx}, \tag{6.5}$$

and

$$J_p(x) = qp\mu_p\mathscr{E} - qD_p\frac{dp}{dx}. \tag{6.6}$$

It should be noted that if Eq. (6.4) is substituted into Eq. (6.5), the RHS of Eq. (6.5) is equal to zero. Similarly, if Eq. (6.3) is substituted into Eq. (6.6), the RHS of Eq. (6.6) is equal to zero. What this means is that the approximations for the electric fields represented by ***Eqs. (6.3) and (6.4) are good approximations only for describing minority-carrier currents***. The $d\phi_p/dx$ term, although very small in a p-region, is entirely responsible for the majority-carrier current in a p-region. In fact, from Eq. (2.64), the hole current density in a p-region is $J_p = -qp\mu_p d\phi_p/dx$. Thus, for describing hole current in a p-region, Eq. (6.2), instead of Eq. (6.3), should be used for the electric field. The electron current in a p-region due to the $d\phi_p/dx$ term, on the other hand, is negligible. Therefore, Eqs. (6.3) and (6.4) are good approximations for describing minority-carrier currents, i.e., for electron current in a p-region and hole current in an n-region. That is, these approximations are applicable to currents in a diode or in a bipolar transistor.

- *Built-in electric field in a nonuniformly doped base region.* Consider the electron current in the p-type base of a forward-biased emitter–base diode. Let $N_B(x)$ be the doping concentration in the base, and, for simplicity, all the dopants are assumed to be ionized. Quasineutrality requires that

$$p_p(x) = N_B(x) + n_p(x). \tag{6.7}$$

Therefore,

$$\frac{dp_p}{dx} = \frac{dN_B}{dx} + \frac{dn_p}{dx}. \tag{6.8}$$

The *built-in electric field* $\mathscr{E}_0$ is defined as the electric field from the nonuniform base dopant distribution alone, ignoring any effect of injected minority carriers. It can be obtained by substituting N_B for p_p in Eq. (6.3), namely

$$\mathscr{E}_0 \equiv \mathscr{E}(n_p = 0) = \frac{kT}{q}\frac{1}{N_B}\frac{dN_B}{dx}. \tag{6.9}$$

Substituting Eq. (6.3) into Eq. (6.5), and using Eqs. (6.8) and (6.9) and the Einstein relationship, we have, for electron current in a nonuniformly doped p-type base region,

$$J_n(x) = qn_p\mu_n\mathscr{E}_0\frac{N_B}{n_p + N_B} + qD_n\left(\frac{2n_p + N_B}{n_p + N_B}\right)\frac{dn_p}{dx}. \tag{6.10}$$

Equation (6.10) suggests that the *effective electric field* $\mathscr{E}_{eff}$ in the p-type base can be written as

$$\mathscr{E}_{eff} = \mathscr{E}_0\frac{N_B}{n_p + N_B}. \tag{6.11}$$

It should be pointed out that Eqs. (6.10) and (6.11) are valid for all levels of electron injection from the emitter, i.e., for all values of n_p.

- *Electric field and current density in the low-injection limit.* At low levels of electron injection from the emitter, i.e., for $n_p << N_B$, $\mathscr{E}_{eff}$ reduces to $\mathscr{E}_0$ and Eq. (6.10) reduces to

$$J_n(x) \approx qn_p\mu_n\mathscr{E}_0 + qD_n\frac{dn_p}{dx}, \tag{6.12}$$

which simply says that the electron current flowing in the base consists of a drift component due to the built-in field from the nonuniform base dopant distribution, and a diffusion component from the electron concentration gradient in the base.

- *Electric field and current density in the high-injection limit.* When the electron injection level is very high, i.e., when $n_p >> N_B$, $\mathscr{E}_{eff}$ becomes very small. The built-in electric field is screened out by the large concentration of injected minority carriers. Therefore, the electron current component associated with the built-in field becomes negligible, and the electron current density approaches

$$J_n(x)\big|_{n_p \gg N_B} \approx q2D_n\frac{dn_p}{dx}. \tag{6.13}$$

That is, at the high-injection limit, the minority-carrier current behaves as if it were purely a diffusion current, but with a diffusion coefficient twice its low-injection value. This is known as the *Webster effect* (Webster, 1954).

6.1.2.2 Heavy-Doping Effect

As discussed in Section 2.1.2.3, the effective ionization energy for impurities in a heavily doped semiconductor decreases with its doping concentration, resulting in a decrease in its effective energy bandgap. For a lightly doped silicon region at thermal equilibrium, Eqs. (2.13) and (2.16) give the relationship between the product p_0n_0 and the energy gap E_g. As the energy gap changes and/or as the densities of states change due the effect of heavy doping, the p_0n_0 product will also change. For modeling purposes, it is convenient to define an *effective intrinsic-carrier concentration* n_{ie} and ***lump all the heavy-doping effects into a parameter called apparent bandgap narrowing,*** ΔE_g, given by the equation

$$p_0(\Delta E_g)n_0(\Delta E_g) \equiv n_{ie}^2 = n_i^2\exp(\Delta E_g/kT). \tag{6.14}$$

The heavy-doping effect increases the effective intrinsic carrier concentration. ***To include the heavy-doping effect, n_i should be replaced by n_{ie}.*** Thus, including heavy-doping effect, the product pn in Eq. (2.67) becomes

$$pn = n_{ie}^2\exp\left[\frac{q(\phi_p - \phi_n)}{kT}\right], \tag{6.15}$$

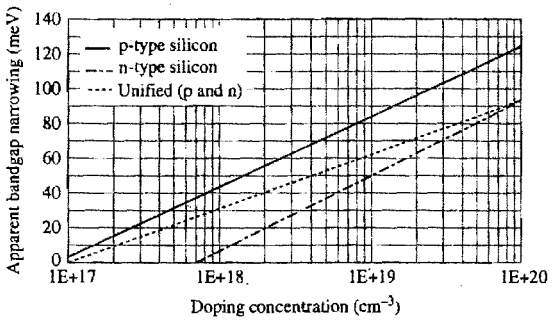


Figure 6.3. Apparent bandgap narrowing as given by the empirical expressions in Eqs. (6.16)–(6.18).

where ϕ_p and ϕ_n are the hole and electron quasi-Fermi potentials, respectively.

It is extremely difficult to determine ΔE_g experimentally and there is considerable scattering in the reported data in the literature (del Alamo *et al.*, 1985a). Careful analyses of the reported data suggest the following *empirical* expressions for the apparent bandgap-narrowing parameter:

$$\Delta E_g(N_d) = 18.7 \ln\left(\frac{N_d}{7 \times 10^{17}}\right) \text{meV} \tag{6.16}$$

for $N_d \geq 7 \times 10^{17}\,\text{cm}^{-3}$, and zero for lower doping levels, for n-type silicon (del Alamo *et al.*, 1985b), and

$$\Delta E_g(N_a) = 9\left(F + \sqrt{F^2 + 0.5}\right) \text{meV}, \tag{6.17}$$

where $F = \ln(N_a/10^{17})$, for $N_a > 10^{17}\,\text{cm}^{-3}$, and zero for lower doping levels, for p-type silicon (Slotboom and de Graaff, 1976; Swirhun *et al.*, 1986). More recently, using a new model that treats both the majority-carrier and minority-carrier mobilities in a unified manner (Klaassen, 1990), Klaassen *et al.* (1992) showed that the heavy-doping effect in both n-type silicon and p-type silicon can be described well by a unified apparent bandgap narrowing parameter. If N represents N_d in n-type silicon and N_a in p-type silicon, then the Klaassen unified apparent bandgap narrowing parameter is given by

$$\Delta E_g(N) = 6.92\left\{\ln\left(\frac{N}{1.3 \times 10^{17}}\right) + \sqrt{\left[\ln\left(\frac{N}{1.3 \times 10^{17}}\right)\right]^2 + 0.5}\right\} \text{meV}. \tag{6.18}$$

Figure 6.3 is a plot of ΔE_g as a function of doping concentration, as given by Eqs. (6.16) to (6.18).

6.1.2.3 Electric Field in a Quasineutral Region with a Nonuniform Energy Bandgap

Aside from the heavy-doping effect, the energy bandgap can also be modified by incorporating a relatively large amount of germanium into silicon. In this case, the bandgap becomes narrower (People, 1986). If both heavy-doping effect and the effect of germanium are included in the parameter ΔE_g in Eq. (6.14), then the product pn given by Eq. (6.15) can be used to describe transport in heavily doped SiGe alloys.

When the energy bandgap is nonuniform, the electric field is no longer simply given by Eqs. (6.3) and (6.4), which include only the effect of nonuniform dopant distribution. When the effect of nonuniform energy bandgap is included, the electric fields are given by (van Overstraeten *et al.*, 1973)

$$\mathscr{E}(\text{p-region}) = \frac{kT}{q}\left(\frac{1}{p_p}\frac{dp_p}{dx} - \frac{1}{n_{ie}^2}\frac{dn_{ie}^2}{dx}\right) \tag{6.19}$$

for a p-type region, and

$$\mathscr{E}(\text{n-region}) = -\frac{kT}{q}\left(\frac{1}{n_n}\frac{dn_n}{dx} - \frac{1}{n_{ie}^2}\frac{dn_{ie}^2}{dx}\right) \tag{6.20}$$

for an n-type region. Derivation of Eq. (6.19) will be shown in Section 7.2.3 in connection with the design of the base region of an n–p–n transistor (see Section 7.2.3).

6.2 Ideal Current–Voltage Characteristics

In Section 2.2.4, the current–voltage characteristics of a p–n diode were derived assuming implicitly that the externally applied voltage appears totally across the immediate junction. All parasitic resistances, and the associated voltage drops due to current flow, were assumed to be negligible. With these assumptions, the currents or current densities in a forward-biased diode increase exponentially with the applied voltage. These are the *ideal current–voltage characteristics*.

In practice, the measured current–voltage characteristics of a bipolar transistor are ideal only over a certain range of applied voltage. At low voltages, the base current is larger than the ideal base current. At large voltages, both the base and the collector currents are significantly smaller than the corresponding ideal currents. In this section, the ideal current–voltage characteristics are discussed. Deviations from the ideal characteristics are discussed in the next section.

It was shown in Section 2.2.5 that, for modern bipolar transistors, ***the base transit time is much smaller than the minority-carrier lifetime in the base, and there is negligible recombination in the base region***. For an n–p–n transistor, neglecting second-order effects, such as avalanche multiplication and generation currents due to defects and/or surface states, the base current is due entirely to the injection of holes from the base into the emitter. Similarly, the collector current is due entirely to the injection of electrons from the emitter into the base. (The effect of avalanche multiplication in the base–collector junction is considered in Section 6.5, where breakdown voltages are discussed. Also, that recombination in the base of modern bipolar transistors is negligible is confirmed in Exercise 6.6).

Referring to Fig. 6.1(a), we see that the base terminal contact is located at the side of the base region. Therefore, the hole current first flows horizontally from the base terminal

into the base region and then bends upward and enters the emitter. The horizontal hole current flow causes a lateral voltage drop within the base region, which in turn causes the forward-bias voltage across the immediate emitter–base junction to vary laterally, with the emitter–base forward bias largest nearest the base contact, and smallest furthest away from the base contact. This is known as *emitter current-crowding effect*. When emitter current crowding is significant, the base and collector current densities are not just a function of x [Fig. 6.1(d)], but also a function of distance from the base contact. Fortunately, as shown in Appendix 16, ***emitter current crowding is negligible in modern bipolar devices*** because of their narrow emitter stripe widths. Therefore, we shall ignore emitter current-crowding effect and assume both the base and collector current densities to be uniform over the entire emitter–base junction area.

Current-Density Equation for Electrons in a p-Type Base

Let us consider the electrons injected from the emitter into the p-type base region of an n–p–n transistor. Instead of starting with Eq. (6.10), it is often convenient to reformulate the electron current density in terms of carrier concentrations (Moll and Ross, 1956). To this end, we start with the electron current density given by Eq. (2.63), namely

$$J_n(x) = -qn_p\mu_n\frac{d\phi_n}{dx}, \tag{6.21}$$

where ϕ_n is the electron quasi-Fermi potential. As we shall show later, the hole current density in the p-type base is small, being smaller than the electron current density by a factor of about 100 (see Section 6.2.3). Also, as indicated in Fig. 6.2, the base region has a reasonably high doping concentration, typically greater than $10^{18}\,\text{cm}^{-3}$ for a modern bipolar transistor. Therefore, the IR drop along the electron-current flow path (which is perpendicular to the intrinsic-base layer) in the p-type base is negligible, which, as discussed in Appendix 4, implies that the hole quasi-Fermi potential ϕ_p is approximately constant. That is, we have

$$\frac{d\phi_p}{dx} \approx 0 \tag{6.22}$$

in the p-type base region. Combining Eqs. (6.21) and (6.22), we obtain

$$J_n(x) \approx qn_p\mu_n\frac{d(\phi_p - \phi_n)}{dx}. \tag{6.23}$$

Now, Eq. (6.15) gives

$$\phi_p - \phi_n = \frac{kT}{q}\ln\left(\frac{p_p n_p}{n_{ie}^2}\right). \tag{6.24}$$

Substituting Eq. (6.24) into Eq. (6.23) and rearranging the terms, we have

$$J_n(x) = qD_n\frac{n_{ie}^2}{p_p}\frac{d}{dx}\left(\frac{n_p p_p}{n_{ie}^2}\right) \tag{6.25}$$

for the electron current in the base. It gives the electron current density in terms of the electron and hole concentrations in the base.

Current-Density Equation for Holes in an n-Type Emitter

The hole current density due to holes injected from the p-type base into the n-type emitter can be derived in a similar manner. The result is

$$J_p(x) = -qD_p\frac{n_{ie}^2}{n_n}\frac{d}{dx}\left(\frac{n_n p_n}{n_{ie}^2}\right). \tag{6.26}$$

Equations (6.25) and (6.26) can be used to calculate the collector and base currents for arbitrary doping profiles, arbitrary energy bandgap grading, and arbitrary injection current levels (Moll and Ross, 1956).

6.2.1 Collector Current

Consider the electrons injected from the emitter into the base. As these electrons reach the collector region, they give rise to a collector current. Referring to Fig. 6.1(d), let $x=0$ denote the depletion-layer edge on the base side of the emitter–base junction, and $x = W_B$ denote the depletion-layer edge on the base side of the base–collector junction. That is, the width of the quasineutral base region is W_B. Since there is negligible recombination in this thin base layer (see Exercise 6.6), the electron current density in steady state in the base is independent of x. Therefore, Eq. (6.25) can be integrated to give

$$J_n\int_0^{W_B}\frac{p_p}{qD_n n_{ie}^2}dx = \left.\frac{n_p p_p}{n_{ie}^2}\right|_{x=W_B} - \left.\frac{n_p p_p}{n_{ie}^2}\right|_{x=0}. \tag{6.27}$$

At $x = 0$, the electron concentration is given by Eq. (2.107), namely,

$$n_p(0) = n_{p0}(0)\exp(qV_{BE}/kT), \tag{6.28}$$

where V_{BE} is the base-emitter forward-bias voltage. At $x = W_B$, the base–collector junction depletion region acts as a sink for the excess electrons in the base region, i.e., $n_p(W_B) = n_{p0}(W_B)$, which is negligible compared to $n_p(0)$. Therefore, the first term on the RHS of Eq. (6.27) can be neglected, and Eq. (6.27) can be rewritten as

$$J_n\int_0^{W_B}\frac{p_p}{qD_n n_{ie}^2}\,dx = -\frac{n_{p0}(0)p_p(0)}{n_{ie}^2(0)}\exp\left(\frac{qV_{BE}}{kT}\right). \tag{6.29}$$

Notice that J_n is negative. This is due to the fact that electrons flowing in the x-direction give rise to a negative current.

Most modern bipolar transistors have a base doping concentration that peaks at or near the emitter–base junction. As long as this peak concentration is large compared to the injected minority-carrier concentration, the majority-carrier concentration near this peak-doping region is about the same as its thermal-equilibrium value. Again referring to the coordinates illustrated in Fig. 6.1(d), this means $p_p(0) \approx p_{p0}(0)$, and Eq. (6.29) is reduced to

$$J_n = -\frac{q\exp(qV_{BE}/kT)}{\int_0^{W_B}(p_p/D_n n_{ie}^2)dx}, \tag{6.30}$$

where we have used the fact that $n_{p0}(0)p_{p0}(0) = n_{ie}^2(0)$.

This electron current is the source of the collector current. Therefore, the collector current I_C is given by

$$I_C = A_E|J_C| = A_E|J_n| = \frac{qA_E\exp(qV_{BE}/kT)}{\int_0^{W_B}(p_p/D_{nB}n_{ieB}^2)dx}, \tag{6.31}$$

where A_E denotes the emitter area, and the subscript B denotes quantities in the base region. [Avalanche multiplication in the base–collector junction will increase the collector current to a value larger than that given by Eq. (6.31). This effect is neglected here but is considered in Section 6.5 in connection with the transistor breakdown voltages.]

The collector current is often written in the form

$$\begin{aligned} I_C &= A_E J_{C0}\exp(qV_{BE}/kT) \\ &= A_E\frac{qn_i^2}{G_B}\exp(qV_{BE}/kT), \end{aligned} \tag{6.32}$$

where J_{C0} is the *saturated collector current density* and G_B is the *base Gummel number* (Gummel, 1961), and n_i is the intrinsic carrier concentration. Comparing Eqs. (6.31) and (6.32) gives

$$J_{C0} = \frac{q}{\int_0^{W_B}(p_p/D_{nB}n_{ieB}^2)dx} \tag{6.33}$$

and

$$G_B = \int_0^{W_B}\frac{n_i^2}{n_{ieB}^2}\frac{p_p}{D_{nB}}dx. \tag{6.34}$$

[In the literature, the base Gummel number is often defined as the total integrated base dose (Gummel, 1961). However, here ***we follow the convention of de Graaff (de Graaff et al., 1977) and define G_B to include both the minority-carrier diffusion coefficient and the effect of heavy doping in the base.*** Thus, $J_{C0} = qn_i^2/G_B$. If heavy-doping effect is negligible and D_{nB} is a constant, then G_B = (total integrated base dose)$/D_{nB}$.]

It should be noted that ***the collector current is a function of the base-region parameters only, and is independent of the properties of the emitter.*** All the effects in the base region, such as bandgap narrowing, bandgap nonuniformity, and dopant distribution, are contained in the parameter G_B. For the special case of a uniformly doped base region at low injection currents, with uniform energy bandgap and negligible heavy-doping effect, the base Gummel number reduces to $N_B W_B/D_{nB}$, and Eq. (6.30) reduces to Eq. (2.131), as expected.

6.2.2 Base Current

Neglecting both base–collector junction avalanche effect and recombination in the base layer, the base current in an n–p–n transistor is equal to the hole current injected from the base into the emitter. Referring to Fig. 6.1(d), let $x = 0$ denote the depletion-layer edge on the emitter side of the emitter–base junction, and $x = -W_E$ denote the location of the ohmic contact to the emitter. W_E is the width of the emitter quasineutral region. Since the emitter is usually so heavily doped that its electron concentration is not affected at all by the hole current level, it is a good approximation to assume $n_n \approx n_{n0} = N_E$, where N_E is the emitter doping concentration. With this approximation, the hole current density in the emitter, i.e., Eq. (6.26), can be rewritten as

$$J_p(x) = -qD_p\frac{n_{ie}^2}{n_{n0}}\frac{d}{dx}\left(\frac{n_{n0}p_n}{n_{ie}^2}\right). \tag{6.35}$$

Equation (6.35) gives the hole current density at any point in the emitter, and $J_p(0)$ is equal to the base current density. It should be noted that ***the base current is a function of the emitter-region parameters only and is independent of the properties of the base region.*** Thus, the base current density changes as the emitter structure and design are changed. In this subsection, we shall use Eq. (6.35) to derive the base current in terms of the more familiar emitter parameters.

6.2.2.1 Shallow or Transparent Emitter

An emitter is considered *shallow* or *transparent* when its width is small compared to its minority-carrier diffusion length. For a shallow emitter, there is negligible recombination in the emitter region except at the emitter contact at $x = -W_E$, and the minority-carrier current density in the emitter is independent of x.

For an n–p–n transistor, the hole current density at the emitter contact is usually written in terms of the surface recombination velocity for holes, S_p, defined by

$$J_p(x = -W_E) \equiv -q(p_n - p_{n0})_{x=-W_E}S_p. \tag{6.36}$$

Notice that J_p is negative because holes flowing in the $-x$ direction give rise to a negative current. Since J_p is independent of x for a transparent emitter, Eq. (6.35) can be rearranged and integrated to give

$$J_p\int_{-W_E}^{0}\frac{n_{n0}}{qD_p n_{ie}^2}dx = -\left.\frac{n_{n0}p_n}{n_{ie}^2}\right|_{x=0} + \left.\frac{n_{n0}p_n}{n_{ie}^2}\right|_{x=-W_E}. \tag{6.37}$$

At $x = 0$, the relation between the hole concentration and the emitter–base voltage is given by Eq. (2.108), namely

$$p_n(0) = p_{n0}(0)\exp(qV_{BE}/kT), \tag{6.38}$$

where V_{BE} is the base–emitter bias voltage. Substituting Eqs. (6.36) and (6.38) into Eq. (6.37), and using the relation $n_{ie}^2 = p_{n0}n_{n0}$, we obtain

$$\begin{aligned} J_p\int_{-W_E}^{0}\frac{n_{n0}}{qD_p n_{ie}^2}dx &= -\exp(qV_{BE}/kT) + 1 - \frac{n_{n0}(-W_E)}{n_{ie}^2(-W_E)qS_p}J_p \\ &\approx -\exp(qV_{BE}/kT) - \frac{n_{n0}(-W_E)}{n_{ie}^2(-W_E)qS_p}J_p, \end{aligned} \tag{6.39}$$

or

$$J_p \approx \frac{-q\exp(qV_{BE}/kT)}{\displaystyle\int_{-W_E}^{0}\frac{n_{n0}}{D_p n_{ie}^2}dx + \frac{n_{n0}(-W_E)}{n_{ie}^2(-W_E)S_p}}. \tag{6.40}$$

Equation (6.40) is valid for a transparent emitter of arbitrary doping profile and arbitrary surface recombination velocity at the emitter contact (Shibib *et al.*, 1979).

Equation (6.40) gives the hole current density entering the emitter. The base current is therefore

$$I_B \equiv A_E|J_B| = A_E|J_p| = \frac{qA_E\exp\left(\dfrac{qV_{BE}}{kT}\right)}{\displaystyle\int_{-W_E}^{0}\frac{N_E}{D_{pE}n_{ieE}^2}dx + \frac{N_E(-W_E)}{n_{ieE}^2(-W_E)S_p}}, \tag{6.41}$$

where A_E is the emitter area, N_E is the emitter doping concentration, and the subscript E denotes parameters in the emitter region.

The base current is often written in the form

$$I_B = A_E J_{B0}\exp(qV_{BE}/kT),$$

or

$$I_B = A_E\frac{qn_i^2}{G_E}\exp\left(\frac{qV_{BE}}{kT}\right), \tag{6.42}$$

where J_{B0} is the *saturated base current density*, and G_E is the *emitter Gummel number* (de Graaff *et al.*, 1977). For a shallow or transparent emitter, Eq. (6.41) gives

$$J_{B0} = \frac{q}{\displaystyle\int_{-W_E}^{0}\frac{N_E}{D_{pE}n_{ieE}^2}dx + \frac{N_E(-W_E)}{n_{ieE}^2(-W_E)S_p}} \tag{6.43}$$

and

$$G_E = \int_{-W_E}^{0}\frac{n_i^2}{n_{ieE}^2}\frac{N_E}{D_{pE}}dx + \frac{n_i^2 N_E(-W_E)}{n_{ieE}^2(-W_E)S_p}. \tag{6.44}$$

- *Transparent emitter with uniform doping concentration and uniform energy bandgap.* Let us consider an n–p–n bipolar transistor with an emitter doping profile as indicated in Fig. 6.2(a). The emitter doping profile is not really uniform or boxlike. Even if we assume the most heavily doped region to be uniform, there is still a transition region where the emitter doping concentration drops from about 10^{20} to about $10^{18}\,\text{cm}^{-3}$ at the emitter–base junction. This transition region plays an important role in determining the emitter–base junction capacitance and the emitter–base junction breakdown voltage. However, as far as the base current is concerned, the effect of this transition region is relatively small (Roulston, 1990). This is due to the fact that the hole diffusion length in this relatively lightly doped transition region is very large compared to the thickness of the region. As a result, the transition region is almost completely transparent to the holes entering the emitter. Therefore, at least for purposes of modeling the base current, it is common to ignore this transition region and simply assume the emitter region to be uniformly doped and boxlike. Besides, such an approximation makes modeling the emitter region relatively simple. For such a uniformly doped transparent emitter with uniform energy bandgap, Eq. (6.43) reduces to

$$J_{B0} = \frac{qD_{pE}n_{ieE}^2}{N_E W_E\left(1 + D_{pE}/W_E S_p\right)}, \tag{6.45}$$

and Eq. (6.44) reduces to

$$G_E = N_E\left(\frac{n_i^2}{n_{ieE}^2}\right)\left(\frac{W_E}{D_{pE}} + \frac{1}{S_p}\right). \tag{6.46}$$

For an ohmic emitter contact, S_p is infinite, and Eq. (6.45) becomes proportional to $1/W_E$, as expected from the properties of a narrow-base diode [cf. Eqs. (2.131) and (2.133)]. The base current increases rapidly as the emitter width, or depth, is reduced.

- *Polysilicon emitter.* The simplest model for describing a polysilicon emitter is to treat the polysilicon-silicon interface located at $x = -W_E$ as a contact with finite surface recombination velocity. In this case, Eq. (6.43) or Eq. (6.45) can be used, depending on whether the single-crystal emitter region is uniformly doped or not. Under certain conditions, a model for the polysilicon emitter can be developed which allows the surface recombination to be evaluated in terms of the properties of the polysilicon layer (Exercise 6.3). In practice, the surface recombination velocity is often used as a fitting parameter to the measured base current. The detailed physics of transport in a polysilicon-emitter is very complicated and is dependent on the polysilicon-emitter fabrication process. Therefore, the surface recombination velocity obtained by fitting to the measured base current is also dependent on the polysilicon-emitter fabrication process. The reader is referred to the vast published literature on polysilicon-emitter physics and technology (Ashburn, 1988; Kapoor and Roulston, 1989).

6.2.2.2 Deep Emitter with Uniform Doping Concentration and Uniform Energy Bandgap

An emitter is *deep*, or *nontransparent*, when its width is large compared to its minority-carrier diffusion length. For a deep emitter, most or all of the injected minority carriers recombine before they reach the emitter contact, and the minority-carrier current is a function of x. The minority-carrier current density given by Eq. (6.35) becomes rather simple if the emitter is assumed to be uniformly doped, has a uniform energy bandgap, and has an ohmic contact at $x = -W_E$. With these assumptions, Eq. (6.35) reduces to

$$J_p(x) = -qD_p\frac{dp_n}{dx}, \tag{6.47}$$

which is simply the hole diffusion current density in the uniformly doped n-side of a diode under low injection, with the hole density given by the equivalent of Eq. (2.122).

The base current density is simply this hole current density at $x = 0$. The base current, therefore, can be obtained from the hole equivalent of Eq. (2.123), namely

$$I_B = A_E|J_p(x=0)| = \frac{qA_E D_{pE} n_{ieE}^2 \exp(qV_{BE}/kT)}{N_E L_{pE} \tanh(W_E/L_{pE})}, \tag{6.48}$$

where A_E is the emitter area, N_E is the emitter doping concentration, and the subscript E denotes parameters in the emitter region. The corresponding base saturation current density and emitter Gummel number are

$$J_{B0} = \frac{qD_{pE} n_{ieE}^2}{N_E L_{pE} \tanh(W_E/L_{pE})}, \tag{6.49}$$

and

$$G_E = \left(\frac{n_i^2}{n_{ieE}^2}\right) \frac{N_E L_{pE} \tanh(W_E/L_{pE})}{D_{pE}}. \tag{6.50}$$

6.2.3 Current Gains

The static *common-emitter* current gain β_0 is defined by

$$\beta_0 \equiv \frac{\partial I_C}{\partial I_B} = \frac{I_C}{I_B} = \frac{J_{C0}}{J_{B0}} = \frac{G_E}{G_B}. \tag{6.51}$$

From Eqs. (6.34) and (6.44),

$$\beta_0 = \frac{\displaystyle\int_{-W_E}^{0} \frac{N_E}{D_{pE} n_{ieE}^2} dx + \frac{N_E(-W_E)}{n_{ieE}^2(-W_E) S_p}}{\displaystyle\int_0^{W_B} \frac{p_p}{D_{nB} n_{ieB}^2} dx} \quad \text{for a transparent emitter,} \tag{6.52}$$

and from Eqs. (6.34) and (6.50),

$$\beta_0 = \frac{N_E L_{pE} \tanh(W_E/L_{pE})}{D_{pE} n_{ieE}^2 \int_0^{W_B} (p_p/D_{nB} n_{ieB}^2) dx} \quad \text{for a deep emitter.} \tag{6.53}$$

The static *common-base current gain* α_0 is defined by

$$\alpha_0 \equiv \frac{\partial I_C}{\partial(-I_E)} = \frac{I_C}{-I_E}, \tag{6.54}$$

where I_E is the emitter current. Here we have defined I_E as the current flowing *into* the emitter, so that $-I_E$ is positive. Since $I_E + I_B + I_C = 0$, we have

$$\alpha_0 = \frac{\beta_0}{1 + \beta_0} \tag{6.55}$$

and

$$\beta_0 = \frac{\alpha_0}{1 - \alpha_0}. \tag{6.56}$$

For modern VLSI bipolar transistors, β_0 is typically about 100. Therefore, α_0 is almost unity. In principle, either α_0 or β_0 can be used to describe the current gain of a bipolar transistor. In practice, β_0 is often used in discussing the device characteristics, device design, and device physics. Throughout this book, we shall use β_0. (However, we shall use α_0 when we consider breakdown voltages in Section 6.5.)

The common-emitter current gain is often quoted as a figure of merit for a bipolar transistor. However, it should be noted that, being the ratio of two currents, the current gain changes as either one of the currents changes. Therefore, to really understand the device design and the device characteristics, both the collector current and the base current, not just the current gain, should be considered. As discussed in the previous subsections, the collector current is a function of only the base parameters, while the base current is a function of only the emitter parameters.

For digital logic circuits, the circuit speed is insensitive to the current gain of the transistors (Ning *et al.*, 1981). However, for many analog circuits, a high current gain is desirable. Most transistors are designed with a current gain of about 100 or larger. For a given bipolar transistor fabrication process, the current gain can be increased or decreased readily by changing the base Gummel number, or the base parameters. Design considerations for the base region will be covered in Section 7.2.

6.2.3.1 Current Gain for Uniformly Doped Deep Emitter and Uniformly Doped Base

For the special case of a uniformly doped emitter with $W_E/L_{pE} \gg 1$, and a uniformly doped base with concentration N_B, Eq. (6.53) reduces to

$$\beta_0 = \frac{n_{ieB}^2 D_{nB} N_E L_{pE}}{n_{ieE}^2 D_{pE} \int_0^{W_B} p_p dx}. \tag{6.57}$$

If the electron current density injected into the base is *low*, then the electron density in the base is also small compared to the hole density, and the hole density is approximately equal to the base doping concentration N_B. In this case, Eq. (6.57) reduces further to

$$\beta_0 = \frac{n_{ieB}^2}{n_{ieE}^2} \frac{D_{nB}}{D_{pE}} \frac{N_E L_{pE}}{N_B W_B}. \tag{6:58}$$

which is independent of current. (The current gain at high currents can be rather complex and will be discussed in Section 6.3.)

It is instructive to estimate the magnitude of the current gain given by Eq. (6.58). If we assume $N_E = 1 \times 10^{20}\,\text{cm}^{-3}$, $N_B = 1 \times 10^{18}\,\text{cm}^{-3}$, and $W_B = 0.1\,\mu\text{m}$ for a typical deep-emitter thin-base n–p–n transistor, then Fig. 6.3 gives $(n_{ieB}/n_{ieE})^2 = \exp[(\Delta E_{gB} - \Delta E_{gE})/kT] \approx 0.19$ at room temperature, Fig. 2.24(a) gives $D_{nB}/D_{pE} = \mu_{nB}/\mu_{pE} \approx 2.6$, $N_E/N_B = 100$, and Fig. 2.24(c) gives $L_{pE}/W_B \approx 4.6$. Substituting these values into Eq. (6.58) gives $\beta_0 = 230$.

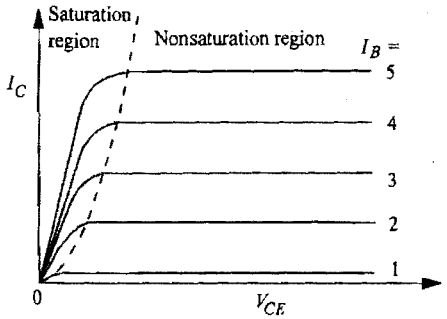


Figure 6.4. Schematic illustration of the ideal I_C-versus-V_{CE} characteristics of an n–p–n transistor. The dashed line is the locus for $V_{CE} = V_{BE}$.

6.2.4 Ideal $I_C - V_{CE}$ Characteristics

Figure 6.4 illustrates the ideal I_C-versus-V_{CE} characteristics of an n–p–n transistor, with I_B as a parameter. Each base current corresponds to a given V_{BE} value. The dashed curve indicates where $V_{CE} = V_{BE}$.

For $V_{CE} < V_{BE}$, the collector–base diode is forward biased and the transistor is said to be in *saturation*. In this case, to first order, the collector current is the difference of the electron current injected from the emitter into the base and the electron current injected from the collector into the base. As a result, the collector current increases with increases in V_{CE}, i.e., as the transistor becomes less saturated.

A transistor is operated in deep saturation when $V_{CE} << V_{BE}$. In general, ***deep saturation is to be avoided*** because the stored charge in the forward-biased collector–base diode increases exponentially with decrease in V_{CE}, and the transistor diffusion capacitance is proportional to the total stored minority charge (see Section 2.2.6). Deep saturation can be and is avoided in all high-speed bipolar circuits. Therefore, only device characteristics in the nonsaturation region are of interest in most applications. In this book, unless stated otherwise, we shall assume the characteristics being considered are for a transistor operated in the nonsaturation region.

However, deep saturation does occur in many relatively slow bipolar circuits. The operation of a bipolar transistor in deep saturation and how deep saturation can be avoided by adding an external resistor to the emitter node are discussed in Section 9.1.3 in connection with bipolar memory circuits.

For $V_{CE} > V_{BE}$, the collector–base diode is reverse biased and the transistor is said to be in its normal *forward-active* mode of operation. All the electrons injected from the emitter into the base are collected by the collector, as recombination in the intrinsic base is negligible in modern transistors, and there is no electron injection from the collector into the base. The collector current is therefore constant, independent of V_{CE}. The current gain is also constant, and the constant-I_B curves are spaced apart by an amount determined by the base-current step, as illustrated in Fig. 6.4.

Note that the schematic in Fig. 6.4 suggests that the collector current is zero when V_{CE} equals zero. This is only a good approximation. Strictly speaking, the collector current in the saturation region has a component due to the injection of holes from the base into the collector. It will be shown later in Section 6.4.1 that in theory the electron current injected from the emitter into the base at $V_{CE} = 0$ cancels exactly the electron current injected from the collector into the base. That this cancellation is almost exact in practical transistors will be shown in Section 7.4.8. Thus, we should expect a small but finite collector current at $V_{CE} = 0$ owing to the injection of holes from the base into the collector. This current is negative because the holes injected from the base are flowing out of the collector. In a linear plot of I_C versus V_{CE} for a typical bipolar transistor, this hole current is usually too small to be noticeable (see Exercise 9.1 in Chapter 9).

The measured current–voltage characteristics of typical bipolar devices are not ideal. The degree of deviation from ideal characteristics depends on the device structure, the device design, the device fabrication process, and on the bias condition of the transistor. The behavior of a typical n–p–n transistor is discussed next.

6.3 Characteristics of a Typical n–p–n Transistor

Figure 6.5 is the Gummel plot of a typical n–p–n transistor. It plots both the collector current I_C and the base current I_B on a logarithmic scale as a function of the forward-bias voltage V_{BE} applied to the emitter and base terminals. The theoretical ideal base and collector currents, discussed in Section 6.2, are indicated by the dashed lines. Figure 6.5

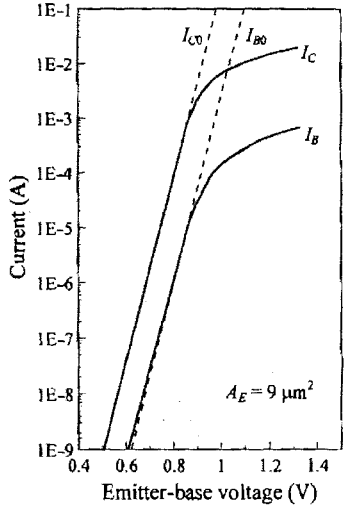


Figure 6.5. Gummel plot of a typical n–p–n bipolar transistor. The dashed lines represent the theoretical ideal base and collector currents. (After Ning and Tang, 1984.)

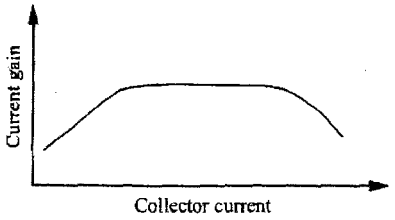


Figure 6.6. Schematic illustration of the current gain I_C/I_B as a function of collector current for a typical bipolar transistor.

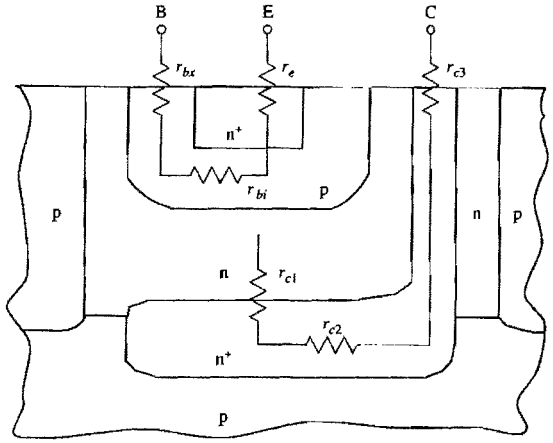


Figure 6.7. Schematic illustrating the parasitic resistances in a typical modern n–p–n transistor.

shows that the measured collector current is ideal except at large V_{BE}, while the measured base current is ideal except at small and at large V_{BE}.

Figure 6.6 illustrates the typical measured current gain, I_C/I_B, as a function of collector current. For the voltage range where both the base and the collector currents are approximately ideal, the current gain is approximately constant. At low currents, the current gain is less than its ideal value because the base current is larger than its ideal value. At high currents, the current gain rolls off with collector current because the percentage by which the collector current is smaller than its ideal value is larger than the percentage by which the base current is smaller than its ideal value. The dominant physical mechanisms responsible for the nonideal behavior of the base and collector currents are discussed in the subsections below.

6.3.1 Effect of Emitter and Base Series Resistances

Figure 6.7 shows schematically the physical origins of the parasitic resistances in a typical n–p–n transistor. These resistances are ignored in Section 6.2 in the description of

the ideal current–voltage characteristics. As the currents flow through these parasitic resistors, voltage drops are developed, which tend to offset the externally applied voltages. The parasitic resistances can therefore be neglected at low currents but can be very important at large currents.

In normal forward-active operation, the base–collector junction is reverse biased. In most bipolar circuits, particularly those designed for high-speed applications, the collector–base junction is designed to remain reverse biased at all times, even at high currents. This is accomplished by employing a heavily doped subcollector layer (to reduce r_{c2}) and a heavily doped reach-through (to reduce r_{c3}) to bring the collector contact to the surface. With the base–collector junction reverse biased, to first order, the collector resistance components shown in Fig. 6.7 have no effect on the current flows in the emitter–base diode, and only the parasitic resistances associated with the emitter and the base need to be considered. (The effect of collector–base voltage on collector current is discussed in the following subsection.)

The emitter series resistance r_e is determined primarily by the emitter contact resistance, since the resistance associated with the thin n^+ emitter region is small. The base resistance r_b can be separated into two components: the intrinsic-base resistance r_{bi}, which is determined by the design of the intrinsic-base region, and the extrinsic-base resistance r_{bx}, which includes all other resistances associated with the base terminal.

The emitter–base diode voltage drop due to the flow of emitter and base currents is

$$\begin{aligned}\Delta V_{BE} &= -I_E r_e + I_B r_b \\ &= I_C r_e + I_B(r_e + r_b),\end{aligned} \tag{6.59}$$

where we have used the fact that $I_E + I_B + I_C = 0$. The relation between the voltage V_{BE} applied to the emitter and base terminals and the voltage V'_{BE} appearing across the immediate emitter–base junction is

$$V'_{BE} = V_{BE} - \Delta V_{BE}. \tag{6.60}$$

To include the effect of the emitter and base series resistances, the equations in Section 6.2 for the ideal collector and base currents should be modified by replacing V_{BE} by V'_{BE}. This results in both the measured collector and base currents, when plotted as a function of V_{BE}, being significantly smaller than the ideal currents at large V_{BE}, as illustrated in Fig. 6.5.

As can be seen from Eq. (6.33), even in the ideal case, the collector saturation current density is a function of the majority-carrier concentration in the base and the base width. Therefore, ***the measured collector current is a function of ΔV_{BE} as well as a function of the base majority-carrier concentration and the base width, which in turn depend on V_{BE}***. The dependence of I_C on V_{BE} is very complex, as can be seen in later subsections.

On the other hand, as can be seen from Eqs. (6.43) and (6.49), the base saturation current density is a function of the emitter parameters only, which, due to the emitter being very heavily doped, do not vary with the minority-carrier injection level. Therefore, at high currents, ***deviation of the base current from its ideal behavior is due to ΔV_{BE} alone*** (Ning and Tang, 1984). The relation between the ideal base current I_{B0} and the measured base current I_B is therefore

$$I_{B0} = I_B \exp(q\Delta V_{BE}/kT), \tag{6.61}$$

which can be used to evaluate the emitter and base series resistances. This is shown in Appendix 15. Many other methods for determining the emitter and base series resistances have been discussed in the literature (Schroder, 1990). Some of these are discussed in Appendix 15 as well.

6.3.2 Effect of Base–Collector Voltage on Collector Current

In many transistors, particularly in modern high-speed transistors where the base width is very small, the measured collector current, and hence the measured current gain, increases as the base–collector reverse-bias voltage is increased. This is due to two effects, or a combination of them. The first effect is the dependence of the quasineutral base width on collector–base voltage. The second effect is the avalanche multiplication in the base–collector junction. We shall discuss these two effects individually in this subsection.

6.3.2.1 Modulation of Quasineutral Base Width by Base–Collector Voltage

As the reverse bias across the base–collector junction is increased, the base–collector junction depletion-layer width increases, and hence the quasineutral base width W_B decreases. This in turn causes the collector current to increase, as can be seen from Eq. (6.31). Thus, instead of as illustrated in Fig. 6.4, where the collector current is independent of collector voltage for $V_{CE} > V_{BE}$, the collector current of a typical bipolar transistor increases with collector voltage, as illustrated in Fig. 6.8.

- *Early voltage.* For circuit modeling purposes, the collector current in the nonsaturation region is often assumed to depend linearly on the collector voltage. The collector voltage at which the linearly extrapolated I_C reaches zero is denoted by $-V_A$. As we shall show later, it is a good and useful approximation to assume that V_A is independent of V_{BE}. This is illustrated in Fig. 6.8. V_A is called the *Early voltage* (Early, 1952). It is defined by

$$V_A + V_{CE} \equiv I_C \left(\frac{\partial I_C}{\partial V_{CE}} \right)^{-1}. \tag{6.62}$$

In practice, except for transistors that tend to punch through (to be discussed later), V_A is much larger than the operation range of V_{CE}. Therefore, V_A can be approximated by

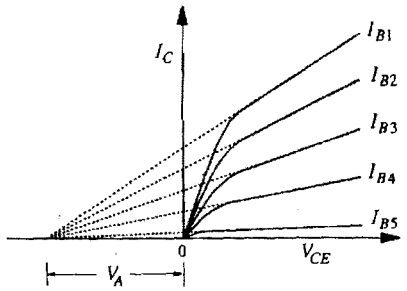


Figure 6.8. Schematic illustrating the approximately linear dependence of I_C on V_{CE}. The linearly extrapolated I_C intersects the V_{CE}-axis at $-V_A$.

$$V_A \approx I_C \left(\frac{\partial I_C}{\partial V_{CE}} \right)^{-1}. \tag{6.63}$$

The collector current is given by Eq. (6.32), which can be written as

$$I_C = A_E J_{C0} \exp(qV_{BE}/kT) = \frac{qA_E \exp(qV_{BE}/kT)}{F(W_B)}, \tag{6.64}$$

where, for convenience, a function F has been introduced (Kroemer, 1985) which is defined by

$$F(W_B) \equiv \frac{q}{J_{C0}} = \int_0^{W_B} \frac{p_p(x)}{D_{nB}(x) n_{ieB}^2(x)} dx. \tag{6.65}$$

The majority-carrier hole charge per unit area in the base is

$$Q_{pB} = q \int_0^{W_B} p_p(x) dx. \tag{6.66}$$

Since V_{BE} is fixed for a given I_B, Eq. (6.63) can be rewritten as

$$\begin{aligned} V_A &\approx I_C \left(\frac{-I_C}{F} \frac{\partial F}{\partial V_{CE}} \right)^{-1} \\ &= \left(\frac{-1}{F} \frac{\partial F}{\partial W_B} \frac{\partial W_B}{\partial Q_{pB}} \frac{\partial Q_{pB}}{\partial V_{CE}} \right)^{-1} \\ &= \left(\frac{-1}{F} \frac{\partial F}{\partial W_B} \frac{\partial W_B}{\partial Q_{pB}} \frac{\partial Q_{pB}}{\partial V_{CB}} \right)^{-1}. \end{aligned} \tag{6.67}$$

Notice that $V_{CB} = -V_{BC}$. As explained in Section 2.2.2.2 in the derivation of Eq. (2.83) for the depletion-layer capacitance for a p–n junction, when the p-side (base) voltage is changed relative to the n-side (collector) by ΔV_{BC}, the p-side depletion charge changes by an amount equal to the change in the majority-hole charge ΔQ_{pB} in the p-side. Therefore,

$$-\frac{\partial Q_{pB}}{\partial V_{CB}} = \frac{\partial Q_{pB}}{\partial V_{BC}} = C_{dBC}, \tag{6.68}$$

where C_{dBC} is the base–collector junction depletion-layer capacitance per unit area [cf. Eq. (2.83)]. The other two derivatives in Eq. (6.67) can be evaluated directly, namely

$$\frac{\partial F}{\partial W_B} = \frac{p_p(W_B)}{D_{nB}(W_B) n_{ieB}^2(W_B)} \tag{6.69}$$

and

$$\frac{\partial W_B}{\partial Q_{pB}} = \left(\frac{\partial Q_{pB}}{\partial W_B} \right)^{-1} = \frac{1}{qp_p(W_B)}. \tag{6.70}$$

Therefore, Eq. (6.67) gives

$$V_A \approx \frac{qD_{nB}(W_B) n_{ieB}^2(W_B)}{C_{dBC}} \int_0^{W_B} \frac{p_p(x)}{D_{nB}(x) n_{ieB}^2(x)} dx. \tag{6.71}$$

For a uniformly doped base, Eq. (6.71) reduces to

$$V_A \approx \frac{Q_{pB}}{C_{dBC}}. \tag{6.72}$$

At sufficiently low collector currents such that the base majority-carrier concentration is approximately the same as its equilibrium value, i.e., $p_p \approx p_{p0} = N_B$, Eq. (6.71) gives

$$V_A \approx \frac{qD_{nB}(W_B)n_{ieB}^2(W_B)}{C_{dBC}} \int_0^{W_B} \frac{N_B(x)}{D_{nB}(x)n_{ieB}^2(x)} dx. \tag{6.73}$$

Equation (6.73) is independent of base current, so that the slope of the curves in Fig. 6.8 intercept the V_{CE}-axis at the same value, namely V_A, as illustrated. It is instructive to estimate the magnitude of Eq. (6.73) for a uniformly doped base. In this case, $V_A \approx qW_BN_B/C_{dBC}$. For a base of $W_B = 0.1$ μm and $N_B = 10^{18}\,\text{cm}^{-3}$, we have $qW_BN_B \approx 1.6 \times 10^{-6}\,\text{C/cm}^2$. For a collector of $N_C = 2 \times 10^{16}\,\text{cm}^{-3}$, then, from Fig. 2.16, $C_{dBC} \approx 4 \times 10^{-8}\,\text{F/cm}^2$. Therefore, $V_A \approx 40$ V. In practice, V_A can vary a lot as the transistor design is "optimized." This will be discussed further later in this section and in Chapter 7. As can be seen in Eq. (6.71), V_A is a function of W_B, which, as discussed earlier, is a function of the collector voltage. Therefore, strictly speaking, the Early voltage is a function of the collector voltage at which the slope is used for extrapolating to $I_C = 0$. In other words, strictly speaking, I_C does not increase lineraly with V_{CE}. However, the linear dependence is a good approximation and is a useful approximation for circuit analyses and modeling purposes. The Early voltage is a figure of merit for devices used in analog circuits. The larger the Early voltage, the more independent is the collector current on collector voltage. Another device figure of merit is the product of the current gain and Early voltage. Using Eqs. (6.33), (6.51), and (6.71), this product can be written as (Prinz and Sturm, 1991)

$$\beta_0 V_A = \frac{q^2 D_{nB}(W_B)n_{ieB}^2(W_B)}{C_{dBC}J_{B0}}, \tag{6.74}$$

where the base saturated current density J_{B0} is a function of the emitter parameters. That is, while V_A is a function of the base parameters only, the product $\beta_0 V_A$ is a function of both the emitter and the base parameters.

- *Emitter–collector punch-through.* As shown in Eq. (6.72), the Early voltage is proportional to the majority-carrier charge in the base. As the collector voltage is increased, the width of the quasineutral base region, and hence the majority-carrier charge in the base, is reduced. For a device with a small majority-carrier base charge or small Early voltage to start with, it does not take much increase in collector voltage before all the majority-carrier base charge is depleted, or before the collector punches through to the emitter. At collector–emitter punch-through, the collector current becomes excessively large, being limited only by the emitter and collector series resistances. The collector current at or close to punch-through is no longer controlled adequately by the base voltage for proper device operation. ***Punch-through must be avoided*** under normal device operation, by designing the device to have a sufficiently large majority-carrier base charge.

6.3.2.2 Base–Collector Junction Avalanche

For a device with a large majority-carrier base charge or large Early voltage to begin with, as the collector voltage is increased, usually the condition of significant base–collector junction avalanche is reached before punch-through is reached. This is certainly the case for transistors where the collector side of the base–collector junction space–charge–layer boundary reaches the subcollector region before punch-through occurs, because as the base–collector junction space-charge-layer boundary reaches the heavily doped subcollector, further increase of the base–collector reverse bias will increase the junction electric field very rapidly.

For an n–p–n transistor, the base–collector junction avalanche process is illustrated in Fig. 6.9(a) (Lu and Chen, 1989). As the electrons injected from the emitter into the base reach the base–collector junction space-charge region, they can cause impact ionization and generate electron–hole pairs. The secondary electrons flow towards the collector terminal, adding to the measured collector current, while the secondary holes flow towards the base terminal, subtracting from the measured base current. If the secondary hole current is large enough, the current measured at the base terminal could be negative (Lu and Chen, 1989). This is illustrated in Fig. 6.9(b).

At very small emitter–base forward biases, the measured base current is positive as usual. The secondary hole current is not large enough to completely offset the usual base current. As the electron current injected from the emitter into the base–collector space-charge region increases with increased emitter–base forward bias, the secondary hole current increases and may reach a point at which the measured base current turns negative. At sufficiently large emitter–base forward biases, as will be discussed in the next subsection, significant base widening can occur and the electric field in the base–collector junction can be reduced. As a result, avalanche multiplication is reduced and the measured base current returns to positive.

The magnitude of base–collector junction avalanche depends on the maximum electric field in the base–collector junction. To minimize base–collector junction avalanche, techniques for reducing the maximum electric field in a p–n junction, such as retrograding the collector doping profile or sandwiching a lightly doped layer between the base and the collector, can be used (Tang and Lu, 1989). The concept is similar to that of a p–i–n diode discussed in Section 2.2.2.

6.3.3 Collector Current Falloff at High Currents

The collector saturation current density for an n–p–n transistor is given by Eq. (6.33), namely

$$J_{C0} = \frac{q}{\displaystyle\int_0^{W_B} \frac{p_p(x)}{D_{nB}(x)n_{ieB}^2(x)} dx}. \tag{6.75}$$

There are a number of physical mechanisms that can cause the denominator in Eq. (6.75) to increase, and hence J_{C0} to decrease, as the collector current density is increased. As J_{C0} falls off, the collector current I_C falls off with it [see Eq. (6.32)]. This collector current

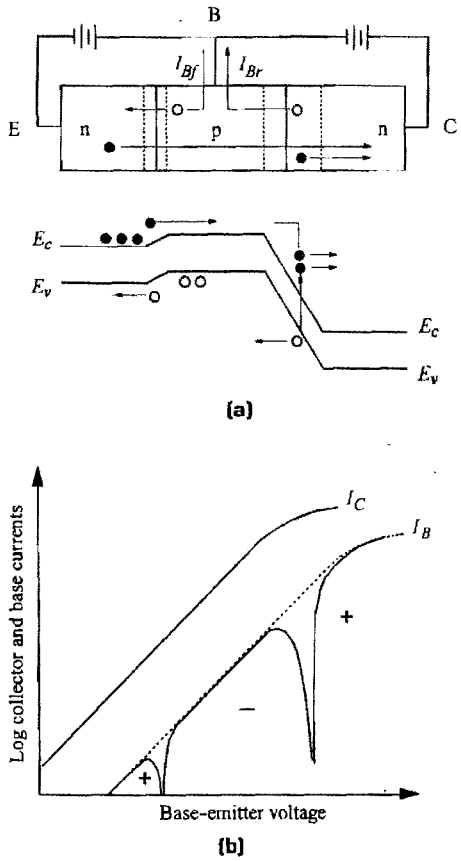


Figure 6.9. (a) Schematics of an n–p–n transistor operated in the forward-active mode with a large base–collector voltage. As electron–hole pairs are generated in the base–collector junction space-charge region, the secondary hole current, I_{Br}, subtracts from the usual forward base current, I_{Bf}. The current measured at the base terminal is $I_B = I_{Bf} - I_{Br}$. (b) Typical Gummel plot of an n–p–n transistor where significant avalanche multiplication occurs in the base–collector junction space-charge region. (After Lu and Chen, 1989.)

falloff at high currents is on top of the effect of emitter and base series resistances discussed in Section 6.3.1. These physical mechanisms are discussed in this subsection.

As electrons are injected into the p-type base, the hole concentration in the base, $p_p(x)$, increases in order to maintain charge neutrality. If this increase in hole concentration is appreciable, J_{C0} decreases. At the same time, as the injected electrons reach the base–collector junction, they add to the space charge in the base–collector junction space-charge region, resulting in widening of the quasineutral base layer. An increase in W_B also causes J_{C0} to decrease. This is known as *base-conductivity modulation effect*. It should be

pointed out that in the literature, $p_p(x)$ in Eq. (6.75) is often approximated by $N_B(x)$. This approximation is good only at small emitter–base biases where the injected minority electron density in the base is small compared to the base doping concentration. At large emitter–base biases, this approximation underestimates the base-conductivity modulation effect.

As W_B increases, the collector side of the base–collector space-charge layer also widens into the collector. At sufficiently high collector current densities, base widening can push the "base–collector junction" deep into the collector region. This is known as *base-widening* or *Kirk effect* (Kirk, 1962), and it also causes J_{C0} to decrease.

Base-conductivity modulation and base-widening effects are not really separate and do not act independently. Dependent on the details of the device design, their combined effect can contribute significantly to the observed saturation of the collector current in a Gummel plot. A combination of base-conductivity modulation and base widening is responsible for the current-gain rolloff at high collector currents depicted in Fig. 6.6. In this subsection we discuss base widening in more detail.

6.3.3.1 Base Widening at Low Currents

Consider the base–collector junction of an n–p–n transistor. For simplicity, let us assume the base region to have a uniform doping concentration N_B, and the collector region to have a uniform doping concentration N_C. When the transistor is turned off, the charge distribution in the base–collector junction is as shown schematically in Fig. 6.10(a), where x_{B0} and x_{C0} are the widths of the depletion regions on the base side and on the collector side, respectively. The relationship between these widths is given by Eq. (2.78), namely,

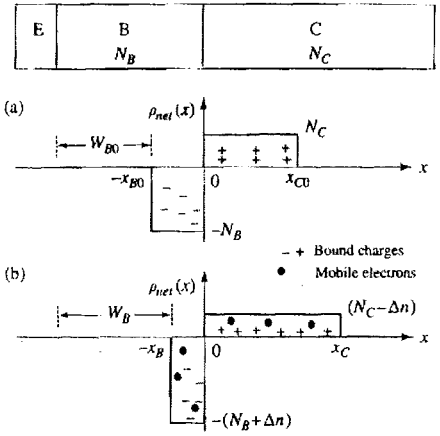


Figure 6.10. Schematics illustrating the charge distribution in the base–collector junction of an n–p–n transistor. (a) Emitter–base diode is not forward biased, and (b) emitter–base diode is forward biased.

$$x_{B0}N_B = x_{C0}N_C. \tag{6.76}$$

The maximum potential drop across the base–collector junction, ψ_{mBC}, is given by Eq. (2.79), which can be rewritten as

$$\psi_{mBC} = \frac{q}{2\varepsilon_{si}}\left(N_B x_{B0}^2 + N_C x_{C0}^2\right). \tag{6.77}$$

When the n–p–n transistor is turned on, electrons are injected into the base and collector regions. These mobile electrons add to the space charge in the base–collector junction region. As long as this additional mobile-electron concentration is small compared with the ionized doping concentrations, the depletion approximation discussed in Section 2.2.2 can be used to estimate its effect. For simplicity, let us assume these mobile electrons traverse the base–collector junction space-charge region at a saturated velocity v_{sat}. The mobile electron concentration Δn in the space-charge region is given by the relation

$$J_C = qv_{sat}\Delta n, \tag{6.78}$$

where J_C is the collector current density. The space-charge concentration on the base side is increased from N_B to $N_B + \Delta n$, and the space-charge concentration on the collector side is decreased from N_C to $N_C - \Delta n$. As a result, the width of the depletion region on the base side is decreased to x_B, and the width of the depletion region on the collector side is increased to x_C, such that

$$x_B(N_B + \Delta n) = x_C(N_C - \Delta n). \tag{6.79}$$

This is illustrated schematically in Fig. 6.10(b). The width of the quasineutral base layer is widened by an amount equal to $x_{B0} - x_B$.

An estimation of the amount of base widening can be made quantitatively if the emitter–base junction is assumed to be forward biased so that the base–collector junction voltage remains unchanged (Ghandhi, 1968). In this case, Eq. (6.77) is replaced by

$$\psi_{mBC} = \frac{q}{2\varepsilon_{si}}\left[(N_B + \Delta n)x_B^2 + (N_C - \Delta n)x_C^2\right]. \tag{6.80}$$

Combining Eqs. (6.77) and (6.80), and assuming $\Delta n/N_C << 1$, we have

$$\begin{aligned} x_C &= x_{C0}\sqrt{\frac{1 + (\Delta n/N_B)}{1 - (\Delta n/N_C)}} \\ &\approx \frac{x_{C0}}{\sqrt{1 - (\Delta n/N_C)}} \end{aligned} \tag{6.81}$$

where we have used the fact that N_B is typically much larger than N_C, so that $\Delta n/N_B << 1$. Similarly

$$\begin{aligned} x_B &= x_{B0}\sqrt{\frac{1 - (\Delta n/N_C)}{1 + (\Delta n/N_B)}} \\ &\approx x_{B0}\sqrt{1 - (\Delta n/N_C)}. \end{aligned} \tag{6.82}$$

6.3.3.2 Base Widening at High Currents

At high current densities, the assumption of Δn being small compared to N_C is no longer valid, and the above equations cannot be used to estimate the base-widening effect. With the mobile-charge concentration comparable to or larger than the fixed ionized-impurity concentration, the depletion approximation is certainly not valid. Furthermore, the excess electrons in the n-type collector can produce a substantial electric field in the collector, according to Eq. (6.4), and the classical concept of a well-defined junction boundary in the base–collector diode is no longer valid. Also, in order to maintain quasineutrality, the excess electrons induce an excess of holes in the n-type collector. The region of the collector with excess holes becomes an extension of the p-type base. In other words, the base region widens into the collector region, until it reaches the subcollector where the excess electron concentration is small compared with the n-type doping concentration. As a result, the high-field region, originally located at the physical base–collector junction, is relocated to near the collector–subcollector intersection (Poon *et al.*, 1969). The numerical simulation results (Poon *et al.*, 1969) shown in Fig. 6.11 illustrate clearly the effects of base widening at high currents. They show that the relocation of the high-field region is accompanied by a buildup of excess electrons and holes in the collector region.

It is instructive to estimate the collector current density at which substantial base widening occurs. The saturated velocity v_{sat} for electrons in silicon is about 1×10^7 cm/s, as indicated in Fig. 2.10. At low collector currents, the maximum electron concentration in the n-type collector region is equal to the collector doping concentration N_C. The maximum electron current density that can be supported by an electron concentration of N_C is $J_{max} = qv_{sat}N_C$. When the injected electron current density approaches J_{max}, the electron concentration has to increase to a value larger than N_C in order to support the injected electron current flow, i.e., there is a density of excess electrons caused by the high electron current density. As the excess electrons build up, there is a build up of excess holes in order to maintain quasineutrality, and a relocation of the high-field region. The results shown in Fig. 6.11 suggest that significant base widening starts at a collector current density of approximately $0.3J_{max}$. This value is consistent with the reported peak cutoff-frequency data for modern VLSI bipolar devices (Crabbé *et al.*, 1993a). Thus, ***to avoid significant base widening, a bipolar transistor should not be operated at collector current densities approaching J_{max}***. For a relatively high N_C of 2×10^{17} cm^{-3}, J_{max} is about 3.2 mA/μm^2. To avoid significant base widening, J_C should be less than about 1 mA/μm^2.

6.3.4 Nonideal Base Current at Low Currents

As shown in Fig. 6.5, for small emitter–base voltages, the base current is larger than its ideal value. The origins of this excess base current are (a) the generation–recombination current in the emitter–base junction depletion region and (b) the tunneling current in the emitter–base junction (Li *et al.*, 1988). The amount of deviation from ideal behavior depends strongly on the transistor structure, device design, and fabrication process. For

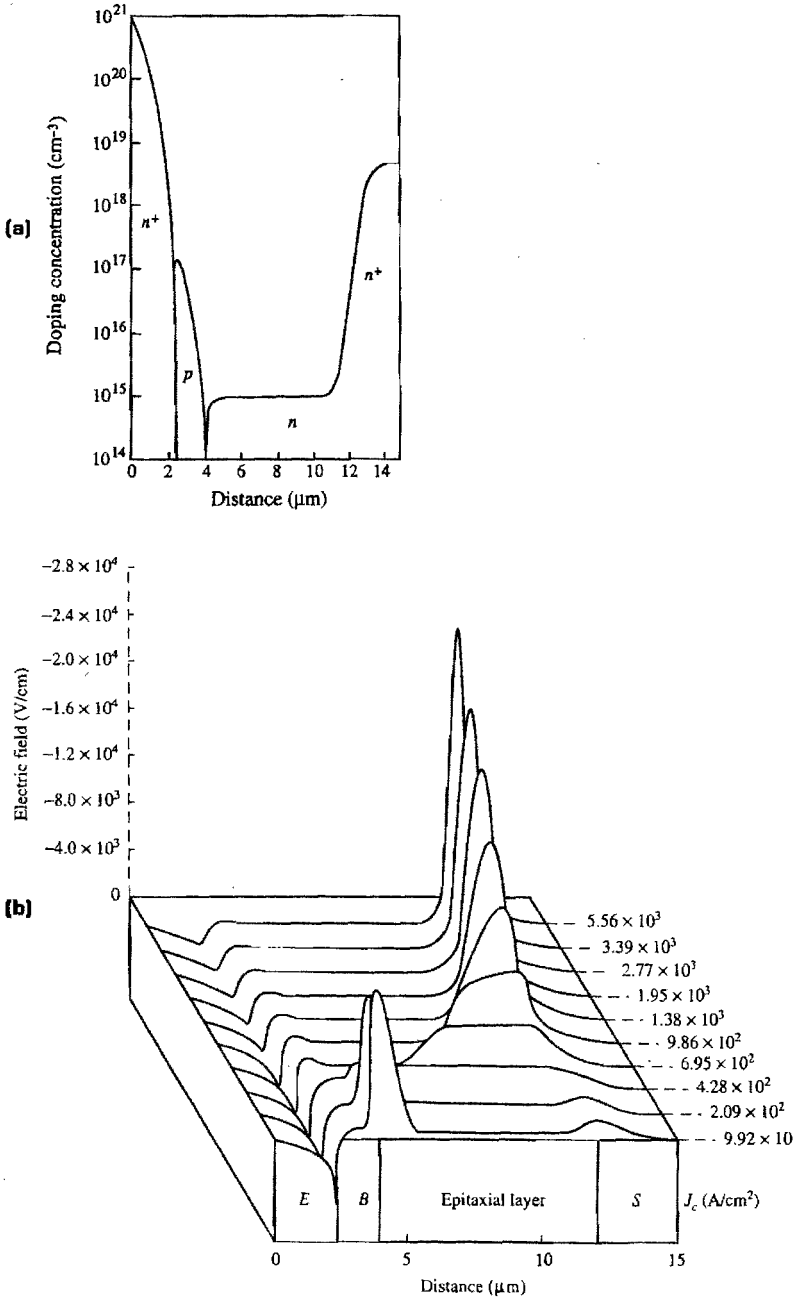


Figure 6.11. Numerical simulation results showing the effects of base widening in an n–p–n transistor at high collector current densities: (a) the doping profiles of the device simulated, (b) relocation of the high-field region from the physical base–collector junction to the collector–subcollector intersection, (c) buildup of excess holes in the collector, and (d) buildup of excess electrons in the collector. (After Poon *et al.*, 1969.)

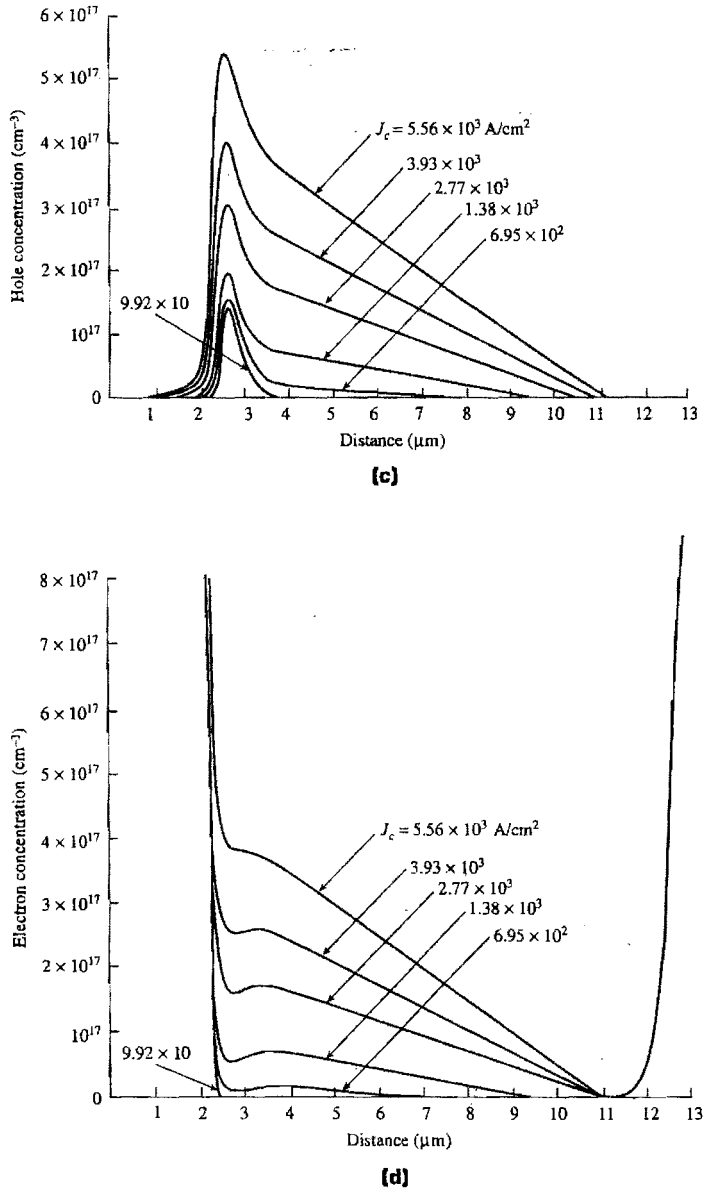


Figure 6.11. (cont.)

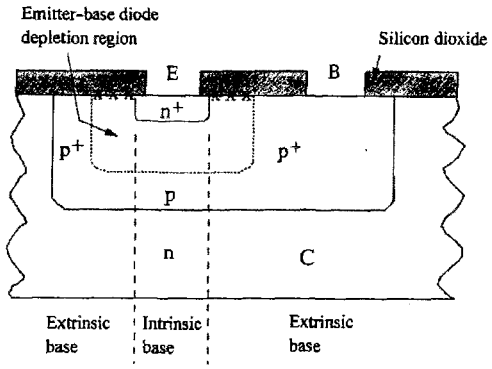


Figure 6.12. Schematic illustrating the cross section of an emitter–base diode. The extrinsic base is usually much more heavily doped than the intrinsic base. The presence of surface states, indicated by x x x, can cause excessive base current, as discussed in the text.

most well-designed bipolar transistors and fabrication processes, this excess current is often negligibly small. In any case, since this excess current is often quite noticeable in experimental devices, particularly before the fabrication process has been optimized, its physical origins are discussed here.

Figure 6.12 illustrates schematically the cross section of an emitter–base diode. The base region directly underneath the emitter is referred to as the *intrinsic base*, and the remaining parts of the base are collectively referred to as the *extrinsic base*. The entire emitter–base diode can be considered as two diodes connected in parallel, one formed by the emitter and the intrinsic base, and the other by the emitter and the extrinsic base. The intrinsic base has been the subject of our discussion so far. The function of the extrinsic base is to provide electrical connection to the intrinsic base from the silicon surface.

To minimize parasitic resistance and to minimize electron injection from the emitter into the extrinsic base region, the extrinsic base is usually doped much more heavily than the intrinsic base. As a result, the collector-current component due to electrons injected from the emitter into the extrinsic base and reaching the collector is negligible compared to the collector-current component due to electrons traversing the intrinsic base. This can be concluded readily from Eq. (6.31). The large width of and high doping concentration in the extrinsic base make its contribution to the collector current very small compared to contribution from the intrinsic base.

Nonetheless, the extrinsic-base–emitter diode can contribute appreciably to the measured base current. This extrinsic-base current has three components, namely (a) the current associated with the injection of holes from the extrinsic base into the emitter, (b) the generation–recombination current, and (c) the tunneling current. The current associated with the injection of holes from the extrinsic base into the emitter has the same dependence on V_{BE} as the current associated with the injection of holes from the intrinsic base into the emitter. Therefore, this current simply adds to the ideal intrinsic-base current

and will not show up as deviation of the measured base current from its ideal behavior. Only the generation–recombination and tunneling currents contribute to the nonideal behavior of the measured base currents. Therefore, only these two components are discussed further here.

6.3.4.1 Base Current Due to Generation–Recombination

Generation-recombination current due to defect centers in silicon is negligible in modern VLSI devices because, unless there is a contamination problem, the concentration of defects that can cause generation–recombination current is negligibly low for all modern VLSI fabrication processes. This may not be true for processes in early development, but it is certainly true by the time a process reaches manufacturing. However, as can be seen from Fig. 6.12, the extrinsic-base–emitter diode has a surface component. The presence of interface states, as indicated in the figure, could give rise to significant surface generation–recombination current, as discussed in Section 2.3.7. The generation–recombination hole current adds to the base current and hence degrades the current gain (Werner, 1976). Surface generation–recombination current, by itself, usually can be recognized by its $\exp(V_{BE}/2kT)$ dependence on V_{BE}, as discussed in Section 2.2.4.10. Fortunately, for properly designed fabrication processes, the density of interface states can be so low that this current component, though usually observable, is not significant in modern bipolar devices.

6.3.4.2 Base Current due to Tunneling

The tunneling current in the emitter–base junction, on the other hand, is expected to increase as the transistor dimensions are scaled down (Stork and Isaac, 1983). The emitter–base junction is a fairly abrupt junction. The emitter is very heavily doped, and the base doping concentration is typically in excess of $1 \times 10^{18}\ \text{cm}^{-3}$, as can be seen from Fig. 6.2. Furthermore, as discussed in Chapters 7 and 8, the peak base doping concentration is increased as the physical dimensions of a bipolar transistor are scaled down, resulting in enhanced tunneling current in the emitter–base diode.

Since the extrinsic base is always more heavily doped than the intrinsic base, the observed tunneling current is usually dominated by the component from the extrinsic-base–emitter diode. Furthermore, the interface states in the extrinsic-base–emitter diode can assist in the tunneling process and thus can enhance the tunneling current very significantly (Li *et al.*, 1988). Figure 6.13 illustrates the typical current–voltage characteristics of a bipolar transistor which has excessive base current due to tunneling in the emitter–base diode. When excessive emitter–base tunneling dominates the base current, the base current is usually much larger than suggested by an $\exp(V_{BE}/2kT)$ dependence. Furthermore, as expected from a tunneling process, the excessive emitter–base tunneling current is nearly independent of temperature (Li *et al.*, 1988). Also, the excessive emitter–base tunneling current increases very rapidly with voltage when the emitter–base diode is reverse biased, as can be seen from Fig. 6.13. Fortunately, excessive tunneling current in the emitter–base diode can be suppressed easily by optimizing the emitter–base diode doping profile and the device fabrication process.

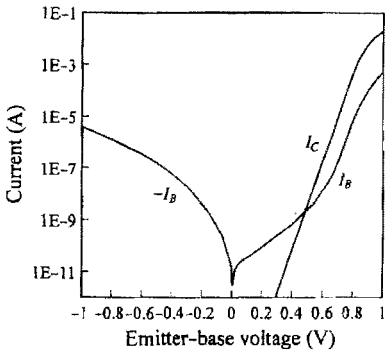


Figure 6.13. Typical voltage–current characteristics of an n–p–n transistor which has excessive tunneling current in the extrinsic-base–emitter diode. (After Li *et al.*, 1988.)

6.4 Bipolar Device Models for Circuit and Time-Dependent Analyses

The merits of a bipolar device should be discussed in the context of the circuit in which it is used. For circuit applications, the device electrical characteristics must be first transformed into equivalent circuit parameters. The merits of a device are then interpreted from the behavior of the circuit or from the characteristics of the equivalent-circuit parameters. In this section, the equivalent-circuit models needed for the discussion of bipolar device design, which will be covered in Chapter 7, and device optimization, which will be covered in Chapter 8, are developed. The models suitable for dc or large-signal analyses will be developed first, followed by the models suitable for small-signal analyses. This is followed by the development of the charge-control model, which is suitable for quasistatic time-dependent analyses.

6.4.1 Basic dc Model

The Ebers–Moll model (Ebers and Moll, 1954) for an n–p–n transistor is shown in Fig. 6.14. It describes an n–p–n transistor as two diodes in series, arranged in the common-base mode. When a voltage V_{BE} is applied to the emitter–base diode, a forward current I_F flows in the emitter–base diode. This current causes a current $\alpha_F I_F$ to flow in the collector, where α_F is the common-base current gain in the forward direction. Similarly, when a voltage V_{BC} is applied across the base–collector diode, a reverse current I_R flows in the collector–base diode, causing a current $\alpha_R I_R$ to flow in the emitter, where α_R is the common-base current gain in the reverse direction. These currents are indicated in Fig. 6.14. They are related by

$$I_E = \alpha_R I_R - I_F, \tag{6.83}$$

$$I_C = \alpha_F I_F - I_R, \tag{6.84}$$

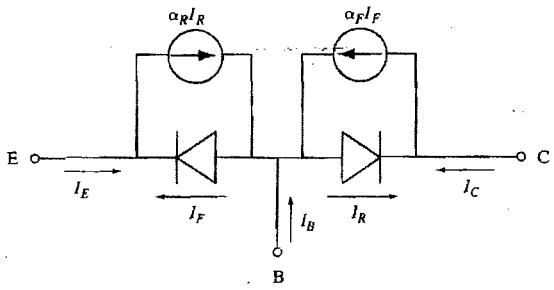


Figure 6.14. Equivalent-circuit representation of the basic dc Ebers–Moll model of an n–p–n transistor.

and

$$I_B = (1 - \alpha_F)I_F + (1 - \alpha_R)I_R. \tag{6.85}$$

The emitter, base, and collector currents are related by $I_E + I_B + I_C = 0$.

From Eq. (2.120) we can write I_F and I_R in the form

$$I_F = I_{F0}[\exp(qV_{BE}/kT) - 1] \tag{6.86}$$

and

$$I_R = I_{R0}[\exp(qV_{BC}/kT) - 1]. \tag{6.87}$$

Therefore, Eqs. (6.83) and (6.84) can be rewritten as

$$I_E = -I_{F0}[\exp(qV_{BE}/kT) - 1] + \alpha_R I_{R0}[\exp(qV_{BC}/kT) - 1] \tag{6.88}$$

and

$$I_C = \alpha_F I_{F0}[\exp(qV_{BE}/kT) - 1] - I_{R0}[\exp(qV_{BC}/kT) - 1]. \tag{6.89}$$

Reciprocity characteristics of the emitter and collector terminals require the off-diagonal coefficients of the equations for I_E and I_C to be equal (Gray *et al.*, 1964; Muller and Kamins, 1977), i.e.,

$$\alpha_R I_{R0} = \alpha_F I_{F0}. \tag{6.90}$$

Alternatively, the reciprocity relationship in Eq. (6.90) can be shown as follows. We note that $\alpha_F I_{F0}$ is the saturated collector current in the forward active mode, and $\alpha_R I_{R0}$ is the saturated collector current in the reverse active mode. For simplicity, we consider a hypothetical transistor having a unit cross-sectional area for both the emitter–base junction and the collector–base junction. In this case, $\alpha_F I_{F0}$ is given by Eq. (6.33) with the integral in the denominator being from $x = 0$ to $x = W_B$ and $\alpha_R I_{R0}$ is also given by Eq. (6.33) but with the integral in the denominator being from $x = W_B$ to $x = 0$. Since the value of an integral of a function is independent of its direction of integration, we have $\alpha_F I_{F0} = \alpha_R I_{R0}$ for our hypothetical transistor having the same emitter–base and collector–base junction areas. It should be noted that no assumption has been made about

the doping and bandgap-narrowing parameters in the base, suggesting that the reciprocity relationship applies to Si-base as well as SiGe-base transistors. (See Section 7.4 for discussion of SiGe-bipolar transistors.) Also, it has been shown through experiments and simulation studies (Rieh *et al.*, 2005) that the saturated collector currents in forward and reverse modes are approximately the same in typical bipolar transistors, even though the collector–base junction area in a typical bipolar transistor is much larger than its emitter–base junction area.

The common-emitter form of the Ebers–Moll model is often more desirable for circuit analyses. To accomplish this, let us define

$$I_{SF} \equiv \alpha_F I_F, \tag{6.91}$$

$$I_{SR} \equiv \alpha_R I_R, \tag{6.92}$$

$$I_{CT} \equiv I_{SF} - I_{SR}, \tag{6.93}$$

$$\beta_F \equiv \frac{\alpha_F}{1-\alpha_F}, \tag{6.94}$$

$$\beta_R \equiv \frac{\alpha_R}{1-\alpha_R}. \tag{6.95}$$

Comparison with Eq. (6.56) shows that β_F and β_R are the common-emitter current gains in the forward and the reverse directions, respectively. Substituting Eqs. (6.90) to (6.95) into Eqs. (6.83) to (6.85) gives

$$I_E = -I_{CT} - \frac{I_{SF}}{\beta_F}, \tag{6.96}$$

$$I_C = I_{CT} - \frac{I_{SR}}{\beta_R}, \tag{6.97}$$

$$I_B = \frac{I_{SF}}{\beta_F} + \frac{I_{SR}}{\beta_R}. \tag{6.98}$$

The equivalent-circuit model for these currents is shown in Fig. 6.15.

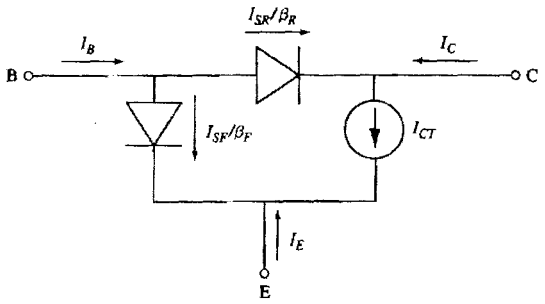


Figure 6.15. Common-emitter equivalent-circuit representation of the dc Ebers–Moll model of an n–p–n transistor.

6.4.2 Basic ac Model

To model the ac behavior of a bipolar transistor, the parasitic internal capacitances and resistances of the transistor must be included. In general, the parasitic resistances can be made rather small by using large device areas and device layout techniques, as well as fabrication process techniques. However, the parasitic capacitances usually can be reduced only by reducing the associated device areas. As a result, the basic behavior of a transistor is determined more by its parasitic capacitances than by its parasitic resistances. For simplicity, we shall first neglect the parasitic resistances and consider only the parasitic capacitances.

As discussed in Section 2.2.6, there are two components in the capacitance of a p–n diode, namely the depletion-layer capacitance and the diffusion capacitance. Let $C_{dBE,tot}$ and $C_{dBC,tot}$ be the depletion-layer capacitances of the emitter–base and collector–base diodes, respectively. Let C_{DE} be the diffusion capacitance associated with forward-biasing the emitter–base diode, and C_{DC} be the diffusion capacitance associated with forward-biasing the collector–base diode. When these capacitances are included, the common-emitter equivalent-circuit model is shown in Fig. 6.16. In Fig. 6.16, the depletion-layer capacitance of the collector-substrate diode, $C_{dCS,tot}$, is also included for completeness.

6.4.2.1 Model for a Transistor Biased in the Forward-Active Mode of Operation

For simplicity, we shall consider only transistors biased in the forward-active mode of operation, i.e., with the emitter–base diodes forward biased and the base–collector diodes reverse biased. (Transistors biased in the reverse-active mode, i.e., with the base–collector diodes forward biased, cannot be switched fast because of the very large diffusion capacitance associated with the forward-biased base–collector diodes. As a result, high-speed circuits usually use transistors biased only in the forward-active mode.) In this case, I_{SR} can be neglected compared to I_{SF}, and $C_{DC} = 0$. The model in Fig. 6.16 then simplifies to that shown in Fig. 6.17.

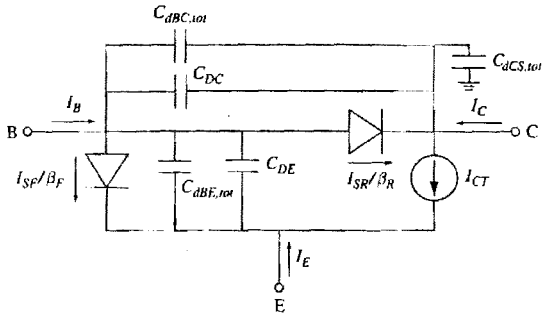


Figure 6.16. Common-emitter equivalent-circuit representation of the ac Ebers–Moll model of a bipolar transistor. Internal capacitances are included.

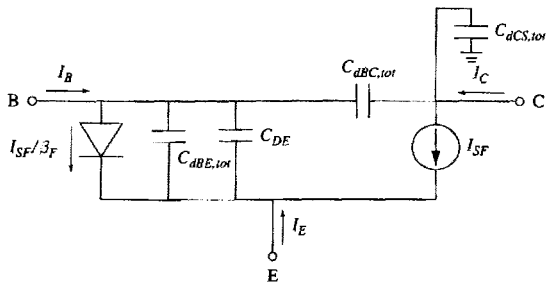


Figure 6.17. Equivalent-circuit representation of the ac Ebers–Moll model of an n–p–n transistor biased in the forward-active mode of operation. Internal capacitances are included.

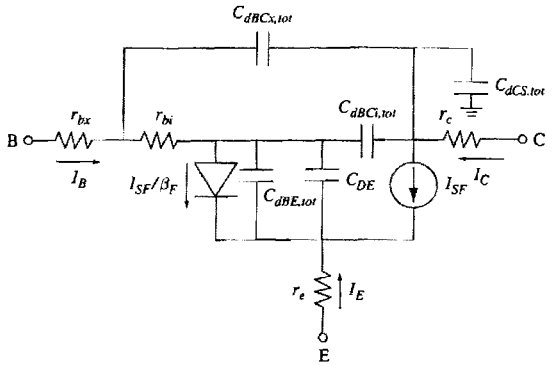


Figure 6.18. Equivalent-circuit representation of the ac Ebers–Moll model of an n–p–n transistor biased in the forward-active mode of operation. Internal parasitic resistance and capacitance are included.

If the internal parasitic resistances indicated in Fig. 6.7 are now included in the equivalent circuit of Fig. 6.17, the resultant equivalent circuit is shown in Fig. 6.18. Here, for purposes of discussion in later chapters, the base resistance is shown as two parts, an intrinsic part r_{bi} and an extrinsic part r_{bx}. The depletion-layer capacitance of the base–collection diode is also separated into an intrinsic part $C_{dBCi,tot}$ and an extrinsic part $C_{dBCx,tot}$.

6.4.3 Small-Signal Equivalent-Circuit Model

Consider a small-signal voltage applied to the input base terminal of a common-emitter equivalent circuit shown in Fig. 6.17 or Fig. 6.18. It will cause small variations in the base and collector currents as well as in the collector terminal voltage. A small-signal equivalent-circuit model provides a relationship among these current and voltage variations. We first develop the small-signal equivalent-circuit model for an intrinsic device

ignoring the transistor parasitic resistances, and then the model for an extrinsic device with these resistances included.

- *Small-signal model when parasitic resistances are negligible.* The Ebers–Moll model for an intrinsic transistor is shown in Fig. 6.17. Let us denote the steady-state base–emitter voltage by V'_{BE} and the collector–emitter voltage by V'_{CE}. The corresponding small-signal voltages are denoted as v'_{be} and v'_{ce}. Here, the convention is such that the primed parameters refer to an intrinsic device, while the unprimed parameters are for an extrinsic device. The corresponding small-signal base and collector currents are i_b and i_c, respectively. The intrinsic transconductance g'_m relates i_c to v'_{be}, i.e.,

$$g'_m = \frac{i_c}{v'_{be}} = \left.\frac{\partial I_C}{\partial V'_{BE}}\right|_{V'_{CE}} = \frac{qI_C}{kT}, \tag{6.99}$$

where we have used the fact that I_C is proportional to $\exp(qV'_{BE}/kT)$. The intrinsic input resistance r'_π relates v'_{be} to i_b, i.e.,

$$r'_\pi = \frac{v'_{be}}{i_b} = \frac{1}{\left.\dfrac{\partial I_B}{\partial V'_{BE}}\right|_{V'_{CE}}} = \frac{kT}{qI_B} = \frac{\beta_0}{g'_m}, \tag{6.100}$$

where we have used the fact that I_B is proportional to $\exp(qV'_{BE}/kT)$ and that $\beta_0 = I_C/I_B$. The intrinsic output resistance r'_0 relates i_c to v'_{ce}, i.e.,

$$r'_0 = \frac{v'_{ce}}{i_c} = \frac{1}{\left.\dfrac{\partial I_C}{\partial V'_{CE}}\right|_{V'_{BE}}} = \frac{V_A}{I_C}, \tag{6.101}$$

where we have used Eq. (6.63) for the Early voltage V_A. The capacitances are designated by

$$C_\mu = C_{dBC,tot}, \tag{6.102}$$

and

$$C_\pi = C_{dBE,tot} + C_{DE}. \tag{6.103}$$

The resulting small-signal equivalent circuit is shown in Fig. 6.19. This is the well-known small-signal hybrid-π model (Gray *et al.*, 1964).

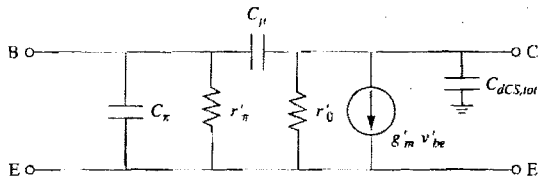


Figure 6.19. Small-signal hybrid-π model of a bipolar transistor when the parasitic resistances are neglected.

- *Small-signal model including parasitic resistances.* When parasitic resistances are included, the device terminal voltages V_C, V_E and V_B are no longer the same as the internal junction voltages V'_C, V'_E and V'_B, respectively, owing to the IR drops in the parasitic resistors r_e, r_c, and r_b. For simplicity, we have lumped r_{bi} and r_{bx} into r_b. The terminal voltages and the internal junction voltages are related by

$$V_C = V'_C + I_C r_c, \tag{6.104}$$

$$V_B = V'_B + I_B r_b, \tag{6.105}$$

and

$$\begin{aligned} V_E &= V'_E + I_E r_e \\ &= V'_E - (I_C + I_B) r_e, \end{aligned} \tag{6.106}$$

where we have used the fact that $I_E + I_C + I_B = 0$. Therefore,

$$\begin{aligned} V_{BE} &= V'_{BE} + I_B r_b + (I_C + I_B) r_e \\ &= V'_{BE} + I_C r_e + I_B (r_b + r_e) \end{aligned} \tag{6.107}$$

and

$$\begin{aligned} V_{CE} &= V'_{CE} + I_C r_c + (I_C + I_B) r_e \\ &\approx V'_{CE} + I_C (r_e + r_c), \end{aligned} \tag{6.108}$$

where in the last equation we have neglected the IR drop due to I_B compared with that due to I_C.

The extrinsic transconductance g_m relates i_c to v_{be}, i.e.,

$$\begin{aligned} g_m &= \frac{i_c}{v_{be}} = \frac{i_c}{v'_{be} + i_c r_e + i_b (r_b + r_e)} \\ &= \left(\frac{1}{g'_m} + r_e + \frac{r_b + r_e}{\beta_0} \right)^{-1} \approx \left(\frac{1}{g'_m} + r_e \right)^{-1} \\ &= \frac{g'_m}{1 + g'_m r_e}, \end{aligned} \tag{6.109}$$

where we have neglected $(r_b + r_e)/\beta_0$ relative to r_e and used Eq. (6.99) for g'_m. Similarly, the extrinsic input resistance r_π relates v_{be} to i_b, i.e.,

$$\begin{aligned} r_\pi &= \frac{v_{be}}{i_b} = \frac{v'_{be} + i_c r_e + i_b (r_b + r_e)}{i_b} \\ &= r'_\pi (1 + g'_m r_e) + (r_b + r_e), \end{aligned} \tag{6.110}$$

where we have used Eq. (6.100) for r'_π. The extrinsic output resistance r_0 relates i_c to v_{ce}, i.e.,

$$\begin{aligned} r_0 &= \frac{v_{ce}}{i_c} = \frac{v'_{ce} + i_c (r_e + r_c)}{i_c} \\ &= r'_0 + (r_e + r_c), \end{aligned} \tag{6.111}$$

where we have used Eq. (6.101) for r'_0.

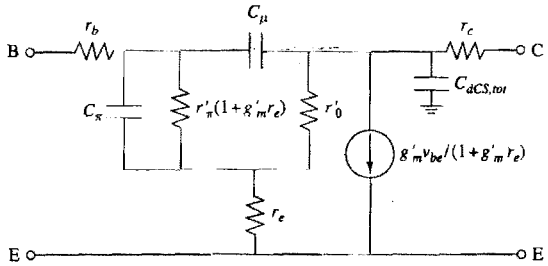


Figure 6.20. Small-signal hybrid-π model of a bipolar transistor including parasitic resistances.

The device capacitance components are still the same as before, with C_μ given by Eq. (6.102) and C_π given by Eq. (6.103). It should be noted that C_μ is determined by V'_{BC}, and not V_{BC}. Similarly, C_π is determined by V'_{BE}, and not by V_{BE}. The equivalent circuit can be deduced from Eqs. (6.109) to (6.111), and is shown in Fig. 6.20.

6.4.4 Emitter Diffusion Capacitance

Consider a small ac signal superimposed on a dc forward bias across the emitter–base diode. The diffusion capacitance C_{DE} is due to the minority carriers in the transistor that can respond to the small signal. Minority carriers that cannot respond to the signal do not contribute to the capacitance. (See Section 2.2.6 and Appendix 6.) Minority carriers are present in the emitter region, the base region, as well as in the space-charge regions of the emitter–base and base–collector diodes. The total minority-carrier charge can therefore be written as the sum of these individual charges:

$$Q_{DE} = |Q_{E,tot,ac}| + |Q_{B,tot,ac}| + |Q_{BE,tot,ac}| + |Q_{BC,tot,ac}|, \tag{6.112}$$

where $Q_{E,tot,ac}$, $Q_{B,tot,ac}$, $Q_{BE,tot,ac}$, and $Q_{BC,tot,ac}$ represent the minority-carrier charge in the emitter, the base, the emitter–base space-charge region, and the base–collector space-charge region, respectively, that can respond to the ac signal and contribute to the diffusion capacitance.

A note about the symbols used to denote minority-charge quantities here and elsewhere in the book is needed. As an illustration, let us consider the minority charge in the base region. In Eq. (2.144), we use Q_B to denote the minority charge per unit area in the base region of a diode. We shall use $Q_{B,tot}$ to denote the total minority charge in the base region. In the case of a simple diode with cross-sectional area A_{diode}, $Q_{B,tot}$ is simply $A_{diode} Q_B$. However, in general, $Q_{B,tot}$ cannot be written simply as $A_{diode} Q_B$. $Q_{B,tot}$ can be determined accurately only by using two-dimensional or three-dimensional numerical simulations. Nonetheless, it is often mathematically convenient to assume such a simple relationship, especially for explaining the basic physics governing device operation. Therefore, we shall use $A_{diode} Q_B$ to mean $Q_{B,tot}$ in many cases. It should be remembered that it is $Q_{B,tot}$ that should be used in quantitative device modeling. Similar comments apply to the other minority-charge quantities Q_E, Q_{BE}, and Q_{BC}.

The subscript 'ac' on the RHS of Eq. (6.112) is to distinguish these quantities from their corresponding steady-state values which can be larger. [As discussed in Section 2.2.6 and in Appendix 6, $Q_{E,ac}$ is equal to 1/2 of the steady-state quantity Q_E for a wide emitter device and $Q_{B,ac}$ is equal to 2/3 of the steady-state quantity Q_B. $Q_{BE,ac}$ and $Q_{BC,ac}$ are usually assumed to be the same as the corresponding steady-state quantities. This is a good assumption because of the high field in the space-charge regions. In a high field space-charge region, the electrons travel at their saturated velocity v_{sat} which is about 1×10^7 cm/s (see Fig. 2.10). For a space-charge layer width of 0.1 μm, the average transit time for the electrons is on the order of 10^{-12} s. This time is short compared to $1/f$, where f is the frequency of a typical signal. That is, electrons in a space-charge region should be able to respond to an ac signal.] Notice that Q_{DE} is the sum of the *absolute values* of the individual minority-carrier charge components, and not the summation of the net charge components. As an illustration of this important distinction, consider a hypothetical transistor having a perfectly symmetrical emitter–base diode, with the n-region doping concentration equal to the p-region doping concentration (not a good transistor design, but a transistor nonetheless). In this case, we have $Q_{B,ac} = -Q_{E,ac}$. The contributions of $Q_{B,ac}$ and $Q_{E,ac}$ to Q_{DE} are $|Q_{B,ac}|$ and $|Q_{E,ac}|$, and not $Q_{B,ac} + Q_{E,ac}$, which is zero. (Also, see the discussion at end of the next subsection.)

If the intrinsic base–emitter forward-bias voltage is $v'_{BE}(t)$, then the emitter diffusion capacitance is

$$C_{DE} \equiv \frac{\partial Q_{DE}}{\partial v'_{BE}}. \tag{6.113}$$

For modeling purposes, it is convenient to consider the collector current $i_C(t)$ being the charging current and rewrite Eq. (6.112) in the form

$$Q_{DE} \equiv \tau_F i_C(t), \tag{6.114}$$

where τ_F is referred to as the *forward transit time*. As we shall show later, at low current densities where base widening is negligible, each of the minority-charge components in Eq. (6.112) is simply proportional to the collector current. In this case, τ_F is independent of the base–emitter bias. If we write the intrinsic base–emitter forward-bias voltage in the form $v'_{BE}(t) = V'_{BE} + v'_{be}(t)$, where V'_{BE} is the dc bias and $v'_{be}(t)$ is the small signal, then

$$i_C(t) \approx I_C\left[1 + \frac{qv'_{be}(t)}{kT}\right] \tag{6.115}$$

and Eqs. (6.113) and (6.114) give

$$C_{DE} = \tau_F \frac{\partial i_C}{\partial v'_{BE}} = \tau_F \frac{qI_C}{kT} = \tau_F g'_m \quad \text{(low current dencity)}, \tag{6.116}$$

where I_C is the steady-state collector current determined by V'_{BE} and g'_m is the intrinsic transconductance given by Eq. (6.99). However, at sufficiently large current densities, base-widening occurs, and, as discussed in Section 6.3.3, the total minority charge in the base, $Q_{B,tot}$, increases rapidly with collector current density. Therefore, we expect $Q_{B,tot,ac}$ to increase rapidly with collector current density as well. In this case, τ_F is no longer independent of the base–emitter bias. Instead, it increases rapidly with collector current density, and Eq. (6.116) is no longer valid.

Comparing Eqs. (6.112) and (6.114), we see that τ_F has contributions from $Q_{E,tot,ac}$, $Q_{B,tot,ac}$, $Q_{BE,tot,ac}$, and $Q_{BC,tot,ac}$. To help distinguish the various contributions, τ_F is often written as the sum of these components, namely,

$$\tau_F = \tau_E + \tau_B + \tau_{BE} + \tau_{BC}. \tag{6.117}$$

In Eq. (6.117), τ_E is the emitter delay time, representing the contribution from $Q_{E,tot,ac}$, τ_B is the base delay time, representing the contribution from $Q_{B,tot,ac}$, τ_{BE} is the base–emitter space-charge region delay time, representing the contribution from $Q_{BE,tot,ac}$, and τ_{BC} is the base–collector space-charge-region delay time, representing the contribution from $Q_{BC,tot,ac}$ (Ashburn, 1988). The emitter is being charged by the base current, so that we expect $|Q_{E,tot,ac}|$ to be proportional to I_B (see Section 2.2.6). For a wide emitter, Eq. (2.166) suggests that $|Q_{E,tot,ac}| = I_B\tau_{pE}/2 = I_C\tau_{pE}/2\beta_0$, where τ_{pE} is the hole lifetime in the n^+ emitter and β_0 is the common-emitter current gain. (Remember, here the Qs include only the portion of minority charge that can follow the ac signal and contribute to the emitter diffusion capacitance. For a wide emitter, this portion is 1/2 of the total minority charge in the emitter.) Similarly, the base is being charged by the collector current, so that we expect $|Q_{B,tot,ac}|$ to be proportional to I_C. However, this is the case only when there is negligible base widening. In this case Eq. (2.165) suggests $|Q_{B,tot,ac}| = 2I_C t_B/3$, where t_B is the base transit time. When base widening occurs, $|Q_{B,tot,ac}|$ increases with I_C at a much faster rate.

The space-charge-region delay time is equal to the average transit time for the corresponding space-charge region. This time is $W_d/2v_{sat}$, where W_d is the depletion layer width and v_{sat} is the saturated electron velocity (Meyer and Muller, 1987). Considerations of these delay-time components in the design of a bipolar transistor will be covered in Chapter 8 (see Section 8.3.3).

6.4.5 Charge-Control Analysis

The behavior of a bipolar transistor is often analyzed in a *charge-control* model where the charges within the various regions of the transistor are related to the currents feeding them. The charge-control model is especially useful for transient analyses. It was used in Section 2.2.5 to describe the discharging of a diode that has been switched from forward bias to reverse bias. In this subsection, we describe the time-dependent behavior of an n–p–n transistor using charge-control analysis.

As we shall show later, the starting point for applying charge-control analysis is after spatial integration of the continuity equation for the physical region of interest. In other words, an entire transistor region is considered as one lumped component. As a result, ***a charge-control analysis does not yield or depend on information about the distribution of the minority charge within the region***. A charge-control method is thus limited to

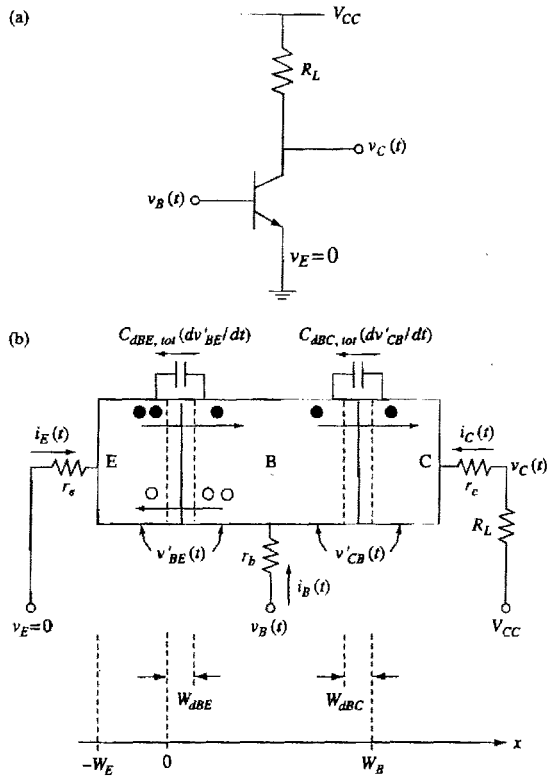


Figure 6.21. (a) Schematic of an n–p–n transistor biased to operate as an amplifier. The input voltage v_B is assumed to be time dependent. (b) Schematic illustrating the resistances and terminal currents in the amplifier. Also illustrated are the displacement currents and the flow of electrons and holes within the transistor. The locations of the emitter contact, the emitter–base boundary, and the base–collector boundary, used in the charge-control model, are also indicated. W_{dBE} and W_{dBC} are the base–emitter and the base–collector junction depletion-layer widths, respectively.

quasistatic situations where all the minority charge within the region of interest can respond fully to a time-dependent voltage. Charge-control analysis is not suitable for situations where the distributed nature of the stored charge is important, e.g., in the derivation of the diffusion capacitance (see Section 2.2.6 and Appendix 6). ***Charge-control method should not be used for small-signal ac analysis of bipolar transistors without great care.***

Consider an n–p–n transistor biased in an amplifier mode. Its circuit schematic is shown in Fig. 6.21(a). The input voltage, which is the base terminal voltage, is assumed to be time dependent. The currents flowing in the transistor are illustrated schematically in Fig. 6.21(b). The time-dependent emitter current, base current and collector current are

denoted by $i_E(t)$, $i_B(t)$, and $i_C(t)$, respectively. The displacement currents in the base–emitter and base–collector junction depletion-layer capacitors are also included.

As the electrons flow through the emitter–base and base–collector junction space-charge regions, they contribute to the mobile charges Q_{BE} and Q_{BC} stored in these regions. (As the holes flow from the base into the emitter, they also contribute to a mobile charge component in the emitter–base space-charge region. However, this hole component, which is proportional to the base current, is small compared with the electron component, which is proportional to the collector current. For simplicity, the hole component of mobile charge stored in the emitter–base space-charge region is ignored.) To facilitate including Q_{BE} and Q_{BC} in the charge-control analysis, we define the *base region* to include the emitter–base space-charge layer and the base–collector space-charge layer. Thus, for our charge-control analysis, the emitter contact is located at $x=-W_E$, the emitter–base boundary is located at $x=0$, and the base–collector boundary is located at $x=W_B$, as illustrated in Fig. 6.21(b).

For mathematical simplicity, let us assume a one-dimensional transistor structure having a cross-section area of A. From Eq. (2.110), the continuity equation for the excess electrons in the p-type base is

$$A\frac{\partial(n_p-n_{p0})}{\partial t}=\frac{1}{q}\frac{\partial i_n(x,t)}{\partial x}-A\frac{n_p-n_{p0}}{\tau_{nB}}, \tag{6.118}$$

where $i_n(x, t)$ is the electron current in the base and τ_{nB} is the electron lifetime in the base. Multiplying both sides of Eq. (6.118) by $-q$ and integrating over the base region, we have

$$-Aq\frac{\partial}{\partial t}\int_0^{W_B}(n_p-n_{p0})dx=-\int_0^{W_B}\frac{\partial i_n(x,t)}{\partial x}dx+A\frac{q}{\tau_{nB}}\int_0^{W_B}(n_p-n_{p0})dx, \tag{6.119}$$

which is the starting equation for charge-control analysis. The excess electron charge per unit area stored in the base is

$$-q\int_0^{W_B}(n_p-n_{p0})dx=Q_{BE}+Q_B+Q_{BC}, \tag{6.120}$$

where Q_{BE}, Q_B, and Q_{BC} are the excess electron charges per unit area stored in the emitter–base space-charge layer, in the quasineutral base layer, and in the base–collector space-charge layer, respectively. Therefore, Eq. (6.119) can be rewritten as

$$A\frac{d}{dt}(Q_B+Q_{BE}+Q_{BC})=i_n(0,t)-i_n(W_B,t)-A\frac{Q_B+Q_{BE}+Q_{BC}}{\tau_{nB}}. \tag{6.121}$$

Similarly, integrating the continuity equation for the excess holes over the emitter region, we obtain

$$A\frac{dQ_E}{dt}=i_p(-W_E,t)-i_p(0,t)-A\frac{Q_E}{\tau_{pE}}, \tag{6.122}$$

where

$$Q_E = q\int_{-W_E}^{0}(p_n - p_{n0})dx \tag{6.123}$$

is the excess hole charge per unit area stored in the emitter, and τ_{pE} is the hole lifetime in the emitter. From the current components illustrated in Fig. 6.21(b), the emitter current is

$$i_E(t) = i_n(0,t) + i_p(0,t) - C_{dBE,tot}\frac{dv'_{BE}(t)}{dt}, \tag{6.124}$$

where $v'_{BE} = v'_B - v'_E$ is the time-dependent intrinsic voltage across the base–emitter junction, and $C_{dBE,tot}$ is the base–emitter junction depletion-layer capacitance. The collector current is

$$i_C(t) = -i_n(W_B,t) + C_{dBC,tot}\frac{dv'_{CB}(t)}{dt}, \tag{6.125}$$

where $v'_{CB} = v'_C - v'_B$ is the time-dependent intrinsic voltage across the base–collector junction, and $C_{dBC,tot}$ is the base–collector junction depletion-layer capacitance. The base current is

$$\begin{aligned} i_B(t) &= -i_E(t) - i_C(t) \\ &= -[i_n(0,t) - i_n(W_B,t)] - i_p(0,t) \\ &\quad + C_{dBE,tot}\frac{dv'_{BE}(t)}{dt} - C_{dBC,tot}\frac{dv'_{CB}(t)}{dt}. \end{aligned} \tag{6.126}$$

Using Eq. (6.121) for $i_n(0,t) - i_n(W_B,t)$ and Eq. (6.122) for $i_p(0,t)$ in Eq. (6.126), we obtain

$$\begin{aligned} i_B(t) =& -A\frac{d(Q_B + Q_{BE} + Q_{BC})}{dt} + A\frac{dQ_E}{dt} - A\frac{Q_B + Q_{BE} + Q_{BC}}{\tau_{nB}} + A\frac{Q_B}{\tau_{pE}} \\ & - i_p(-W_E,t) + C_{dBE,tot}\frac{dv'_{BE}(t)}{dt} - C_{dBC,tot}\frac{dv'_{CB}(t)}{dt} \\ =& -\frac{d(Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot})}{dt} + \frac{dQ_{E,tot}}{dt} - \frac{Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot}}{\tau_{nB}} + \frac{Q_{E,tot}}{\tau_{pE}} \\ & - i_p(-W_E,t) + C_{dBE,tot}\frac{dv'_{BE}(t)}{dt} - C_{dBC,tot}\frac{dv'_{CB}(t)}{dt}, \end{aligned} \tag{6.127}$$

where, as explained in Section 6.4.4, we have replaced AQ_B by $Q_{B,tot}$, AQ_{BE} by $Q_{BE,tot}$, etc., to make the applicability of Eq. (6.127) not limited to a one-dimensional bipolar transistor but to a bipolar transistor of arbitrary device structure. Equation (6.127) is the charge-control model for the base current of a bipolar transistor. It states that the base current feeds the excess minority charge in the emitter and the base, the base–emitter diode depletion-layer capacitance, the base–collector diode depletion-layer capacitance, the recombination current in the base, the recombination current in the emitter, and the hole recombination current at the emitter contact.

The base-current equation can be reduced to a more useful form by noting the relationship between $v'_{CB}(t)$ and $v'_{BE}(t)$. We have

$$\begin{aligned} v'_{CB}(t) &= v'_C(t) - v'_B(t) \\ &= -[v'_B(t) - v'_E(t)] + [v'_C(t) - v'_E(t)] \\ &\approx -v'_{BE}(t) + v_{CE}(t) - i_C(t)(r_e + r_c) \\ &= -v'_{BE}(t) + V_{CC} - i_C(t)(R_L + r_e + r_c), \end{aligned} \tag{6.128}$$

where we have used Eq. (6.108) which relates v'_{CE} to v_{CE}, and the fact that $v_{CE} = V_{CC} - i_C R_L$. Therefore,

$$\frac{dv'_{CB}(t)}{dt} = -\frac{dv'_{BE}(t)}{dt} - (R_L + r_e + r_c)\frac{di_C(t)}{dt}. \tag{6.129}$$

Substituting Eq. (6.129) into Eq. (6.127), we have

$$\begin{aligned} i_B(t) =& -\frac{d(Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot})}{dt} + \frac{dQ_{E,tot}}{dt} - \frac{Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot}}{\tau_{nB}} + \frac{Q_{E,tot}}{\tau_{pE}} \\ & - i_p(-W_E,t) + (C_{dBE,tot} + C_{dBC,tot})\frac{dv'_{BE}(t)}{dt} + C_{dBC,tot}(R_L + r_e + r_c)\frac{di_C(t)}{dt} \\ =& -\frac{d(Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot})}{dt} + \frac{dQ_{E,tot}}{dt} - \frac{Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot}}{\tau_{nB}} + \frac{Q_{E,tot}}{\tau_{pE}} \\ & - i_p(-W_E,t) + \frac{(C_{dBE,tot} + C_{dBC,tot})}{g'_m}\frac{di_C(t)}{dt} + C_{dBC,tot}(R_L + r_e + r_c)\frac{di_C(t)}{dt}, \end{aligned} \tag{6.130}$$

where we have used Eq. (6.99) for the intrinsic transconductance g'_m.

In the steady state, Eq. (6.130) gives

$$I_B = \frac{I_C}{\beta_0} = \left[\frac{-(Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot})}{\tau_{nB}} + \frac{Q_{E,tot}}{\tau_{pE}}\right]_{\text{steady state}} - i_p(-W_E). \tag{6.131}$$

That is, in the steady state, the base current is simply equal to the sum of the recombination currents in the base, in the emitter, and at the emitter contact. If we assume the time dependence is quasistatic such that the steady-state relationship in Eq. (6.131) between the collector current and the sum of the recombination currents holds for the nonsteady state as well, i.e., if we assume

$$\frac{i_C(t)}{\beta_0} = \frac{-(Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot})}{\tau_{nB}} + \frac{Q_{E,tot}}{\tau_{pE}} - i_p(-W_E,t), \tag{6.132}$$

then Eq. (6.130) becomes

$$\begin{aligned} i_B(t) =& -\frac{d(Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot})}{dt} + \frac{dQ_{E,tot}}{dt} + \frac{i_C(t)}{\beta_0} \\ & + \left[\frac{C_{dBE,tot} + C_{dBC,tot}}{g'_m} + C_{dBC,tot}(R_L + r_e + r_c)\right]\frac{di_C(t)}{dt}. \end{aligned} \tag{6.133}$$

This is the differential equation relating the time-dependent base and collector currents for a transistor with a resistive load R_L (Ghandhi, 1968).

In Section 6.4.4, we made the distinction between the total stored minority charge in a transistor and the portion of minority charge capable of responding to an ac signal and contributing to the emitter diffusion capacitance. In the literature, often this distinction is not made. In that case, the first two terms in Eq. (6.133) related to the change in the total stored charge with time are replaced by the term $\tau_F di_C(t)/dt$ [see Eqs. (6.112) and (6.114)]. This is equivalent to using the charge-control method to derive the emitter diffusion capacitance, which, as discussed in Section 6.4.4 and in Appendix 6, overestimates the emitter diffusion capacitance.

In writing Eq. (6.112), we indicated that the minority charge responsible for the emitter diffusion capacitance is the sum of the *absolute* amounts coming from the various regions of the bipolar transistor. In Eq. (6.133), the minority charge $Q_{B,tot} + Q_{BE,tot} + Q_{BC,tot}$ in the base is due to electrons while the minority charge $Q_{E,tot}$ in the emitter is due to holes. Therefore, the two dQ/dt terms, including their signs and the fact that an electron has a charge $-q$ while a hole has a charge q, actually add together to give the derivative of the sum of the absolute amounts of minority charges with respect to time. In other words, the charge-control analysis automatically gives the correct summing of the minority charges in the various regions of a bipolar transistor for determining their contributions to the emitter diffusion capacitance. If it were not for the subtle difference between Q and Q_{ac}, Eq. (6.133) would have led to the correct emitter diffusion capacitance. Indeed, if we assume substituting $\tau_F di_C(t)/dt$ for the first two terms in Eq. (6.133) to be valid, then Eq. (6.133) would give the expected frequency-dependent behavior of a bipolar transistor (Ghandhi, 1968).

6.5 Breakdown Voltages

The breakdown voltages of a bipolar transistor are often characterized by applying a reverse bias across two of the three device terminals, with the third device terminal left open-circuit. These breakdown voltages are usually denoted by

BV_{EBO} = emitter–base breakdown voltage with the collector open-circuit,
BV_{CBO} = collector–base breakdown voltage with the emitter open-circuit,
BV_{CEO} = collector–emitter breakdown voltage with the base open-circuit.

Since bipolar transistors are usually operated with the emitter-base junction zero-biased or forward biased, their BV_{EBO} values are not important as long as they do not adversely affect the other device parameters. On the other hand, BV_{CBO} and BV_{CEO} must be adequately large for the intended circuit application. BV_{CBO} and BV_{CEO} are often determined, respectively, from the measured common-base and common-emitter current–voltage characteristics. The measurement setups for an n–p–n transistor, and the corresponding I–V characteristics, are illustrated schematically in Fig. 6.22.

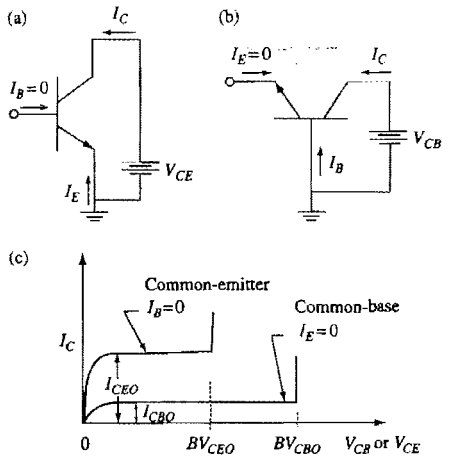


Figure 6.22. Circuit schematics for measuring (a) BV_{CEO} and (b) BV_{CBO} of an n–p–n transistor. (c) Common-emitter I_C–V_{CE} characteristics at $I_B = 0$, and common-base I_C–V_{CB} characteristics at $I_E = 0$.

6.5.1 Common-Base Current Gain in the Presence of Base–Collector Junction Avalanche

Consider an n–p–n transistor biased in the forward-active mode, as illustrated in Fig. 6.23(a). The corresponding energy-band diagram and the electron and hole current flows inside the transistor are illustrated in Fig. 6.23(b), where the locations of the emitter–base junction and the base–collector space-charge layer, where avalanche multiplication takes place, are also indicated. The emitter current I_E is equal to the sum of the hole current entering the emitter from the base and the electron current entering the base from the emitter, i.e.,

$$I_E = A_E[J_n(0) + J_p(0)], \tag{6.134}$$

where A_E is the emitter area. It should be noted that I_E, defined as the current entering the emitter, is a negative quantity for an n–p–n transistor, since both J_n and J_p are negative.

As the electrons traverse the base layer, some of them can recombine within the base layer. Only those electrons reaching $x = W_B$ contribute to the collector current. In the presence of avalanche multiplication in the reverse-biased base–collector junction, the electron current exiting the base–collector space-charge layer is a factor of M larger than that entering the space-charge layer, where M is the avalanche multiplication factor (see Section 2.5.1). That is,

$$J_n(W_B + W_{dBC}) = MJ_n(W_B). \tag{6.135}$$

The collector current I_C is equal to the electron current exiting the base–collector space-charge layer, i.e.,

$$I_C = -A_E J_n(W_B + W_{dBC}). \tag{6.136}$$

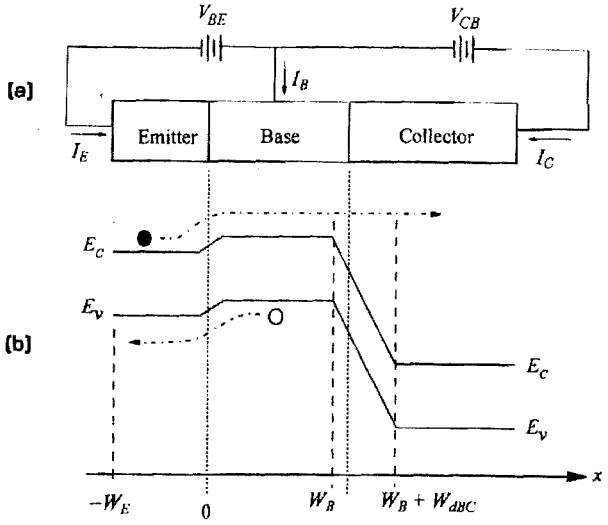


Figure 6.23. (a) Schematic illustrating the voltages and currents in an n–p–n transistor biased in the forward-active mode. (b) The corresponding energy-band diagram and illustration of the electron and hole flows inside the transistor. Also indicated are the locations of the emitter–base junction and the base–collector space-charge layer.

The minus sign in Eq. (6.136) is due to the fact that, as defined, I_C is a current entering the collector, I_C is a positive quantity for an n–p–n transistor.

Using Eqs. (6.134) to (6.136), we can rewrite the static common-base current gain α_0 [cf. Eq. (6.54)] as

$$\begin{aligned}\alpha_0 &\equiv \frac{\partial I_C}{\partial(-I_E)} \\ &= \frac{\partial J_n(W_B + W_{dBC})}{\partial\left[J_n(0) + J_p(0)\right]} \\ &= \frac{\partial J_n(0)}{\partial\left[J_n(0) + J_p(0)\right]} \frac{\partial J_n(W_B)}{\partial J_n(0)} \frac{\partial J_n(W_B + W_{dBC})}{\partial J_n(W_B)} \\ &= \gamma \alpha_T M, \end{aligned} \tag{6.137}$$

where the *emitter injection efficiency* γ is defined by

$$\gamma \equiv \frac{\partial J_n(0)}{\partial\left[J_n(0) + J_p(0)\right]} = \frac{J_n(0)}{J_n(0) + J_p(0)}, \tag{6.138}$$

and the *base transport factor* α_T is defined by

$$\alpha_T \equiv \frac{\partial J_n(W_B)}{\partial J_n(0)} = \frac{J_n(W_B)}{J_n(0)}. \tag{6.139}$$

When base–collector junction avalanche effect is negligible, we have $M \approx 1$, and the common-base current gain is

$$\alpha_0 = \gamma \alpha_T \quad (\text{when } M = 1). \tag{6.140}$$

If we further assume that recombination in the thin base is negligible (see Exercise 6.6), then the common-base current gain is simply

$$\alpha_0 = \gamma = \frac{\partial J_n(0)}{\partial\left[J_n(0) + J_p(0)\right]} \quad (\text{when } M = 1 \text{ and } \alpha_T = 1). \tag{6.141}$$

[*Note*: Throughout this chapter, by equating the collector current to the electron current entering the intrinsic base, i.e., Eq. (6.31), and by equating the base current to the hole current entering the emitter, i.e., Eq. (6.41), we have implicitly made the assumptions that $M = 1$ and $\alpha_T = 1$. That is, we have implicitly assumed that $\alpha_0 = \gamma$.]

6.5.2 Saturation Currents in a Transistor

If we define I_{EBO} and I_{CBO} by (Ebers and Moll, 1954)

$$I_{EBO} \equiv I_{F0}(1 - \alpha_R \alpha_F) \tag{6.142}$$

and

$$I_{CBO} \equiv I_{R0}(1 - \alpha_R \alpha_F), \tag{6.143}$$

then Eqs. (6.88) and (6.89) give

$$I_E = -I_{EBO}[\exp(qV_{BE}/kT) - 1] - \alpha_R I_C \tag{6.144}$$

and

$$I_C = -I_{CBO}[\exp(qV_{BC}/kT) - 1] - \alpha_F I_E. \tag{6.145}$$

The physical meaning of I_{EBO} and I_{CBO} is apparent from these equations. I_{EBO} is the saturation current of the emitter–base diode when the collector is open-circuit, i.e., it is the emitter current when the emitter–base diode is reverse biased and $I_C = 0$. This is the current one measures in measuring BV_{EBO}. Similarly, I_{CBO} is the saturation current of the collector–base diode when the emitter is open-circuit, i.e., it is the collector current when the base–collector diode is reverse biased and $I_E = 0$. This is the current one measures in measuring BV_{CBO}. I_{CBO} is indicated in Fig. 6.22(c).

Let us apply Eq. (6.145) to the BV_{CEO} measurement setup shown in Fig. 6.22(a). We note that when V_{CE} is near BV_{CEO}, the collector–base diode is reverse biased. Also, at $I_B = 0$, $I_C = -I_E$. Therefore, Eq. (6.145) gives, for the common-emitter configuration with the base–collector junction reverse biased and at $I_B = 0$,

$$I_C = \frac{I_{CBO}}{1 - \alpha_F}. \tag{6.146}$$

This is the saturation current in the common-emitter configuration. We shall denote this current by I_{CEO}, i.e.,

$$I_{CEO} = \frac{I_{CBO}}{1 - \alpha_0}, \tag{6.147}$$

where we have used the fact that $\alpha_F = \alpha_0$. This current is also indicated in Fig. 6.22(c). It is clear from Eq. (6.147) that I_{CEO} is significantly larger than I_{CBO}, since α_0 is usually less than but close to unity. This is indicated in Fig. 6.22(c).

6.5.3 Relation Between BV_{CEO} and BV_{CBO}

As pointed out in Section 2.5.1, the breakdown voltages in VLSI devices are usually determined experimentally, rather than calculated from some model. The avalanche multiplication factor M in a reverse-biased diode is often expressed in terms of its breakdown voltage BV using the *empirical* formula (Miller, 1955)

$$M(V) = \frac{1}{1 - (V/BV)^m}, \tag{6.148}$$

where V is the reverse-bias voltage and m is a number between 3 and 6 depending on the material and its resistivity. Thus, for the reverse-biased collector–base diode, we have

$$M(V_{CB}) = \frac{1}{1 - (V_{CB}/BV_{CBO})^m}. \tag{6.149}$$

Equation (6.147) implies that I_{CEO} becomes infinite when $\alpha_0 = 1$. From Eq. (6.137), this means that when the collector voltage reaches BV_{CEO},

$$\gamma\alpha_T M(V_{CB}) = \gamma\alpha_T M(BV_{CEO}) = 1. \tag{6.150}$$

Equations (6.149) and (6.150) give

$$\frac{BV_{CEO}}{BV_{CBO}} = (1 - \gamma\alpha_T)^{1/m}. \tag{6.151}$$

Since $1 - \gamma\alpha_T \ll 1$, Eq. (6.151) indicates that BV_{CEO} can be substantially smaller than BV_{CBO}. This is illustrated in Fig. 6.22(c). Another way of comparing these breakdown voltages is to note that it takes M approaching infinity to cause collector–base breakdown, while it takes M only slightly larger than unity to cause collector–emitter breakdown (see Exercise 6.7). From Eq. (6.140), $\gamma\alpha_T = \alpha_0(M = 1) = \beta_0/(1 + \beta_0)$, where we have used Eq. (6.55) and β_0 is the current gain at negligible collector–base junction avalanche. Thus, Eq. (6.151) can also be written as

$$\frac{BV_{CEO}}{BV_{CBO}} = \left(\frac{1}{1 + \beta_0}\right)^{1/m} \approx \left(\frac{1}{\beta_0}\right)^{1/m}. \tag{6.152}$$

Equation (6.152) shows that ***there is a tradeoff between the collector–emitter breakdown voltage and the current gain*** of a transistor.

It should be noted that the relationship between BV_{CEO} and BV_{CBO} in Eq. (6.152) is valid only when collector–base junction breakdown is governed by the intrinsic-base–collector diode, and not the extrinsic-base–collector diode. This may or may not be the case in a typical transistor, depending on the device structure and the fabrication process employed. (The BV_{CBO} of a modern bipolar transistor, with its extrinsic base formed independently of the intrinsic base and its collector optimized for minimal capacitance, is usually determined by the intrinsic-base–collector diode rather than the extrinsic-base–collector diode. The design and characteristics of modern bipolar transistors are covered in Chapter 7.) Figure 6.24 is a plot of BV_{CEO} versus BV_{CBO} based on data reported in recent literature for n–p–n transistors. It shows that for modern n–p–n transistors, BV_{CEO} is typically a factor of 2 to 4 smaller than BV_{CBO}.

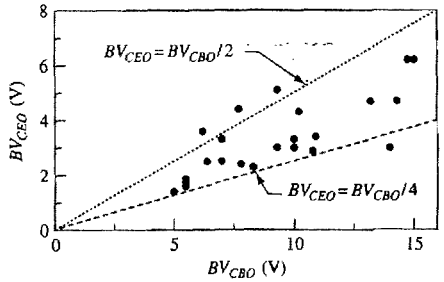


Figure 6.24. Reported BV_{CEO} versus BV_{CBO} data for recently published n–p–n transistors.

Excercises

6.1 The electric field in an n-type semiconductor is given by Eq. (6.20), i.e.

$$\mathscr{E}(\text{n-region}) = -\frac{kT}{q}\left(\frac{1}{n_n}\frac{dn_n}{dx} - \frac{1}{n_{ie}^2}\frac{dn_{ie}^2}{dx}\right).$$

Derive this equation, stating clearly the approximations made in the derivation.

6.2 The hole current density in the n-side of a p–n diode is given by Eq. (6.26), i.e.,

$$J_p(x) = -qD_p\frac{n_{ie}^2}{n_n}\frac{d}{dx}\left(\frac{n_n p_n}{n_{ie}^2}\right).$$

Derive this equation, stating clearly the approximations made in the derivation.

6.3 For a polysilicon emitter with the emitter–base junction located at $x = 0$ and the silicon–polysilicon interface located at $x = -W_E$, the emitter Gummel number is given by Eq. (6.46), namely

$$G_E = N_E\left(\frac{n_i^2}{n_{ieE}^2}\right)\left(\frac{W_E}{D_{pE}} + \frac{1}{S_p}\right).$$

One model (Ning and Isaac, 1980) for relating S_p to the properties of the polysilicon layer is to assume that there is no interfacial oxide, so that the transport of holes

through the interface is simply determined by the properties of the polysilicon layer. Let W_{E1} be the thickness of the polysilicon layer, and let D_{pE1} and L_{pE1} be the hole diffusion coefficient and hole diffusion length, respectively, in the polysilicon. Assume an ohmic metal–polysilicon contact.

(a) Let $\Delta p_n(-W_E)$ be the excess hole concentration at the polysilicon–silicon interface, and let x' denote the distance from the polysilicon–silicon interface, i.e., $x' = -(x + W_E)$. Show that the excess hole distribution in the polysilicon layer, $\Delta p_n(x')$, is given by [cf. Eq. (2.119)]

$$\Delta p_n(x') = \Delta p_n(-W_E)\frac{\sinh\left[(W_{E1} - x')/L_{pE1}\right]}{\sinh\left(W_{E1}/L_{pE1}\right)}.$$

(b) The relationship between S_p and the hole current density entering the polysilicon layer is given by Eq. (6.36). Show that

$$S_p = \frac{D_{pE1}}{L_{pE1}\tanh\left(W_{E1}/L_{pE1}\right)}.$$

6.4 Consider an n–p–n transistor with negligible parasitic resistances (which will be included in Exercise 6.5). Equations (6.144) and (6.145) give

$$I_E = -I_{EBO}\left[\exp\left(qV'_{BE}/kT\right) - 1\right] - \alpha_R I_C$$

and

$$I_C = -I_{CBO}\left[\exp\left(qV'_{BC}/kT\right) - 1\right] - \alpha_F I_E,$$

where V'_{BE} and V'_{BC} are the internal base–emitter and base–collector junction bias voltages. If the transistor is operated in saturation, i.e., both V'_{BE} and V'_{BC} are positive, show that the internal collector–emitter voltage, $V'_{CE} = V'_C - V'_E$, is related to the currents by

$$V'_{CE} = \frac{kT}{q}\ln\left[\frac{\alpha_F(I_{EBO} - I_E - \alpha_R I_C)}{\alpha_R(I_{CBO} - I_C - \alpha_F I_E)}\right].$$

[*Hint*: Use Eqs. (6.90), (6.142), and (6.143) to show that $I_{CBO}/I_{EBO} = \alpha_F/\alpha_R$.]

6.5 From the expression for V'_{CE} in Exercise 6.4, show that if the emitter and collector series resistances r_e and r_c are included, and if the saturation currents I_{EBO} and I_{CBO} are negligible, the voltage drop across the collector and emitter terminals, $V_{CE} = V_C - V_E$, is given by

$$V_{CE} = \frac{kT}{q}\ln\left[\frac{I_B + I_C(1 - \alpha_R)}{\alpha_R[I_B - I_C(1 - \alpha_F)/\alpha_F]}\right] + r_e(I_B + I_C) + r_c I_C$$

and that for open-circuit collector

$$V_{CE}(I_C = 0) = \frac{kT}{q}\ln\left(\frac{1}{\alpha_R}\right) + r_c I_B.$$

[The emitter resistance r_e is often determined from a plot of the saturation open-collector voltage, $V_{CE}(I_C = 0)$, as a function of I_B (Ebers and Moll, 1954; Filensky and Beneking, 1981). The collector resistance r_c can be determined in a similar way by interchanging the emitter and collector connections.]

6.6 For an n–p–n transistor, the base transport factor α_T is given in Eq. (6.139), i.e.,

$$\alpha_T \equiv \frac{J_n(x = W_B)}{J_n(x = 0)},$$

where the intrinsic-base layer is located between $x = 0$ and $x = W_B$. For a uniformly doped base, the excess-electron distribution is given by Eq. (2.119), namely

$$n_p - n_{p0} = n_{p0}[\exp(qV_{BE}/kT) - 1]\frac{\sinh[(W_B - x)/L_{nB}]}{\sinh(W_B/L_{nB})}.$$

If the electron current in the base is due to diffusion current only, show that

$$\alpha_T = \left(\cosh\frac{W_B}{L_{nB}}\right)^{-1}.$$

Use Fig. 2.24(c) to estimate α_T for a uniformly doped base with $N_B = 1 \times 10^{18}\ \text{cm}^{-3}$ and $W_B = 100$ nm, and show that our assumption of negligible recombination in the intrinsic base is justified.

6.7 If M is the avalanche multiplication factor for the base–collector junction, and β_0 is the common-emitter current gain at negligible base–collector junction avalanche, show that the collector–emitter breakdown occurs when

$$M - 1 = \frac{1}{\beta_0}.$$

[*Hint*: Use Eqs. (6.140) and (6.150).] (It is interesting to note that since β_0 is typically about 100, collector–emitter breakdown occurs when M is only slightly larger than unity. That is, it does not take much base–collector junction avalanche to cause collector–emitter breakdown.)

7 Bipolar Device Design

Bipolar device design can be considered in two parts. The first part deals with designing bipolar transistors in general, independent of their intended application. In this case, the goal is to reduce as much as possible, consistent with the start-of-the-art fabrication technology, all the internal resistance and capacitance components of the transistor. The second part deals with designing a bipolar transistor for a specific circuit application. In this case, the optimal device design point depends on the application. The design of a bipolar transistor in general is covered in this chapter, and the optimization of a transistor for a specific application is discussed in Chapter 8.

7.1 Design of the Emitter Region

It was shown in Section 6.2 that the emitter parameters affect only the base current, and have no effect on the collector current. In theory, a device designer can vary the emitter design to vary the base current. In practice, this is rarely done, for two reasons. First, for digital-circuit applications, as long as the current gain is not unusually low or the base current unusually high, the performance of a bipolar transistor is rather insensitive to its base current (Ning *et al.*, 1981). For many analog-circuit applications, once the current gain is adequate, the reproducibility of the base current is more important than its magnitude. Therefore, there is really no particular reason to tune the base current of a bipolar device by tuning the emitter design, once a low and reproducible base current is obtained. Second, as can be seen in Appendix 2, the emitter is formed towards the end of the device fabrication process. Any change to the emitter process to tune the base current could affect the doping profile of the other device regions and hence could affect the other device parameters. As a result, once a bipolar technology is ready for manufacturing, its emitter fabrication process is usually fixed. All that a device designer can do to alter the device and circuit characteristics in this bipolar technology is to change the base and the collector designs, which often can be accomplished independently of the emitter process and hence has no effect on the base current.

The objective in designing the emitter of a bipolar transistor is then to achieve a low but reproducible base current while at the same time minimizing the emitter series resistance. As illustrated in Fig. 6.2, the commonly used bipolar transistors have either a diffused (or implanted-and-diffused) emitter or a polysilicon emitter. The design of both types of emitters is discussed in this section.

7.1.1 Diffused or Implanted-and-Diffused Emitter

A diffused or implanted-and-diffused emitter is formed by predoping a surface region of the silicon above the intrinsic base and then thermally diffusing the dopant to a desired depth. As shown in Eq. (6.48), for a diffused emitter, the base current is inversely proportional to the emitter doping concentration. Therefore, to minimize both the base current and the emitter series resistance, a diffused emitter is usually doped as heavily as possible. For n–p–n transistors, arsenic, instead of phosphorus, is usually used as the dopant, because arsenic gives a more abrupt doping profile than phosphorus. A more abrupt emitter doping profile leads to a shallower emitter junction, and, as we shall see later, a shallow emitter junction is needed for achieving a thin intrinsic base. Also, a shallower emitter has a smaller vertical junction area and associated capacitance. A diffused emitter typically has a peak doping concentration of about 2×10^{20} cm^{-3}, as indicated in Fig. 6.2(a).

A diffused emitter is contacted either directly by a metal, or by a metal via a metal silicide layer. Commonly used silicides for emitter contact include platinum silicide and titanium silicide. If the fabrication process leaves negligible residual oxide on the emitter prior to contact formation, the resultant contact resistivity, as discussed in Section 2.4.4, is a function of the metal or metal silicide used, as well as a function of the emitter doping concentration at the contact. For a doping concentration of 2×10^{20} cm^{-3} at the contact, a specific contact resistivity of about $(1–2) \times 10^{-7}$ Ω-cm^2 should be achievable.

Using the resistivity values of silicon shown in Fig. 2.9, the specific series resistivity of a 0.5-μm-deep silicon region, with an averaged doping concentration of 1×10^{20} cm^{-3}, is about 4×10^{-8} Ω-cm^2. Therefore, the series resistance of a diffused emitter is dominated by its metal–silicon contact resistance; the series resistance of the doped-silicon region itself is negligible in comparison. For a diffused emitter of 1 μm^2 in area, the emitter series resistance is typically about 10–20 Ω.

It can be inferred from Fig. 6.1(b) that the intrinsic-base width W_B is related to the emitter junction depth x_{jE} and the base junction depth x_{jB} by

$$W_B = x_{jB} - x_{jE}. \tag{7.1}$$

As we shall see in Section 7.2, one of the objectives in the design of the intrinsic base is to minimize its width. For W_B to be well controlled, reproducible, and thin, x_{jE} should be as small as possible. If x_{jE} is much larger than W_B, then W_B is given by the difference of two large numbers and hence will have large fluctuation.

Commonly used metal silicides are formed by depositing a layer of the appropriate metal on the silicon surface and then reacting the metal with the underlying silicon to form silicide. The emitter width W_E is therefore reduced when metal silicide is used for emitter contact, because silicon in the emitter is consumed in the metal silicide formation process. As shown in Section 6.2.2, once W_E is less than the minority-carrier diffusion length, the base current increases as $1/W_E$. As a result, the base current, and hence the current gain, of a bipolar transistor with a shallow diffused emitter varies with the emitter contact process (Ning and Isaac, 1980).

Referring to the minority-carrier diffusion lengths shown in Fig. 2.24(c), we see that the junction depth of a diffused n-type emitter should be larger than 0.3 μm in order to

have adequately controllable and reproducible base-current characteristics. ***Diffused emitters are therefore not suitable for base widths of less than 100 nm.***

7.1.2 Polysilicon Emitter

Practically all modern high-performance bipolar transistors with base widths of 100 nm or smaller employ a polysilicon emitter. In this case, the emitter is formed by doping a polysilicon layer heavily and then activating the doped polysilicon layer just sufficiently to obtain reproducible base current and low emitter series resistance. The emitter junction depth, measured from the silicon–polysilicon interface, can be as small as 25 nm (Warnock, 1995). Consequently, with polysilicon-emitter technology, base widths of 50 nm or less can be obtained. The polysilicon-emitter process recipes are usually considered proprietary. However, there is a vast amount of literature on the physics of polysilicon-emitter devices (Ashburn, 1988; Kapoor and Roulston, 1989). Interested readers are referred to these publications.

The base current of a polysilicon-emitter transistor is given by Eqs. (6.42) to (6.44), with a surface recombination velocity, S_p, appropriate for the particular process used for forming the polysilicon emitter. The S_p is usually used as a fitting parameter to the measured base current. In general, the base current of a polysilicon-emitter transistor is sufficiently low so that current gains in excess of 100 are readily achievable.

As will be shown in Chapter 8, the maximum speed of a modern bipolar transistor is determined primarily by its diffusion capacitance. In Section 2.2.6, the diffusion capacitance due to minority-carrier storage in the emitter was shown to be small compared to that due to minority-carrier storage in the base. Therefore, the maximum speed of a bipolar transistor is relatively insensitive to the emitter component of its diffusion capacitance. In other words, as long as the desired base-region characteristics are obtained, the details of the emitter region have relatively little effect on the maximum speed of a bipolar transistor (Ning *et al.*, 1981). Nonetheless, a polysilicon-emitter fabrication process should be designed to give low emitter series resistance and adequate emitter–base breakdown voltage, as well as the desired base-region characteristics.

The series resistance of a polysilicon emitter includes the polysilicon–silicon contact resistance, resistance of the polysilicon layer, and resistance of the metal–polysilicon contact. The specific resistivity of a metal–polysilicon contact is about the same as that of a metal–silicon contact. For arsenic-doped polysilicon emitters, the reported specific silicide–polysilicon contact resistivity is typically (2–6) $\times 10^{-7}$ Ω-cm^2, depending on the arsenic concentration (Iinuma *et al.*, 1995). It is large compared to the series resistance of the polysilicon layer itself. The polysilicon–silicon contact resistance, on the other hand, is a strong function of the polysilicon-emitter fabrication process and can vary by large amounts (Chor *et al.*, 1985). In fact, polysilicon-emitter technology is still an area of active development. The recently published data (Iinuma *et al.*, 1995; Uchino *et al.*, 1995; Kondo *et al.*, 1995; Shiba *et al.*, 1996) suggest that a total emitter specific resistivity, which includes contributions from both the polysilicon–silicon interface and the metal–polysilicon contact, of 7–50 Ω-μm^2 should be obtainable (see Exercise 7.7).

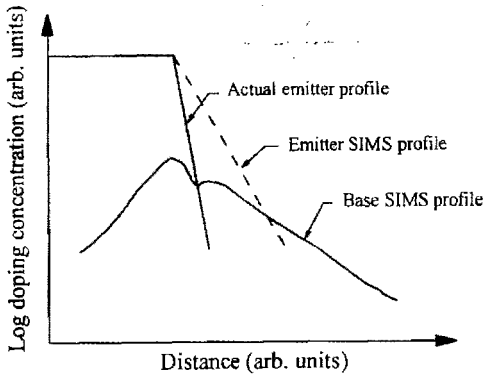


Figure 7.1. Schematic illustrating the measured SIMS doping profiles of the emitter and base of a modern n–p–n transistor. The measured emitter SIMS profile is usually less abrupt than the real one.

The small junction depth of a polysilicon emitter implies a relatively small perimeter, or vertical, extrinsic-base–emitter junction area. The total emitter–base junction capacitance of a polysilicon emitter is therefore much smaller than that of a diffused emitter. For a 0.3-μm emitter stripe, the total emitter–base junction capacitance of a polysilicon emitter can be less than $\frac{1}{3}$ of that of a diffused emitter.

It should be pointed out that the junction of a polysilicon emitter is so shallow that the commonly used secondary-ion mass spectroscopy (SIMS) technique for measuring dopant concentration profiles often indicates an emitter junction deeper than it really is. The real emitter junction depth can be obtained from the p-type base SIMS profile, which shows a dip where the n-type and p-type doping concentrations are equal (Hu and Schmidt, 1968). This is illustrated schematically in Fig. 7.1.

7.2 Design of the Base Region

It was shown in Section 6.2 that the base-region parameters affect only the collector current, not the base current. The base current is determined by the emitter parameters. It has been demonstrated experimentally (Ning *et al.*, 1981), and will be discussed in Chapter 8, that the performance of a bipolar circuit is determined primarily by the collector current, not the base current, at least for circuits where the bipolar transistors do not saturate. Therefore, as long as current gain is adequate, which is the case with a polysilicon emitter, the focus in designing or optimizing a bipolar transistor should be on the collector current, and not on the base current. In other words, the focus should be on the intrinsic base when there is negligible base widening, and on both the intrinsic base and the collector when base widening is not negligible.

The design of the base of a bipolar transistor can be very complex, because of the tradeoffs that must be made between the ac and dc characteristics, which depend on the intended application, and because of the tradeoffs that must be made between the desired

device characteristics and the complexity of the fabrication process for realizing the design. In this section, the relationship between the physical and electrical parameters of the base is derived, and the design tradeoffs are discussed. Optimization of the base design for various circuit applications will be covered in Chapter 8.

Referring to Fig. 6.12, we can divide the base region into two parts. The part directly underneath the emitter is the intrinsic base, and the part connecting the intrinsic base to the base terminal is the extrinsic base. As a first-order but good approximation, the intrinsic base is what determines the collector current characteristics, and hence the intrinsic performance of a transistor. The discussions and the collector current characteristics derived in Chapter 6 are all for the intrinsic base. Effects of the extrinsic base were ignored.

The extrinsic base is an integral part of any bipolar transistor. It is a parasitic component in that it does not contribute appreciably to the collector current, at least for properly designed transistors. In general, designing the extrinsic base is very simple: the extrinsic-base area and its associated capacitance and series resistance should all be as small as possible. How this is accomplished depends on the fabrication process used. A major focus in bipolar-technology research and development has been to minimize the parasitic resistance and capacitance associated with the extrinsic base. The interested reader is referred to the vast literature on the subject (Warnock, 1995; Nakamura and Nishizawa, 1995; Asbeck and Nakamura, 2001; and the references therein), and to Appendix 2, which outlines the fabrication process for one of the most widely used modern bipolar transistors.

Any adverse effect of the extrinsic base on the breakdown voltages of the emitter–base and base–collector diodes should be minimized. This is accomplished by having the dopant distribution of the extrinsic base not extending appreciably into the intrinsic base. If the extrinsic base encroaches appreciably on the intrinsic base, the encroached-on intrinsic-base region will appear to be wider, as well as more heavily doped, than the rest of the intrinsic base. Extrinsic-base encroachment on the intrinsic base, therefore, will lead to a smaller collector current as well as degraded dc and ac characteristics (Lu *et al.*, 1987; Li *et al.*, 1987).

For an optimally designed bipolar process, extrinsic-base encroachment is usually negligible. As a result, the extrinsic base usually has little effect on the collector current. Therefore, only the design of the intrinsic base will be discussed further in this section. We first consider the design of a Si-base in this section. The design of a SiGe-base is covered in Section 7.4.

7.2.1 Relationship between Base Sheet Resistivity and Collector Current Density

As shown in Fig. 6.5, the collector current of a typical bipolar transistor is ideal, i.e., varying as $\exp(qV_{BE}/kT)$, for V_{BE} less than about 0.9 V. For this ideal region, the saturated collector current density for an n–p–n transistor is given by Eq. (6.33), which is repeated here:

$$J_{C0} = \frac{q}{\int_0^{W_B} \frac{p_p(x)}{D_{nB}(x)n_{ieB}^2(x)}dx}, \tag{7.2}$$

where p_p, D_{nB}, and n_{ieB} are the hole density, the electron diffusion coefficient, and the effective intrinsic-carrier concentration, respectively, in the p-type base region. The effective intrinsic-carrier concentration is given by Eq. (6.14). It can be used to allow for heavy-doping effect as well as any bandgap-engineering effect by properly adjusting the bandgap-narrowing parameter ΔE_g. We shall first consider the case where n_{ieB} is used to allow for heavy-doping effect in the base. The case of using n_{ieB} to allow for base-bandgap engineering will be covered in Section 7.4.

For device design purposes, it is often convenient to assume that both D_{nB} and n_{ieB} are slowly varying functions of x and hence can be approximated by some average values. That is, Eq. (7.2) is often written as

$$J_{C0} \approx \frac{q\bar{D}_{nB}\bar{n}_{ieB}^2}{\int_0^{W_B} p_p(x)dx}. \tag{7.3}$$

At low currents, the hole concentration in the base is equal to the base doping concentration $N_B(x)$, and Eq. (7.3) can be further simplified to

$$J_{C0} \approx \frac{q\bar{D}_{nB}\bar{n}_{ieB}^2}{\int_0^{W_B} N_B(x)dx}. \tag{7.4}$$

The integral in the denominator of Eq. (7.4) is simply the total *integrated base dose*. [In the literature, the denominator in Eq. (7.4) is sometimes referred to as the base Gummel number (Gummel, 1961). However, in this book we follow the convention of de Graaff (de Graaff *et al.*, 1977), where the base Gummel number G_B is defined by Eq. (6.34).] Thus, the collector current density at low currents is approximately inversely proportional to the total integrated base dose. Using ion-implantation techniques for doping the intrinsic base, the integrated base dose, and hence the collector current density, can be controlled quite precisely and reproducibly.

The sheet resistivity of the intrinsic base, R_{Sbi}, is

$$R_{Sbi} = \left(q\int_0^{W_B} p_p(x)\mu_p(x)dx\right)^{-1}. \tag{7.5}$$

Again, for device design purposes, it is convenient to assume an average mobility and rewrite Eq. (7.5) as

$$R_{Sbi} \approx \left(q\bar{\mu}_p\int_0^{W_B} p_p(x)dx\right)^{-1}. \tag{7.6}$$

Substituting Eq. (7.6) into Eq. (7.3), we obtain

$$J_{C0} \approx q^2\bar{D}_{nB}\bar{\mu}_p\bar{n}_{ieB}^2R_{Sbi}. \tag{7.7}$$

That is, ***the collector current density is approximately proportional to the intrinsic-base sheet resistivity***. This direct correlation is valid for R_{Sbi} between 500 and 20×10^3 $\Omega/\square$, which is the range of interest in most bipolar device designs (Tang, 1980).

7.2.2 Intrinsic-Base Dopant Distribution

For a desired intrinsic-base sheet resistivity, the detailed intrinsic-base dopant distribution depends on the fabrication process used. Most modern bipolar-transistor processes employ ion implantation, followed by thermal annealing and/or thermal diffusion, to form the intrinsic base. In this case, the intrinsic-base doping profile is approximately a *Gaussian* distribution, often with an exponentially decreasing tail. As will be shown in Section 7.3, the collector doping concentration of a modern bipolar transistor is relatively high, often in excess of 1×10^{17} cm^{-3}. This concentration is usually high compared with the tail of the base dopant distribution. As a result, the lightly doped tail is often clipped off by the collector doping profile and has little effect on the collector current. Therefore, for simplicity of discussion and analysis, we shall ignore the tail of the base dopant distribution.

If the Gaussian base dopant distribution peaks at the emitter–base junction located at $x = 0$, then the base doping concentration can be described by

$$N_B(x) = N_{B\max} \exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{7.8}$$

where σ and $N_{B\max}$ are the standard deviation and peak concentration, respectively, of the distribution. For most bipolar device designs, the peak doping concentration in the base is approximately 10–100 times that in the collector. Here, for purposes of discussion, we assume $N_{B\max}/N_C = 100$. This implies a base width of

$$W_B \approx 3\sigma \tag{7.9}$$

for the Gaussian base dopant distribution.

With the advent of silicon epitaxy processes, instead of implanting dopant ions into silicon, the intrinsic base can be formed by epitaxial growth of a thin, *in situ* doped silicon layer on top of the collector. In this case, the base dopant distribution depends on the *in situ* doping process used. The simplest distribution is an approximately uniform, or *boxlike*, distribution.

Figure 7.2 illustrates a box profile of $N_B = 1 \times 10^{18}$ cm^{-3}, and a Gaussian profile of the same integrated base dose and the same base width. It shows that the peak concentration of the Gaussian profile is more than twice that of the box profile. The emitter–base depletion-layer capacitance of a Gaussian base doping profile is therefore larger than that of a boxlike base doping profile. Also, with a higher base doping concentration at the emitter–base junction, high-field effects at the emitter–base junction are also more severe for a Gaussian-profile base than for a box-profile base.

In general, for the same base width and integrated base dose, a base doping profile with a higher peak concentration, at or close to the emitter–base junction, will lead to a larger emitter–base junction capacitance. However, this does not imply that a box-profile base is necessarily preferred over a base with a peak concentration located at or near the emitter–base junction, for there are many other factors or parameters, such as base transit time and ease of fabrication, that must also be considered. (A boxlike base doping profile will certainly lead to a larger Early voltage, as will be shown in Section 8.5.1.) Before we

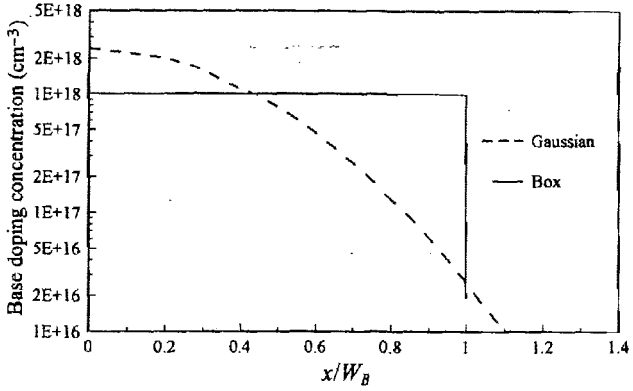


Figure 7.2. Schematic illustration of a boxlike doping profile and a Gaussian doping profile for the same base width and the same integrated base dose. The peak doping concentration of the Gaussian profile is approximately 2.4 times that of the box profile, and the base width is approximately equal to 3σ.

discuss the dependence of the base transit time on the physical parameters of the base, we need to discuss the electric field in a quasineutral base region, since the transport of minority carriers in the base depends on the electric field in it.

7.2.3 Electric Field in the Quasineutral Intrinsic Base

The electron current density in the base region of an n–p–n transistor is given by Eq. (6.25). Since this is also the collector current density, we can write

$$J_C = qD_{nB}\frac{n_{ieB}^2}{p_p}\frac{d}{dx}\left(\frac{n_p p_p}{n_{ieB}^2}\right), \tag{7.10}$$

where the subscript B has been added to indicate that the parameters are for the base region. It is valid for arbitrary base doping profile and arbitrary bandgap variation in the base. The dependence on bandgap variation is implicit in the effective intrinsic-carrier concentration, which is given by Eq. (6.14), i.e.,

$$n_{ieB}^2 = n_i^2 \exp(\Delta E_{gB}/kT), \tag{7.11}$$

where ΔE_{gB} is the bandgap-narrowing parameter in the base.

The electric field in the base region can be derived by decomposing Eq. (7.10) into the more familiar drift and diffusion components. To this end, Eq. (7.10) can be rewritten as

$$J_C = qD_{nB}n_p\left(\frac{1}{p_p}\frac{dp_p}{dx} - \frac{1}{n_{ieB}^2}\frac{dn_{ieB}^2}{dx}\right) + qD_{nB}\frac{dn_p}{dx}. \tag{7.12}$$

From Eq. (6.5), the collector current density can also be written in its usual form of

$$J_C = qn_p\mu_{nB}\mathscr{E} + qD_{nB}\frac{dn_p}{dx}. \tag{7.13}$$

Comparison of Eqs. (7.12) and (7.13) shows that the electric field for minority-carrier electrons in the p-type base region is given by

$$\mathscr{E}(\text{p-base}) = \frac{kT}{q}\left(\frac{1}{p_p}\frac{dp_p}{dx} - \frac{1}{n_{ieB}^2}\frac{dn_{ieB}^2}{dx}\right). \tag{7.14}$$

This dependence of the electric field on majority-carrier concentration and effective intrinsic-carrier concentration was stated without derivation in Eq. (6.19). Using Eq. (7.11), Eq. (7.14) can also be written as

$$\mathscr{E}(\text{p-base}) = \frac{kT}{q}\frac{1}{p_p}\frac{dp_p}{dx} - \frac{1}{q}\frac{d\Delta E_{gB}}{dx}, \tag{7.15}$$

which relates the electric field to the majority-carrier concentration and bandgap narrowing in the base region.

7.2.3.1 Electric Field in the Quasineutral Base Region at Low Currents

At low injection currents, $p_p \simeq N_B$, and Eq. (7.14), or Eq. (7.15), gives the built-in electric field in the base caused by the base doping profile and base-bandgap variation. It should be noted that as N_B increases, ΔE_{gB} increases, and hence dN_B/dx and $d\Delta E_{gB}/dx$ have the same sign. Equation (7.15) shows that ***the electric field due to the heavy-doping effect always tends to offset the electric field due to the dopant distribution***.

When transistors are designed with base widths much larger than 100 nm, the peak base doping concentration is usually about 10^{17} cm^{-3} or smaller. For such low concentrations, the effect of heavy doping is negligible. In this case, a graded base doping profile can result in a substantial electric field in the base, which enhances the drift component of the collector current traversing the base layer. These are so-called *drift transistors*. They have higher cutoff frequencies than transistors with a more uniform base doping profile (Sze, 1981; Ghandhi, 1968).

Modern bipolar transistors, however, have peak base doping concentrations larger than 10^{18} cm^{-3}, as indicated in Fig. 6.2. For these devices, the electric field due to the $d\Delta E_{gB}/dx$ term must be included, which could substantially cancel the electric field due to the dN_B/dx term. As a result, the net electric field in the quasineutral intrinsic base of a modern bipolar transistor can be relatively small. In other words, ***the drift-transistor concept is less important in modern thin-base bipolar device design than in the design of wide-base bipolar transistors***. (This will be demonstrated below for a Gaussian-profile base design.)

- *Electric field in an intrinsic base with a box profile.* For a boxlike base doping profile, both dN_B/dx and $d\Delta E_{gB}/dx$ are equal to zero. (Here we assume ΔE_{gB} is due to heavy doping alone. ΔE_{gB} due to base bandgap grading will be covered in Section 7.4.) There is no electric field in a box-profile base region at low injection currents.

- *Electric field in an intrinsic base with a Gaussian doping profile.* For the Gaussian base profile given by Eq. (7.8), the electric field at low currents is

$$\mathscr{E}(\text{Gaussian-base}) = \left(-\frac{kT}{q}\frac{x}{\sigma^2}\right) + \left(-\frac{1}{q}\frac{d\Delta E_{gB}}{dx}\right). \tag{7.16}$$

The first term in Eq. (7.16) is negative. The Gaussian base doping profile, with its concentration larger near the emitter-base junction and lower towards the base–collector junction, has a *graded-base* electric field in a direction to drive the electrons across the base layer. However, the second term in Eq. (7.16) is positive, since ΔE_{gB} is larger near the emitter–base junction, where the base doping concentration is large, than near the base–collector junction, where the base doping concentration is very small.

For the Gaussian base profile shown in Fig. 7.2, the electric field components as well as the total electric field at low injection levels, i.e., for $p_p \simeq N_B$, are shown in Fig. 7.3. The bandgap-narrowing parameter given by Eq. (6.17) is used, and the base width is assumed to be 100 nm. Figure 7.3 shows that, for this specific Gaussian base doping profile, the effect of heavy doping almost completely offsets the effect of nonuniform dopant distribution, except for the region near the base–collector junction, where the base doping concentration is relatively small and hence the effect of heavy doping is negligible. This lightly doped base region near the base–collector junction is most likely depleted in normal device operation and hence does not form part of the quasineutral base. Therefore, the net electric field in the entire quasineutral base region is quite negligible.

For non-Gaussian base doping profiles, the cancellation of the electric field components may not be as complete as suggested in Fig. 7.3. Nonetheless, the cancellation is substantial. The reader is referred to the literature (Suzuki, 1991) for more examples of similar calculations.

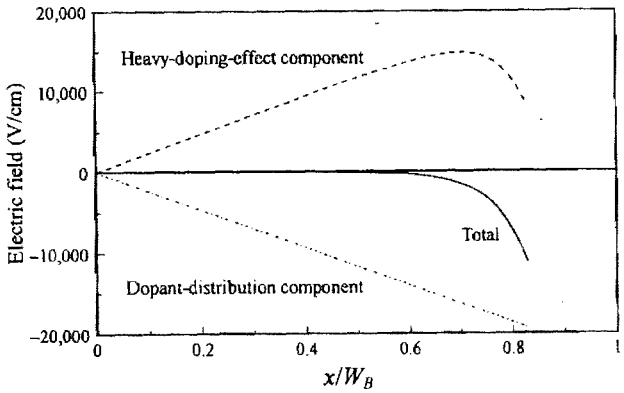


Figure 7.3. Electric fields in the quasineutral intrinsic-base region with a Gaussian doping profile. The total electric field is the sum of the dopant-distribution and the bandgap-narrowing components. The Gaussian-profile parameters are $\sigma = W_B/3$, $W_B = 100$ nm, and $N_{B\max} = 2.4 \times 10^{18}$ cm^{-3}.

7.2.4 Base Transit Time

The base transit time t_B is an important and often used figure of merit for bipolar transistors. It was shown in Section 2.2.5 that for a thin-base diode, t_B is the time needed to fill the base region of the diode with minority carriers, and it is also the average time for minority carriers injected from the emitter to traverse the thin base. For a bipolar transistor, the charging current for the base is the current flowing from the emitter into the base, i.e., the collector current. The base transit time for a bipolar transistor is therefore defined by

$$t_B \equiv |Q_B|/|J_C|, \tag{7.17}$$

where J_C is the collector current density and Q_B is the minority-carrier charge per unit area stored in the base region. For an n–p–n transistor, Q_B is given by

$$Q_B = -q\int_0^{W_B} [n_p(x) - n_{p0}(x)]dx. \tag{7.18}$$

It should be noted that Q_B is negative for an n–p–n transistor since the minority carriers in the p-type base are electrons.

Since recombination in a thin base is negligible, J_C is independent of x. Therefore, Eq. (7.10) can be rearranged and then integrated to give

$$\begin{aligned} J_C\int_x^{W_B}\frac{p_p(x')}{qD_{nB}(x')n_{ieB}^2(x')}dx' &= \int_x^{W_B} d\left(\frac{p_p(x')n_p(x')}{n_{ieB}^2(x')}\right) \\ &= -\frac{p_p(x)n_p(x)}{n_{ieB}^2(x)} + \frac{p_p(W_B)n_p(W_B)}{n_{ieB}^2(W_B)} \\ &= -\frac{p_p(x)n_p(x)}{n_{ieB}^2(x)} + \frac{p_{p0}(x)n_{p0}(x)}{n_{ieB}^2(x)} \\ &\approx -\frac{p_p(x)}{n_{ieB}^2(x)}[n_p(x) - n_{p0}(x)], \end{aligned} \tag{7.19}$$

where we have used the boundary condition that the density of excess electrons at W_B is zero, i.e., $n_p(W_B) = n_{p0}(W_B)$, $p_p(W_B) = p_{p0}(W_B)$, the low-injection approximation of $p_p(x) \approx p_{p0}(x)$, and the fact that $p_{p0}(x)n_{p0}(x) = n_{ieB}^2(x)$. Equation (7.19) can be rearranged to give

$$n_p(x) - n_{p0}(x) \approx -\frac{J_C}{q}\frac{n_{ieB}^2(x)}{p_p(x)}\int_x^{W_B}\frac{p_p(x')}{D_{nB}(x')n_{ieB}^2(x')}dx'. \tag{7.20}$$

Substituting Eqs. (7.18) and (7.20) into Eq. (7.17), we obtain

$$t_B = \int_0^{W_B}\frac{n_{ieB}^2(x)}{p_p(x)}\int_x^{W_B}\frac{p_p(x')}{D_{nB}(x')n_{ieB}^2(x')}dx'dx. \tag{7.21}$$

Equation (7.21) is the general expression for the base transit time (Kroemer, 1985). It includes high-current effects, through the dependence of p_p and W_B on minority-carrier concentration, as well as nonuniform-bandgap effects, through the parameter n_{ieB}.

7.2.4.1 Base Transit Time at Low Currents

At low currents, the hole concentration is approximately equal to the base doping concentration, and Eq. (7.21) reduces to

$$t_B \approx \int_0^{W_B}\frac{n_{ieB}^2(x)}{N_B(x)}\int_x^{W_B}\frac{N_B(x')}{D_{nB}(x')n_{ieB}^2(x')}dx'dx. \tag{7.22}$$

If the dopant distribution and the dependencies of mobility and bandgap narrowing on doping concentration are known, Eq. (7.22) can be used to calculate the base transit time at low currents (Suzuki, 1991).

In the *uniform-bandgap approximation* for the base transit time at low currents, n_{ieB} is independent of x and Eq. (7.22) reduces to (Moll and Ross, 1956)

$$t_B \approx \int_0^{W_B}\frac{1}{N_B(x)}\int_x^{W_B}\frac{N_B(x')}{D_{nB}(x')}dx'dx. \tag{7.23}$$

For a box profile, both N_B and D_{nB} are constant, and Eq. (7.23) reduces further to

$$t_B \approx \frac{W_B^2}{2D_{nB}}, \tag{7.24}$$

which, as expected, is the same as Eq. (2.147) for the transit time for a uniformly doped thin-base diode.

Equation (7.24) suggests that ***an effective way to reduce the base transit time is to reduce the intrinsic-base width***. However, as the base width is reduced, the base doping concentration must be increased appropriately to avoid emitter–collector punch-through or the Early voltage becoming unacceptably small.

7.3 Design of the Collector Region

The cross section of the physical structure of a modern n–p–n bipolar transistor is illustrated schematically in Fig. 7.4. The collector includes all the n-type regions underneath and surrounding the p-type base. It can be subdivided into four parts. The part directly underneath the emitter and intrinsic base (shaded in Fig. 7.4) is the active region of the collector. This region is usually referred to simply as the *collector*. It is the region referred to in all the transport and current equations in this book. The horizontal heavily doped (n^+) region underneath the collector is called the *subcollector*, and the heavily doped (n^+) vertical region connecting the subcollector to the collector terminal on the silicon surface is called the *reach-through*. The remaining n-type regions make up the parasitic collector, which is usually relatively lightly doped (n^-) in order to minimize the total base–collector junction capacitance.

To first order, both the subcollector and the reach-through are there only to reduce the series resistance between the collector terminal and the active collector. However, as will be discussed later, the proximity of the subcollector to the intrinsic base, that is, the thickness of the active collector layer, has a strong effect on the base–collector breakdown voltage and on the collector current characteristics at high current densities.

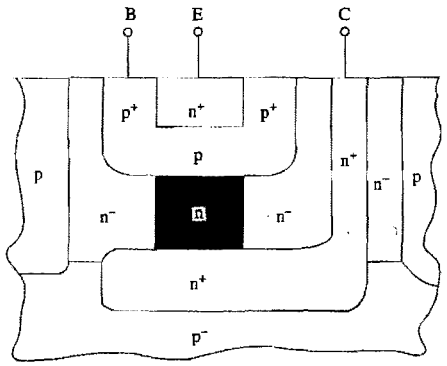


Figure 7.4. Schematic cross section of a modern n–p–n bipolar transistor, illustrating the various doped regions. This transistor employs p-type regions for isolation.

Just like the extrinsic base, the parasitic collector is an unavoidable part of a bipolar transistor structure. In general, designing the parasitic collector is very simple: the parasitic-collector area and its associated capacitance should be as small as possible. As can be seen from Fig. 7.4, the parasitic collector and the extrinsic base form a p–n diode. Therefore, a bipolar technology that gives a small extrinsic-base area will have a small parasitic-collector area as well. For a given extrinsic-base area, the parasitic-collector doping concentration should be as low as possible, in order to achieve the smallest capacitance and the largest breakdown voltage for the extrinsic-base–collector diode. Interested readers are referred to the vast literature on the research and development of bipolar technology, which describes methods for reducing the extrinsic-base area and reducing the parasitic-collector doping concentration (Warnock, 1995; Nakamura and Nishizawa, 1995), and to Appendix 2 for the outline of a process for making a commonly used modern n–p–n transistor.

The collector (we shall use the terms collector and active collector interchangeably whenever there is no confusion) and the subcollector are the only regions that affect the intrinsic characteristics of a bipolar transistor. Therefore, these are the only regions that will be discussed further in this section.

For bipolar transistors operated in the forward-active mode, i.e., with the base–collector junction reverse-biased at all times, the collector acts simply as a sink for the carriers injected from the emitter and traversing the base layer. As discussed in Sections 6.3.2 and 6.3.3, how the collector and subcollector affect the collector current depends on whether or not the collector current density is large enough to cause significant base widening. Thus, the design of the collector will be discussed in two parts, one when base widening is negligible, and the other when base widening is significant. (The collector parameters have no effect on the base current, which, as shown in Section 6.2.2, depends only on the emitter parameters.)

It should be pointed out that one of the most important trends in VLSI technology development is the continued miniaturization of devices while simultaneously increasing

their operating current densities. For bipolar technology, emitter areas of much less than 1 μm^2 can be fabricated readily today, while device currents of 1 mA and larger are desired in many bipolar circuits. That is, the collector current densities in many modern bipolar transistors can easily exceed 1 $mA/\mu m^2$. At these high current densities, ***base widening can easily occur***, and special attention should be paid to the design of the collector and subcollector to minimize its effects.

7.3.1 Collector Design When There Is Negligible Base Widening

As discussed in Section 6.3.3, to maintain negligible base widening, the collector current density J_C should be small compared to the maximum current density, J_{max}, that can be supported by the collector doping concentration N_C. That is, J_C should satisfy the condition that

$$|J_C| \ll J_{max} = qv_{sat}N_C \qquad \text{for negligible base widening,} \tag{7.25}$$

where v_{sat} is the saturated velocity for electrons in silicon, which is about 1×10^7 cm/s (see Fig. 2.10). Published data suggest that base widening becomes quite appreciable in modern n–p–n transistors when $J_C > 0.3J_{max}$.

For $N_C = 1 \times 10^{16}$ cm^{-3}, one has $J_{max} = 0.16$ $mA/\mu m^2$, and the allowed J_C is only about 0.05 $mA/\mu m^2$, which is much too small for the modern bipolar devices. To increase the collector current density without increasing base-widening effect, N_C must be increased proportionately. However, as N_C is increased, the base–collector junction capacitance is increased, and other device characteristics, such as base–collector junction avalanche, can be adversely affected. Therefore, ***tradeoffs have to be made in the design of the collector***. These design tradeoffs are discussed below.

7.3.1.1 Tradeoff in Early Voltage

The Early voltage of a bipolar transistor is inversely proportional to the base–collector junction depletion-layer capacitance per unit area, C_{dBC} [cf. Eq. (6.71)]. As N_C is increased to allow a larger collector current density, C_{dBC} is increased and V_A will decrease. Therefore, there is a tradeoff between the current-density capability of a transistor and its Early voltage.

7.3.1.2 Tradeoff in Base-Collector Junction Avalanche Effect

As discussed in Section 6.3.2, base–collector junction avalanche occurs when the electric field in the junction space-charge region becomes too large. Excessive base–collector junction avalanche can cause the base and collector currents to increase out of control and hence can affect the functionality of the circuits using these transistors. Indeed, when base–collector junction avalanche runs away, device breakdown occurs. Bipolar circuits typically operate with a power supply voltage of 3.3 or 5 V. These voltages are sufficiently high that significant base–collector junction avalanche can easily occur unless care has been taken in the collector design to minimize it (Lu and Chen, 1989).

There are several ways to reduce avalanche multiplication in the base–collector junction. The most straightforward way is to reduce N_C, but that will proportionately reduce the allowed collector current density. Alternatively, the base and/or the collector

doping profiles, at or near the base–collector junction, can be designed to reduce the electric field in the junction.

Referring to Fig. 7.2, the Gaussian base doping profile, with its graded dopant distribution near the base–collector junction, has a lower electric field in the base–collector junction than the boxlike base doping profile. In practice, ion implantation of boron usually results in an exponential tail in the base doping profile, as can be seen from Fig. 6.2. This tail is caused by a combination of channeling effect during ion implantation and defect-induced enhanced-diffusion effect during postimplantation thermal annealing. The ion-implanted base profile is therefore always graded.

If the intrinsic base is formed by epitaxial growth and is doped *in situ*, its doping profile can be much more boxlike. For the same collector doping profile, such a base doping profile will result in a larger electric field in the base–collector junction. However, this does not imply that a graded base doping profile is preferred over a boxlike profile. This point will be discussed further in Chapter 8 in connection with the optimization of a device design.

The collector doping profile can also be retrograded (i.e., graded with its concentration increasing with distance into the silicon) to reduce the electric field in the base–collector junction (Lee *et al.*, 1996). Retrograding of the collector doping profile can be achieved readily by high-energy ion implantation. The transistor doping profiles illustrated in Fig. 6.2 show collectors with retrograded doping profiles.

Qualitatively, grading the base doping profile, and/or retrograding the collector doping profile, is similar to sandwiching an i-layer between the base and collector doped regions. Introducing a thin i-layer between the p- and n-regions of a diode is quite effective in reducing the electric field in the junction, as discussed in Section 2.2.2.

Reducing base–collector junction avalanche, either by reducing the collector doping concentration or by grading the base doping profile and/or retrograding the collector doping profile, reduces the base–collector junction depletion-layer capacitance as well. This should help to improve the device and circuit performance (Lee *et al.*, 1996). However, as can be seen from Eqs. (6.81) and (6.82), these techniques for reducing the base–collector junction capacitance also lead to more base widening, or to base widening occurring at a lower collector current density. Thus, reducing base–collector junction avalanche can reduce the current-density capability, and hence the maximum speed, of a bipolar transistor (Lu and Chen, 1989). The tradeoff between base–collector junction avalanche effect and device and circuit speed will be discussed further in Chapter 8.

7.3.2 Collector Design When There Is Appreciable Base Widening

As mentioned earlier, the operating current densities of a modern bipolar transistor could easily be in excess of 1 mA/μm^2, if base-widening effect were not a concern. Unfortunately, at these high current densities, base widening does occur. The challenge in designing the collector when base widening is unavoidable is to minimize the deleterious effects of base widening.

As shown in Section 6.3.3, when base widening occurs, there are excess minority carriers stored in the collector, and, as shown in Section 6.4.4, these excess minority

carriers contribute to the emitter diffusion capacitance. As will be shown in Chapter 8, when a bipolar transistor is operated with significant base widening, it is its emitter diffusion capacitance that limits its circuit speed and cutoff frequency. To minimize emitter diffusion capacitance, the total excess minority carriers stored in the collector should be minimized. To accomplish this goal, in addition to retrograding the collector doping profile as discussed in the previous subsection, the total collector volume available for minority-carrier storage should also be minimized. That is, the thickness of the collector layer should be minimized. This is easily accomplished by reducing the thickness of the epitaxial layer grown after the subcollector region is formed (see Appendix 2).

However, thinning the collector can lead to an increase in the base–collector junction depletion-layer capacitance, if the collector thickness is comparable to the base–collector depletion-layer width. Thus, when operated at low current densities, where base widening is negligible, a circuit using thin-collector transistors could run slower than a circuit using thick-collector transistors. However, at high current densities, circuits with thin-collector transistors often run faster than circuits with thick-collector transistors (Tang *et al.*, 1983). Also, when the collector–base space-charge layer extends all the way to the subcollector, base–collector junction avalanche will increase, and the base–collector junction breakdown voltage will decrease.

Designing the collector of a modern bipolar transistor is therefore a complex tradeoff process. The important point to remember is that ***base widening occurs readily in modern bipolar devices, and optimizing the tradeoff in the collector design is key to realizing the maximum performance of these devices.***

7.4 SiGe-Base Bipolar Transistors

The energy bandgap of Ge (≈ 0.66 eV) is significantly smaller than that of Si (≈ 1.12 eV). By incorporating Ge into the base region of a Si bipolar transistor, the energy bandgap of the base region, and hence the accompanied device characteristics, can be modified (Iyer *et al.*, 1987). When Ge is incorporated into Si, the Si energy bandgap becomes smaller primarily owing to shifting of the valence band edge (People, 1986; Van de Walle and Martin, 1986). The larger the Ge concentration the smaller the energy bandgap. A SiGe-base bipolar transistor is usually designed to have a graded Ge distribution in the base, i.e. with lower Ge concentration at the emitter end and larger Ge concentration at the collector end, in order to establish a drift field which drives electrons across the quasi-neutral base layer (Patton *et al.*, 1990; Harame *et al.*, 1995a, b).

The emitter of a typical SiGe-base bipolar transistor is the same as that of a regular Si-base bipolar transistor. In both transistors, it is simply a polysilicon emitter. As for the Ge distribution in the base, several variations of a graded Ge profile have been studied. The most commonly used profile is that of a triangular or linearly graded Ge distribution. This profile assumes a Ge distribution which is zero at the emitter end of the quasineutral base and increases at a constant rate across the base layer. It leads to a simple graded base bandgap that decreases linearly from the emitter end to the collector end.

In a SiGe-base bipolar device fabrication process, Ge is incorporated into a starting base layer prior to the polysilicon-emitter formation step. Depending on the details of the base and emitter formation steps, Ge may or may not end up in the single-crystalline region of the emitter. Once Ge ends up in the single-crystalline portion of the emitter, the Ge profile within the quasineutral base can become quite complex. In particular, the Ge distribution at the emitter end of the quasineutral base will depend on the depth of the single-crystalline emitter region. Therefore, a trapezoidal Ge profile, with a low but finite Ge concentration near the emitter end and a higher Ge concentration at the collector end, gives a more general description of the Ge distribution in a typical SiGe-base transistor. A SiGe-base transistor having a trapezoidal Ge distribution in its base can be modeled with close-form solutions. Furthermore, a triangular Ge profile and a constant-Ge profile can be treated as special cases of a trapezoidal profile.

In Section 7.4.1, the properties of a polysilicon-emitter SiGe-base transistor having a linearly graded base bandgap, corresponding to a simple triangular Ge profile, are discussed and compared to those of a polysilicon-emitter Si-base transistor. A triangular profile describes very well the basic properties of a typical polysilicon-emitter SiGe-base bipolar transistor. For readers who desire only a first-order explanation of the difference between a SiGe-base transistor and a Si-base transistor, this simple description should be adequate.

In the remaining sections, the properties of a SiGe-base bipolar transistor having a trapezoidal Ge distribution in the base are discussed in greater depth. These sections are intended for those readers interested in understanding the more subtle properties of a SiGe-base bipolar transistor. The models developed in these sections can also be used for optimizing the Ge distribution, beyond the simple triangular distribution, for improved device characteristics.

The presence of Ge in the emitter changes the properties of the emitter region, which in turn can change the base current characteristics. The effect on base current due to the presence of Ge in the emitter is considered in Section 7.4.2. The collector current, Early voltage and base transit time are modeled in Section 7.4.3 for a transistor having a trapezoidal Ge distribution, and in Section 7.4.4 for a transistor having a constant Ge distribution.

For a given device fabrication process, there is always a distribution in emitter depth and base width caused by process variation. A methodology for evaluating the effect of emitter depth variation on device characteristics is developed in Section 7.4.5. The results are then applied to the optimization of a Ge profile in Section 7.4.6. There are also subtle but interesting effects in a SiGe-base transistor that are either absent or relatively unimportant in a Si-base transistor. They are discussed in Sections 7.4.7 and 7.4.8. Finally, Section 7.4.9 is devoted to a discussion of the heterojunction nature of a SiGe-base bipolar transistor, contrasting a SiGe-base transistor with a traditional wide-gap-emitter *heterojunction bipolar transistor* (HBT).

7.4.1 Transistors Having a Simple Linearly Graded Base Bandgap

It is shown in Appendix 17 that a simple triangular Ge distribution in the base of a Si–SiGe n^+–p diode produces a bandgap grading in the base such that the valence-band edge

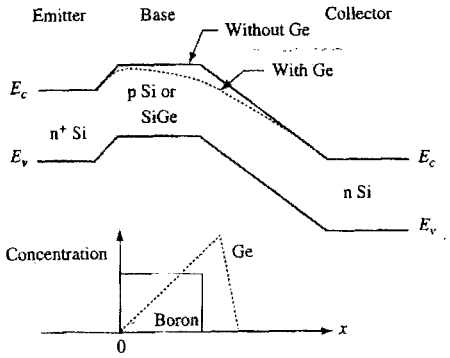


Figure 7.5. Schematic illustration of the energy bands of a SiGe-base n–p–n transistor (dotted) and a Si-base n–p–n transistor (solid). Both transistors are assumed to have the same base doping profile. The base bandgaps of the two transistors are the same near the base–emitter junction. The base bandgap of the SiGe-base transistor narrows gradually towards the base–collector junction.

in the base is essentially spatially constant, while the conduction-band edge has a downward slope towards the p-type SiGe contact, i.e. E_c decreases with distance x from the emitter–base junction. As a result, the energy-band diagram for a SiGe-base bipolar transistor having a triangular Ge distribution in the base is as illustrated in Fig. 7.5.

As shown in Section 6.2.2, the base current is determined by the emitter parameters only, and is independent of the base parameters. A SiGe-base bipolar transistor typically has the same polysilicon emitter as a Si-base transistor. Also, it is shown in Appendix 17 that the presence of Ge in the base does not change the energy barrier for hole injection from the base into the emitter. Therefore, the base current of a SiGe-base transistor should be the same as that of a Si-base transistor. This is indeed the case for most SiGe-base transistors. (Even when Ge ends up in the single-crystalline emitter region, the effect on base current is still small, as will be explained in Section 7.4.2.)

Since base current is not affected by the presence of Ge in the base, we need to consider only the effect of Ge in the base on collector current. The base bandgap-narrowing parameter in Eq. (7.11) can be extended to include bandgap narrowing caused by the presence of Ge. That is, the effective intrinsic-carrier concentration in the base containing Ge can be written in the form (Kroemer, 1985)

$$n_{ieB}^2(\mathrm{SiGe},x) = n_{ieB}^2(\mathrm{Si},x)\gamma(x)\exp\left[\frac{\Delta E_{gB,SiGe}(x)}{kT}\right], \tag{7.26}$$

where $n_{ieB}(\mathrm{Si}, x)$ is the effective intrinsic-carrier concentration without Ge, $\Delta E_{gB,SiGe}(x)$ is the local bandgap narrowing in the base due to the presence of Ge, and the parameter

$$\gamma(x) \equiv \frac{(N_c N_v)_{SiGe}}{(N_c N_v)_{Si}} \tag{7.27}$$

is introduced to account for any change in the density of states caused by the presence of Ge (Harame *et al.*, 1995a,b). Effects due to heavy doping are contained in the parameter $n_{ieB}(\mathrm{Si}, x)$. In addition to reducing the bandgap energy, the incorporation of Ge into Si also lifts the degeneracy of the valence-band and conduction-band edges (People, 1985). The result is a reduction in the densities of states N_c and N_v. That is, $\gamma(x) < 1$ except where the Ge concentration is zero. For the Ge distribution being considered here, the Ge-induced bandgap narrowing is zero at the emitter–base junction and increases linearly to $\Delta E_{g\max}$ at the base–collector junction:

$$\Delta E_{gB,SiGe}(x) = \frac{x}{W_B}\Delta E_{g\ \max}. \tag{7.28}$$

7.4.1.1 Collector Current and Current Gain

The collector current density in a SiGe-base transistor can be obtained simply by substituting $n_{ieB}(\mathrm{SiGe}, x)$ for $n_{ieB}(\mathrm{Si}, x)$ and $D_{nB}(\mathrm{SiGe}, x)$ for $D_{nB}(\mathrm{Si}, x)$ in the equations derived earlier for the collector current density in a Si-base transistor. Thus, the saturated collector current density given in Eq. (6.33) becomes

$$J_{C0}(\mathrm{SiGe}) = \frac{q}{\displaystyle\int_0^{W_B} \frac{p_p(x)}{D_{nB}(\mathrm{SiGe}, x) n_{ieB}^2(\mathrm{SiGe}, x)} dx}. \tag{7.29}$$

For a boxlike base doping profile and at low current densities, $p_p(x) \approx N_B$ and is independent of x. $D_{nB}(\mathrm{Si}, x)$ and $n_{ieB}(\mathrm{Si}, x)$ are also independent of x. Therefore, Eqs. (7.26) and (7.29) give the ratio of the collector current with Ge to that without Ge as

$$\begin{aligned}\frac{J_{C0}(\mathrm{SiGe})}{J_{C0}(\mathrm{Si})} &= \frac{W_B}{\displaystyle\int_0^{W_B} \frac{1}{\gamma(x)\eta(x)} \exp\left[-\Delta E_{gB,SiGe}(x)/kT\right] dx} \\ &\approx \frac{\overline{\gamma\eta} W_B}{\displaystyle\int_0^{W_B} \exp\left[-\Delta E_{gB,SiGe}(x)/kT\right] dx},\end{aligned} \tag{7.30}$$

where the ratio parameter

$$\eta(x) \equiv \frac{D_{nB}(\mathrm{SiGe}, x)}{D_{nB}(\mathrm{Si}, x)} \tag{7.31}$$

accounts for the effect of Ge on electron mobility in the base. $D_{nB}(\mathrm{SiGe})$ is proportional to $\mu_{nB}(\mathrm{SiGe})$ which is found to be a function of both base doping concentration and Ge concentration (Kay and Tang, 1991). Therefore, $\eta(x)$ is a function of base doping concentration and Ge concentration as well. In writing the last part of Eq. (7.30), we have made an assumption that $\gamma(x)$ and $\eta(x)$ inside the integral can be replaced by some average values $\bar{\gamma}$ and $\bar{\eta}$. It should be noted that Eq. (7.30) is valid for any arbitrary dependence of $\Delta E_{gB,SiGe}(x)$ on x.

For the simple linearly graded bandgap described by Eq. (7.28), Eq. (7.30) can be integrated to give

$$\frac{J_{C0}(\mathrm{SiGe})}{J_{C0}(\mathrm{Si})} = \frac{\overline{\gamma\eta}\Delta E_{g\ \max}/kT}{1 - \exp\left(-\Delta E_{g\ \max}/kT\right)}. \tag{7.32}$$

The value of $\gamma(x)$ varies between 1 where there is no Ge to about 0.4 where the Ge concentration is 20% (Prinz *et al.*, 1989). For a typical SiGe base where the base doping concentration is in excess of 1×10^{18} cm^{-3}, $\eta(x)$ is about unity (Kay and Tang, 1991; Manku and Nathan, 1992). Therefore, the product $\overline{\gamma\eta}$ is not far from unity. In the literature, the corrections to density of states and to electron mobility are often ignored in discussing the effect of Ge on collector current, which is equivalent to setting $\overline{\gamma\eta}$ to unity in Eq. (7.32). Equation (7.32) is the well-known result for a simple triangular Ge distribution (Harame *et al.*, 1995a, b).

As discussed earlier, the base current of a SiGe-base transistor is the same as that of a Si-base transistor. The current gain ratio is therefore the same as the collector current ratio, namely

$$\frac{\beta_0(\mathrm{SiGe})}{\beta_0(\mathrm{Si})} = \frac{J_{C0}(\mathrm{SiGe})}{J_{C0}(\mathrm{Si})} = \frac{\overline{\gamma\eta}\Delta E_{g\ \max}/kT}{1 - \exp\left(-\Delta E_{g\ \max}/kT\right)}. \tag{7.33}$$

Readily achievable values of $\Delta E_{g\max}$ are in the range of 100–150 meV, which means a SiGe-base transistor typically has a collector current and current gain that are 4 to 6 times those of a Si-base transistor having the same base dopant distribution. As shown in Eq. (7.7), the collector current density, and hence the current gain, is proportional to the intrinsic-base sheet resistivity. The enhanced current gain of a SiGe-base transistor can be used to tradeoff for a smaller intrinsic-base sheet resistivity. The resultant smaller intrinsic-base sheet resistivity increases Early voltage as well [cf. Eq. (6.73)].

7.4.1.2 Early Voltage

The effect of base-bandgap grading on Early voltage can be obtained from Eq. (6.73) in a similar manner by replacing $n_{ieB}(\mathrm{Si}, x)$ by $n_{ieB}(\mathrm{SiGe}, x)$ and $D_{nB}(\mathrm{Si}, x)$ by $D_{nB}(\mathrm{SiGe}, x)$. The result is

$$\begin{aligned}&V_A(\mathrm{SiGe}) \\ &\approx \frac{qD_{nB}(\mathrm{SiGe}, W_B) n_{ieB}^2(\mathrm{SiGe}, W_B)}{C_{dBC}} \int_0^{W_B} \frac{N_B(x)}{D_{nB}(\mathrm{SiGe}, x) n_{ieB}^2(\mathrm{SiGe}, x)} dx,\end{aligned} \tag{7.34}$$

where C_{dBC} is the base–collector junction depletion-layer capacitance per unit area. Using Eq. (6.72) for $V_A(\mathrm{Si})$ and substituting Eqs. (7.26) and (7.27) into Eq. (7.34) we obtain the ratio of the Early voltage of a SiGe-base transistor to that of a Si-base transistor having the same boxlike boron distribution in the base as

$$\frac{V_A(\text{SiGe})}{V_A(\text{Si})} = \frac{D_{nB}(\text{SiGe}, W_B)n_{ieB}^2(\text{SiGe}, W_B)}{W_B}\int_0^{W_B}\frac{dx}{D_{nB}(\text{SiGe}, x)n_{ieB}^2(\text{SiGe}, x)}$$
$$\approx \frac{D_{nB}(\text{SiGe}, W_B)}{\bar{D}_{nB}(\text{SiGe})}\frac{\gamma(W_B)}{\bar{\gamma}}\frac{\exp(\Delta E_{g\ \max}/kT)}{W_B}\int_0^{W_B}\exp\left[-\Delta E_{gB,SiGe}(x)/kT\right]dx$$
$$\approx \frac{\exp(\Delta E_{g\ \max}kT)}{W_B}\int_0^{W_B}\exp\left[-\Delta E_{gB,SiGe}(x)/kT\right]dx, \quad (7.35)$$

where, in writing the last part of the equation, we have made a further assumption that the average values of D_{nB} and γ are about the same as their values at the base–collector junction. It should be noted that Eq. (7.35) is valid for any arbitrary dependence of $\Delta E_{gB,SiGe}(x)$ on x.

For the simple linearly graded base bandgap described by Eq. (7.28), Eq. (7.35) can be integrated to give

$$\frac{V_A(\text{SiGe})}{V_A(\text{Si})} = \frac{kT}{\Delta E_{g\ \max}}\left[\exp(\Delta E_{g\ \max}/kT) - 1\right]. \quad (7.36)$$

Equation (7.36) is the well-known result for a simple triangular Ge distribution (Harame *et al.*, 1995a). It shows that the Early voltage increases approximately exponentially with $\Delta E_{g\max}/kT$ when $\Delta E_{g\max}/kT > 1$. For a typical value of $\Delta E_{g\max} = 100$ meV, the Early voltage is increased by a factor of 12 at room temperature.

Combining Eqs. (7.33) and (7.36), the ratio of $\beta_0 V_A$ for a SiGe-base transistor to that for a Si-base transistor is

$$\frac{\beta_0(\text{SiGe})V_A(\text{SiGe})}{\beta_0(\text{Si})V_A(\text{Si})} = \bar{\gamma}\bar{\eta}\exp(\Delta E_{g\ \max}kT). \quad (7.37)$$

The same result could have been obtained from Eq. (6.74) by using Eq. (7.26) for $n_{ieB}(\text{SiGe}, W_B)$. Again, in the literature, the product $\bar{\gamma}\bar{\eta}$ is often assumed to be unity and dropped. For $\Delta E_{g\max} = 100$ meV, the $\beta_0 V_A$ product is increased by almost a factor of 50 at room temperature.

7.4.1.3 Base Transit Time

The graded base bandgap introduces an electric field which drives electrons across the p-type base layer. For a total bandgap narrowing of 100 meV across a base layer of 100 nm, a SiGe-base transistor has a built-in electric field of 10^4 V/cm in the base due to the presence of Ge alone. This field is in addition to the electric fields due to base dopant distribution and heavy-doping effect, which have been discussed earlier in Section 7.2.3. As can be seen from comparing this field with the fields plotted in Fig. 7.3, the electric field due to the presence of a graded Ge distribution can be comparable to the maximum fields due to dopant distribution and heavy-doping effect. Consequently, the base transit time of a SiGe-base transistor can be significantly smaller than that of a Si-base transistor having the same base dopant distribution. The base transit time at low current densities

can be derived by substituting $n_{ieB}(\text{Si}, x)$ with $n_{ieB}(\text{SiGe}, x)$ and $D_{nB}(\text{Si}, x)$ with $D_{nB}(\text{SiGe}, x)$ in Eq. (7.22). The result is

$$t_B(\text{SiGe}) \approx \int_0^{W_B}\frac{n_{ieB}^2(\text{SiGe}, x)}{N_B(x)}\int_x^{W_B}\frac{N_B(x')}{D_{nB}(\text{SiGe}, x')n_{ieB}^2(\text{SiGe}, x')}dx'dx$$
$$= \int_0^{W_B}\frac{\gamma(x)n_{ieB}^2(\text{Si}, x)}{N_B(x)}\exp\left[\Delta E_{gB,SiGe}(x)/kT\right]$$
$$\times\int_x^{W_B}\frac{N_B(x')}{\gamma(x')\eta(x')D_{nB}(\text{Si}, x')n_{ieB}^2(\text{Si}, x')}\exp\left[-\Delta E_{gB,SiGe}(x')/kT\right]dx'dx. \quad (7.38)$$

Again, for a boxlike base dopant distribution, $N_B(x)$, $D_{nB}(\text{Si}, x)$, and $n_{ieB}(\text{Si}, x)$ are all independent of x, and Eq. (7.38) simplifies to

$$t_B(\text{SiGe}) \approx \frac{1}{\bar{\eta}D_{nB}(\text{Si})}\int_0^{W_B}\exp\left[\Delta E_{gB,SiGe}(x)/kT\right]$$
$$\times\int_x^{W_B}\exp\left[-\Delta E_{gB,SiGe}(x')/kT\right]dx'dx, \quad (7.39)$$

where, consistent with the approximations made earlier for collector current and Early voltage, we have made the assumption that $\gamma(x)$ and $\eta(x)$ are relatively slow varying functions of x and hence can be replaced by their average values $\bar{\gamma}$ and $\bar{\eta}$. The base transit time for a Si-base transistor is $t_B(\text{Si}) = W_B^2/2D_n(\text{Si})$, given in Eq. (7.24). Therefore, the ratio of the base transit time of a SiGe-base transistor to that of a Si-base transistor having the same base width and boxlike base dopant distribution is

$$\frac{t_B(\text{SiGe})}{t_B(\text{Si})} = \frac{2}{\bar{\eta}W_B^2}\int_0^{W_B}\exp\left[\Delta E_{gB,SiGe}(x)/kT\right]$$
$$\times\int_x^{W_B}\exp\left[-\Delta E_{gB,SiGe}(x')/kT\right]dx'dx. \quad (7.40)$$

Equation (7.40) is valid for any arbitrary dependence of $\Delta E_{gB,SiGe}(x)$ on x.

For the simple linearly graded base bandgap described by Eq. (7.28), Eq. (7.40) can be integrated to give

$$\frac{t_B(\text{SiGe})}{t_B(\text{Si})} = \frac{2kT}{\bar{\eta}\Delta E_{g\ \max}}\left\{1 - \frac{kT}{\Delta E_{g\ \max}}\left[1 - \exp(-\Delta E_{g\ \max}/kT)\right]\right\}. \quad (7.41)$$

Again, in the literature, the diffusion coefficient correction factor $\bar{\eta}$ is often set to unity and dropped. For $\Delta E_{g\max} = 100$ meV, the low-current base transit time of a SiGe-base transistor is about 2.5 times smaller than that of a Si-base transistor having the same base width and base dopant distribution. Equation (7.41) is the well-known result for a simple triangular Ge distribution (Harame *et al.*, 1995a, b).

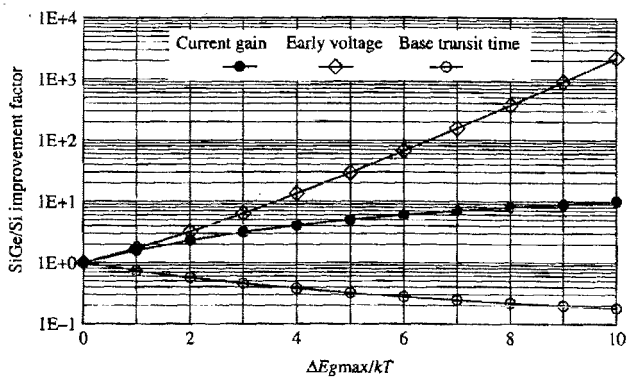


Figure 7.6. Relative improvement factors for current gain, Early voltage, and base transit time of a SiGe-base bipolar transistor over a Si-base bipolar transistor, as a function of the maximum base bandgap narrowing. A linearly graded Ge profile is assumed. Also, $\bar{\gamma}$ and $\bar{\eta}$ are set to unity.

7.4.1.4 Emitter Delay Time

It will be shown in Chapter 8 that the cutoff frequency f_T of a bipolar transistor is limited by the forward transit time τ_F of which the emitter delay time τ_E is one of the components. It will also be shown in Chapter 8 [see Eq. (8.16)] that τ_E is inversely proportional to the current gain. Thus, with a significantly larger current gain, a SiGe-base transistor has a much smaller emitter delay time than a Si-base transistor of the same emitter design.

7.4.1.5 SiGe-Base Bipolar as a High-Frequency Transistor

It will be shown in Chapter 8 that some of the desirable attributes of a high-frequency bipolar transistor are: small transit times, small base resistance, and large output resistance or Early voltage. Figure 7.6 is a plot of the improvement factors for current gain [Eq. (7.33)], Early voltage [Eq. (7.36)], and base transit time [Eq. (7.41)], of a SiGe-base bipolar transistor relative to a Si-base bipolar transistor having the same base width and base dopant distribution, plotted as a function of $\Delta E_{gmax}/kT$ using $\bar{\gamma} = 1$ and $\bar{\eta} = 1$. It shows that incorporating a linearly graded Ge distribution into the base of a bipolar transistor can greatly improve its current gain, Early voltage, and base transit time. As discussed in the previous subsection, the larger current gain also implies a smaller emitter delay time. Alternatively, the larger current gain can be traded off for a smaller intrinsic-base resistance. Thus, compared to a Si-base bipolar transistor, a SiGe-base bipolar transistor is much superior in frequency performance.

7.4.2 Base Current When Ge Is Present in the Emitter

Polysilicon emitter is employed in the fabrication of all modern silicon bipolar transistors, including SiGe-base transistors. In the fabrication of a SiGe-base transistor, typically a p-type base layer containing the desired Ge distribution is first formed before an n^+ emitter polysilicon layer is formed on top. There is usually a Ge-free cap in the starting base layer to avoid exposing Ge to any oxidizing ambient in the polysilicon-emitter formation process. During the polysilicon-emitter formation process, n-type dopant from the polysilicon layer diffuses into the starting base layer, forming a thin single-crystalline n^+ emitter region of depth x_{jE} (see Section 7.1.2 and Fig. 6.2). The final value of x_{jE} is a function of emitter annealing condition (temperature and time), emitter dopant species (arsenic or phosphorus), emitter stripe width, and whether or not metal silicide is formed on top of the emitter polysilicon layer (Kondo *et al.*, 2001). Depending on the thickness of the starting Ge-free cap, the final emitter–base junction may or may not extend into the Ge-containing region of the base, and hence Ge may or may not be present in the single-crystalline emitter region. Since any change in the emitter parameters can affect the base current, we want to consider what happens to the base current of a polysilicon-emitter SiGe-base bipolar transistor when Ge from the starting base layer ends up in the single-crystalline emitter region.

In the literature, there are reports of intentionally introducing Ge into the single-crystalline emitter region (Huizing *et al.*, 2001) as well as intentionally adding Ge to the emitter polysilicon layer (Martinet *et al.*, 2002; Kunz *et al.*, 2002; Kunz *et al.*, 2003). Often the stated objectives are to reduce current gain of a SiGe-base transistor. The merits of these and other approaches for reducing current gain of a SiGe-base transistor will also be discussed.

7.4.2.1 Ge-Induced Bandgap Narrowing in the Emitter

The Ge distribution in the emitter and base regions of a polysilicon-emitter SiGe-base bipolar transistor having a boxlike base dopant distribution is illustrated in Fig. 7.7 for the case of a trapezoidal Ge distribution. The Ge distribution causes a bandgap narrowing of ΔE_{g0} near the emitter–base junction and a peak bandgap narrowing of ΔE_{gmax} at the base–collector junction. For the case illustrated in Fig. 7.7, the emitter depth x_{jE} is larger than the starting Ge-free cap thickness, W_{cap}, resulting in a bandgap narrowing of $\Delta E_{g0be} > \Delta E_{g0}$ at the emitter–base junction. There is also Ge present within the single-crystalline emitter region, causing a narrowing of the bandgap in the region. Since base current is determined by the emitter parameters, the Ge-induced bandgap variation in the emitter affects the base current. In the next subsection, we examine the base current when there is Ge in the emitter (Ning, 2003a).

7.4.2.2 Base Current When There Is Ge in the Single-Crystalline Emitter Region

It is shown in Section 6.2.2 that a polysilicon emitter can be modeled as a shallow or transparent emitter having a finite surface recombination velocity at the emitter contact, i.e., at the polysilicon–silicon interface. Consistent with the convention used in Section 6.2.2, Fig. 7.8 shows the coordinates for modeling the current flows in the emitter region of the emitter–base diode of Fig. 7.7. The p–n junction is assumed to be located at the origin "0". The emitter is contacted by a polysilicon layer, with the polysilicon–silicon interface located at $x = -W_E$, i.e., $W_E = x_{jE}$.

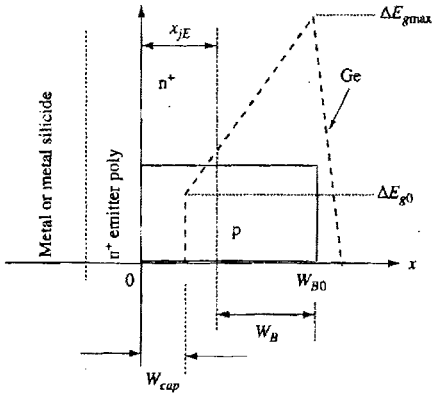


Figure 7.7. Schematic illustrating the emitter and base regions of a polysilicon-emitter SiGe-base bipolar transistor with a trapezoidal Ge distribution. The starting base layer thickness is W_{B0}, including a Ge-free cap layer of thickness W_{cap}. The quasineutral base width is W_B after polysilicon-emitter drive in. The base width is a function of emitter depth x_{jE}, given by $W_B = W_{B0} - x_{jE}$. The emitter–base space-charge region thickness is assumed to be zero, for simplicity of illustration. With $x_{jE} > W_{cap}$, there is no residual Ge-free region in the final quasineutral base layer, but there is Ge in the single-crystalline n^+ emitter region.

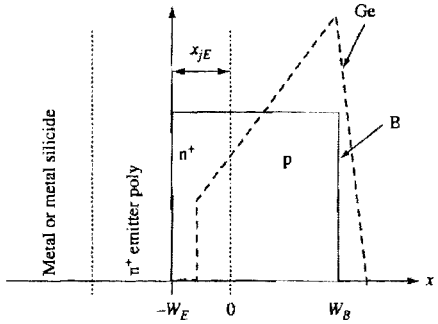


Figure 7.8. Coordinates for modeling the current flows in the emitter of a polysilicon-emitter SiGe-base bipolar transistor.

Following Eqs. (6.43) and (6.44), the saturated base current density in a SiGe-base bipolar transistor can be written as

$$J_{B0}(\text{SiGe}) = \frac{qn_i^2}{G_E(\text{SiGe})}, \tag{7.42}$$

with the emitter Gummel number as

$$G_E(\text{SiGe}) = \int_{-W_E}^{0} \frac{n_i^2}{n_{ieE}^2(\text{SiGe}, x)} \frac{N_E(x)}{D_{pE}(\text{SiGe}, x)} dx + \frac{n_i^2}{n_{ieE}^2(\text{SiGe}, -W_E)} \frac{N_E(-W_E)}{S_p(\text{SiGe})}, \tag{7.43}$$

where S_p(SiGe) is the surface recombination velocity for holes at the polysilicon–silicon interface, and $N_E(x)$, D_{pE}(SiGe, x), and n_{ieE}(SiGe, x) are the doping concentration, hole diffusion coefficient, and effective intrinsic-carrier concentration, respectively, in the single-crystalline emitter region. The surface recombination velocity S_p(SiGe) depends on the transport of holes through the polysilicon–silicon interface and inside the polysilicon layer. For example, it is shown in Ex. 6.3 and in the literature (Ning and Isaac, 1980) that for a Si-base bipolar transistor S_p(Si) depends only on the transport of holes inside the polysilicon layer when there is no appreciable hole barrier at the polysilicon–silicon interface. In this simple case, S_p(Si) is given by

$$S_p(\text{Si}) = \frac{D_{pE,poly}}{L_{pE,poly} \tanh\left(\frac{W_{E,poly}}{L_{pE,poly}}\right)}, \tag{7.44}$$

where $D_{pE,poly}$ and $L_{pE,poly}$ are the hole diffusion coefficient and hole diffusion length, respectively, in the emitter polysilicon, and $W_{E,poly}$ is the thickness of the emitter polysilicon layer.

It should be noted that regardless of the details of the physical model for S_p, the operation of a polysilicon-emitter transistor is based on the experimentally confirmed fact that the hole current is determined primarily by the surface recombination velocity of holes at the polysilicon–silicon interface and is relatively insensitive to the transport of holes within the shallow single-crystalline emitter region. That is, the operation of a polysilicon-emitter transistor is based on the assumption that G_E is determined primarily by the term containing S_p in Eq. (7.43). In other words, for a polysilicon-emitter SiGe-base bipolar transistor,

$$G_E(\text{SiGe}) \approx \frac{n_i^2 N_E(-W_E)}{n_{ieE}^2(\text{SiGe}, -W_E) S_p(\text{SiGe})}. \tag{7.45}$$

Following the same procedure used in Section 7.4.1 to model the SiGe base region, we can write the emitter parameter n_{ieE}(SiGe, x) in the form

$$n_{ieE}^2(\text{SiGe}, x) = n_{ieE}^2(\text{Si}, x)\gamma_E(x) \exp\left[\frac{\Delta E_{gE,SiGe}(x)}{kT}\right], \tag{7.46}$$

where n_{ieE}(Si, x) is the effective intrinsic-carrier concentration without Ge and $\Delta E_{gE,SiGe}(x)$ is the local bandgap narrowing due to the presence of Ge. Also, the parameter

$$\gamma_E(x) \equiv \frac{(N_c N_v)_{SiGe}}{(N_c N_v)_{Si}} \tag{7.47}$$

is to account for any change in the densities of states in the emitter due to the presence of Ge. Effects due to heavy doping are contained in the parameter n_{ieE}(Si, x). From Eqs. (7.42),

(7.45) and (7.46) we can write the ratio of the base current of a polysilicon-emitter SiGe-base bipolar transistor to that of a polysilicon-emitter Si-base bipolar transistor as (Ning, 2003a)

$$\frac{J_{B0}(\text{SiGe})}{J_{B0}(\text{Si})} \approx \frac{S_p(\text{SiGe})}{S_p(\text{Si})} \gamma_E(-W_E) \exp\left[\Delta E_{gE,SiGe}(-W_E)/kT\right]. \tag{7.48}$$

As discussed earlier, there is a Ge-free cap in the starting base layer prior to the emitter formation steps. That is, the Ge concentration is zero at or near the polysilicon–silicon interface. Therefore, $\Delta E_{gE,SiGe}(-W_E) = 0$ and $\gamma_E(-W_E) = 1$ for a typical polysilicon-emitter SiGe-base transistor. Furthermore, we expect $S_p(\text{Si}) \approx S_p(\text{SiGe})$ in this case because there is no Ge at or near the interface and there is no Ge inside the emitter polysilicon layer. Equation (7.48) then suggests that, for a typical polysilicon-emitter SiGe-base bipolar transistor, the base current should be insensitive to the Ge distribution in the starting base layer, even when Ge ends up inside the single-crystalline region of the emitter. This explains why the measured base current of a polysilicon-emitter SiGe-base transistor and that of a polysilicon-emitter Si-base control are approximately the same (Prinz and Sturm, 1990; Harame *et al.*, 1995a, b; Oda *et al.*, 1997).

7.4.2.3 Non-Transparent "Polysilicon Emitter"

In an attempt to reduce or control the current gain in a SiGe-base bipolar transistor, sometimes designers intentionally introduce a thin Ge-containing layer within the single-crystalline emitter region of a polysilicon-emitter SiGe-base bipolar transistor (Huizing *et al.*, 2001). In this case, the thin Ge-containing layer creates a local potential well for holes, causing a significant increase in Auger recombination of electrons and holes within the single-crystalline emitter region. It results in a significant increase in base current. For such a transistor, even though a polysilicon layer is used to form a "polysilicon emitter," the single-crystalline part of the emitter is not transparent because of the large recombination in it. As a result, the conventional transparent-emitter model described in Section 6.2.2 for a polysilicon emitter does not apply. That is, Eqs. (7.43) and (7.45), which are derived based on the assumption that the single-crystalline emitter region is transparent, are no longer valid. Instead, the base current should be evaluated from Eqs. (6.35) and (6.36).

Reducing current gain leads to an increase in emitter delay time [see Eq. (8.16)]. Thus far, there is no reported data suggesting that adding a high-recombination region within the single-crystalline emitter region, or using any similar techniques for reducing current gain, will lead to a transistor of better performance. As a result, such non-transparent "polysilicon-emitter" devices will not be discussed any further.

7.4.2.4 Polycrystalline Silicon–Germanium Emitter

In some studies (Martinet *et al.*, 2002; Kunz *et al.*, 2002; Kunz *et al.*, 2003), polycrystalline silicon-germanium (polySiGe) instead of polysilicon is used to form the emitter in an attempt to reduce current gain in a SiGe-base bipolar transistor. The energy bandgap of a polySiGe layer is smaller than that of a polysilicon layer. The reduced bandgap increases

the injection of holes from the base into the polySiGe emitter region. In addition, the value of S_p for a polySiGe emitter could be quite different from that for a polysilicon emitter.

As discussed in Section 7.1, the polysilicon emitter was developed to overcome the limitation of the diffused emitter. The polysilicon emitter enables the scaling of bipolar transistors to base widths of less than 100 nm. Thin-base bipolar transistors using diffused emitters have excessively large and varying base currents, causing current gains to be too small and to have large variations. Thin-base transistors using polysilicon emitters do not have such problems. SiGe-base transistor designers often want to reduce current gain as a means to increase BV_{CEO} [cf. Eq. (6.152)]. Using the polySiGe emitter in place of the polysilicon emitter indeed leads to an increase in base current, hence smaller current gain and somewhat larger BV_{CEO}.

However, as pointed out in Section 6.2.3, it is important to recognize that current gain can be changed by changing the collector current, the base current, or both. Also, it is important to note that, compared to a Si-base bipolar transistor, the larger current gain in a polysilicon-emitter SiGe-base bipolar transistor is due entirely to an increase in the collector current, and not to any significant change in the base current. It will be shown in Section 7.4.6 that it is possible to reduce collector current, and hence current gain, of a SiGe-base transistor without affecting its transit time advantage over a Si-base transistor. This is accomplished by optimizing the Ge profile in the base.

Another effective approach to reduce collector current and current gain of a transistor is to reduce its intrinsic-base sheet resistivity [cf. Eq. (7.7)]. It will be shown in Chapter 8 that reducing base resistance leads to improved device and circuit performance. Therefore, if a smaller current gain is desired, a device designer should consider reducing the intrinsic-base sheet resistivity of the transistor. This can be accomplished easily by increasing the base doping concentration.

As pointed out earlier, reducing current gain leads to an increase in emitter delay time. Furthermore, there is no theory or experimental results to suggest that replacing a polysilicon emitter with a polySiGe emitter will lead to improved device speed. Therefore, we will not consider the polySiGe emitter any further.

7.4.3 Transistors Havivg a Trapezoidal Ge Distribution in the Base

Various Ge profiles have been analyzed and/or tested out experimentally by various groups (e.g., see Cressler *et al.*, 1993a, Harame *et al.*, 1995a,b, and Washio *et al.*, 2002). Here we focus on the trapezoidal Ge profile illustrated in Fig. 7.7 because close-form equations for the various transistor parameters can be readily obtained for it. The close-form equations enable us to discuss more clearly the device physics and operation, as well as device design optimization. Besides, a trapezoidal profile is more general than the simple triangular profile discussed in Section 7.4.1.

Even though a simple triangular Ge distribution may be the design target, the Ge profile in the quasineutral base at the end of the fabrication process is often more like a trapezoid than a triangle. For instance, if the Ge concentration is ramped down a bit

more slowly than intended during device fabrication, some Ge can be present in the cap region which is intended to be Ge-free. When that happens, the emitter–base junction will be located at a point where the Ge concentration is finite instead of zero. The resultant Ge distribution in the quasineutral base will have a trapezoidal profile instead of a triangular profile. In this case, a model for a trapezoidal Ge profile gives a more accurate description of the SiGe-base transistor than a model for a simple triangular Ge profile.

As illustrated in Fig. 7.7, for a given Ge distribution in the starting base layer of thickness W_{B0}, which includes a Ge-free cap layer of thickness W_{cap}, the quasineutral base width W_B is a function of the emitter depth x_{jE}, namely $W_B = W_{B0} - x_{jE}$. (Note that $x = W_{B0}$ is the location of the collector end of the quasineutral base and $x = x_{jE}$ is the emitter end of the quasineutral base. Whether the value of W_{B0} changes or not during device fabrication, the width of the quasineutral base is always given by $W_B = W_{B0} - x_{jE}$. For modern polysilicon-emitter SiGe-base transistors fabricated using emitter formation processes of low thermal budgets, W_{B0} usually changes less than x_{jE} during the device fabrication process.)

Figure 7.7 depicts the case of $x_{jE} > W_{cap}$, which means there is no residual Ge-free region in the final base layer. If we have $x_{jE} < W_{cap}$ instead, the final base layer would contain a residual Ge-free cap of thickness $W_{cap} - x_{jE}$. Here we want to extend the SiGe-base bipolar transistor model to include emitter depth as a parameter. With emitter depth included, the model can be used to evaluate the effect of emitter depth on device characteristics. We shall consider both the case of $x_{jE} > W_{cap}$ and the case of $x_{jE} < W_{cap}$.

7.4.3.1 Collector Current for a Trapezoidal Ge Distribution

The ratio of the collector current of a SiGe-base transistor to that of a Si-base transistor with the same boxlike base dopant distribution is given by Eq. (7.30). It can easily be adapted to include the effect of emitter depth by noting that the quasineutral base starts at $x = x_{jE}$ and ends at $x = W_{B0}$. From Eq. (7.30), we can write the collector current ratio as a function of emitter depth as

$$\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})} \approx \frac{\overline{\gamma}\overline{\eta}(W_{B0} - x_{jE})}{\int_{x_{jE}}^{W_{B0}} \exp\left[-\Delta E_{gB,SiGe}(x)/kT\right] dx}. \tag{7.49}$$

- *Case of Ge in the emitter* (i.e., $x_{jE} > W_{cap}$). This is the situation depicted in Fig. 7.7. The Ge distribution in the quasineutral base has a simple trapezoidal profile, with a base bandgap narrowing of ΔE_{g0be} at the emitter end of the base given by

$$\Delta E_{g0be}(x_{jE}) = \Delta E_{g0} + \left(\frac{x_{jE} - W_{cap}}{W_{B0} - W_{cap}}\right)(\Delta E_{g\max} - \Delta E_{g0}). \tag{7.50}$$

The base bandgap narrowing as a function of position in the base is given by

$$\Delta E_{gB,SiGe}(x) = \Delta E_{g0be}(x_{jE}) + \left(\frac{x - x_{jE}}{W_{B0} - x_{jE}}\right)[\Delta E_{g\max} - \Delta E_{g0be}(x_{jE})]. \tag{7.51}$$

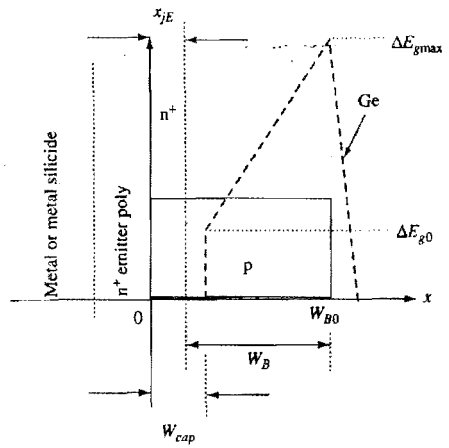


Figure 7.9. Schematic illustrating the emitter and base regions of a polysilicon-emitter SiGe-base bipolar transistor having the same base dopant and Ge profiles as in Fig. 7.7, but with $x_{jE} < W_{cap}$.

Substituting Eqs. (7.50) and (7.51) into Eq. (7.49), we obtain

$$\left.\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})}\right|_{x_{jE} > W_{cap}}$$

$$= \frac{\overline{\gamma}\overline{\eta}(W_{B0} - x_{jE})}{\exp\left[\dfrac{-\Delta E_{g0be}(x_{jE})}{kT}\right]\displaystyle\int_{x_{jE}}^{W_{B0}} \exp\left\{\frac{[\Delta E_{g\max} - \Delta E_{g0be}(x_{jE})](x_{jE} - x)}{(W_{B0} - x_{jE})kT}\right\} dx}$$

$$= \overline{\gamma}\overline{\eta}\left[\frac{\Delta E_{g\max} - \Delta E_{g0be}(x_{jE})}{kT}\right]\exp\left[\frac{\Delta E_{g0be}(x_{jE})}{kT}\right]$$

$$\times\left\{1 - \exp\left[\frac{\Delta E_{g0be}(x_{jE}) - \Delta E_{g\max}}{kT}\right]\right\}^{-1}. \tag{7.52}$$

- *Case of no Ge in the emitter* (i.e., $x_{jE} < W_{cap}$). When $x_{jE} < W_{cap}$, there is a residual Ge-free layer of thickness $W_{cap} - x_{jE}$ in the base. This is the situation depicted in Fig. 7.9. The base bandgap narrowing parameter is given by

$$\Delta E_{gB,SiGe}(x) = \Delta E_{g0} + \left(\frac{x - W_{cap}}{W_{B0} - W_{cap}}\right)(\Delta E_{g\max} - \Delta E_{g0}) \quad x > W_{cap}$$

$$= 0 \quad x < W_{cap}. \tag{7.53}$$

Substituting Eq. (7.53) into Eq. (7.49) we obtain

$$\left.\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})}\right|_{x_{jE}<W_{cap}} = \frac{\overline{\gamma\eta}(W_{B0}-x_{jE})}{W_{cap}-x_{jE}+\int_{W_{cap}}^{W_{B0}}\exp\left[-\Delta E_{gB,SiGe}(x)/kT\right]dx}$$

$$= \frac{\overline{\gamma\eta}}{\left[\frac{W_{cap}-x_{jE}}{W_{B0}-x_{jE}}\right]+\left[\frac{W_{B0}-W_{cap}}{W_{B0}-x_{jE}}\right]\left[\frac{kT}{\Delta E_{g\max}-\Delta E_{g0}}\right]\exp\left[\frac{-\Delta E_{g0}}{kT}\right]\left[1-\exp\left(\frac{\Delta E_{g0}-\Delta E_{g\max}}{kT}\right)\right]} \tag{7.54}$$

When $x_{jE} = W_{cap}$, Eq. (7.54) has the same form as Eq. (7.52), as it should. Also, when $x_{jE} = W_{cap}$ and $\Delta E_{g0} = 0$, Eq. (7.54) reduces to (7.32), as it should.

7.4.3.2 Early Voltage for a Trapezoidal Ge Distribution

The same procedures can be followed to obtain equations for the Early voltage ratio. The ratio of the Early voltage of a SiGe-base transistor to that of a Si-base transistor with the same boxlike base dopant distribution is given by Eq. (7.35). It can be adapted to include the effect of emitter depth, by noting that the quasineutral base starts at $x = x_{jE}$ and ends at $x = W_{B0}$. The result is

$$\frac{V_A(\text{SiGe}, x_{jE})}{V_A(\text{Si}, x_{jE})} \approx \frac{\exp\left[\Delta E_{gB,SiGe}(W_{B0})/kT\right]}{W_{B0}-x_{jE}} \times \int_{x_{jE}}^{W_{B0}}\exp\left[-\Delta E_{gB,SiGe}(x)/kT\right]dx. \tag{7.55}$$

- *Case of Ge in the emitter* (i.e., $x_{jE} > W_{cap}$). For the case of Ge in the emitter, substituting Eqs. (7.50) and (7.51) into Eq. (7.55), we obtain

$$\left.\frac{V_A(\text{SiGe}, x_{jE})}{V_A(\text{Si}, x_{jE})}\right|_{x_{jE}>W_{cap}} = \left[\frac{kT}{\Delta E_{g\max}-\Delta E_{g0be}(x_{jE})}\right] \times \left\{\exp\left[\frac{\Delta E_{g\max}-\Delta E_{g0be}(x_{jE})}{kT}\right]-1\right\}. \tag{7.56}$$

It should be noted that the Early voltage ratio in this case depends on the bandgap energy difference $[\Delta E_{g\max} - \Delta E_{g0be}(x_{jE})]$ across the quasineutral base layer. Equation (7.56) has the same form as Eq. (7.36), where $\Delta E_{g0be}(x_{jE}) = 0$.

- *Case of no Ge in the emitter* (i.e., $x_{jE} < W_{cap}$). For the case with no Ge in the emitter, there is a residual Ge-free layer of thickness $W_{cap} - x_{jE}$ in the base. Substituting Eq. (7.53) into Eq. (7.55), we obtain

$$\left.\frac{V_A(\text{SiGe}, x_{jE})}{V_A(\text{Si}, x_{jE})}\right|_{x_{jE}<W_{cap}} = \left(\frac{W_{cap}-x_{jE}}{W_{B0}-x_{jE}}\right)\exp\left(\frac{\Delta E_{g\max}}{kT}\right) + \left(\frac{W_{B0}-W_{cap}}{W_{B0}-x_{jE}}\right)\left(\frac{kT}{\Delta E_{g\max}-\Delta E_{g0}}\right) \times \left\{\exp\left[\frac{\Delta E_{g\max}-\Delta E_{g0}}{kT}\right]-1\right\}. \tag{7.57}$$

When $x_{jE} = W_{cap}$, Eqs. (7.56) and (7.57) have the same form, as they should. Also, when $x_{jE} = W_{cap}$ and $\Delta E_{g0} = 0$, Eqs. (7.57) reduces to Eq. (7.36), as it should.

7.4.3.3 Base Transit Time for a Trapezoidal Ge Distribution

The base transit time ratio can be derived in the same manner. The ratio of the base transit time of a SiGe-base transistor to that of a Si-base transistor having the same boxlike base dopant distribution is given by Eq. (7.40). It can be adapted to include the effect of emitter depth by noting that the quasineutral base starts at $x = x_{jE}$ and ends at W_{B0}. The result is

$$\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})} = \frac{2}{\bar{\eta}(W_{B0}-x_{jE})^2}\int_{x_{jE}}^{W_{B0}}\exp\left[\Delta E_{gB,SiGe}(x)/kT\right] \times \int_{x}^{W_{B0}}\exp\left[-\Delta E_{gB,SiGe}(x')/kT\right]dx'dx. \tag{7.58}$$

- *Case of Ge in the emitter* (i.e., $x_{jE} > W_{cap}$). For the case of Ge in the emitter, substituting Eqs. (7.50) and (7.51) into Eq. (7.58), we obtain

$$\left.\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})}\right|_{x_{jE}>W_{cap}} = \frac{2}{\bar{\eta}}\left[\frac{kT}{\Delta E_{g\max}-\Delta E_{g0be}(x_{jE})}\right]-\frac{2}{\bar{\eta}}\left[\frac{kT}{\Delta E_{g\max}-\Delta E_{g0be}(x_{jE})}\right]^2 \times \left\{1-\exp\left[\frac{\Delta E_{g0be}(x_{jE})-\Delta E_{g\max}}{kT}\right]\right\}. \tag{7.59}$$

It should be noted that, just like the Early voltage ratio in Eq. (7.56), the transit time ratio depends on the bandgap energy difference $[\Delta E_{g\max} - \Delta E_{g0be}(x_{jE})]$. Equation (7.59) reduces to Eq. (7.41) when $\Delta E_{g0be}(x_{jE}) = 0$, as expected.

- *Case of no Ge in the emitter* (i.e., $x_{jE} < W_{cap}$). For the case of no Ge in the emitter, there is a residual Ge-free layer of thickness $W_{cap} - x_{jE}$ in the base. Substituting Eq. (7.53) into Eq. (7.58), we obtain

$$\left.\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})}\right|_{x_{jE}<W_{cap}} = \frac{1}{\bar{\eta}}\frac{(W_{cap}-x_{jE})^2}{(W_{B0}-x_{jE})^2}+\frac{2}{\bar{\eta}}\frac{(W_{B0}-W_{cap})^2}{(W_{B0}-x_{jE})^2}\left(\frac{kT}{\Delta E_{g\max}-\Delta E_{g0}}\right) + \frac{2}{\bar{\eta}}\left(\frac{W_{cap}-x_{jE}}{W_{B0}-x_{jE}}\right)\left(\frac{W_{B0}-W_{cap}}{W_{B0}-x_{jE}}\right)\left(\frac{kT}{\Delta E_{g\max}-\Delta E_{g0}}\right)$$

$$\times \left\{1 - \exp\left[\frac{\Delta E_{g0} - \Delta E_{g\max}}{kT}\right]\right\} \exp\left(\frac{-\Delta E_{g0}}{kT}\right)$$

$$-\frac{2}{\bar{\eta}} \frac{(W_{B0} - W_{cap})^2}{(W_{B0} - x_{jE})^2} \left(\frac{kT}{\Delta E_{g\max} - \Delta E_{g0}}\right)^2 \left\{1 - \exp\left[\frac{\Delta E_{g0} - \Delta E_{g\max}}{kT}\right]\right\}. \quad (7.60)$$

Note that Eq. (7.60) depends on the energy difference ($\Delta E_{g\max} - \Delta E_{g0}$) as well as on ΔE_{g0}. This should be contrasted with the case of Ge in the emitter above [Eq. (7.59)]. The added dependence on ΔE_{g0} can be used to tailor the Ge profile to further improve base transit time. This will be discussed later in Section 7.4.6.2.

7.4.4 Transistors Having a Constant Ge Distribution in the Base

So far our discussions have focused on SiGe-base bipolar transistors having a graded base bandgap. In the literature, most reported SiGe-base transistors are of the graded-base-bandgap type. However, SiGe-base transistors having a spatially constant base bandgap, corresponding to a spatially constant Ge distribution in the quasineutral base, are also used quite widely. A spatially constant Ge distribution is equivalent to setting $\Delta E_{g0} = \Delta E_{g\max}$ in Fig. 7.7.

A cursory examination of the current ratio in Eq. (7.49) suggests that a spatially constant Ge distribution in the base leads to a J_{C0} that is larger by approximately a factor of $\exp(\Delta E_{g0}/kT)$ than a Si-base transistor having the same base width and base dopant distribution. The Early voltage ratio in Eq. (7.55) suggests that there should be no improvement in Early voltage. Also, Eq. (7.58) suggests that there should be no improvement in base transit time other than indirectly through the factor $\bar{\eta}$. Yet, in the literature, there are ample experimental data showing SiGe-base transistors having supposedly constant Ge distribution in the base to be superior to Si-base transistors in both Early voltage and base transit time (e.g., see Schüppen and Dietrich, 1995, Schüpper *et al.*, 1996, Hobart *et al.*, 1995, and Deixler *et al.*, 2001). In this section, we extend the models developed in the previous sections to examine the properties of a constant-Ge SiGe-base transistor more closely.

- *Case of Ge in the emitter* (i.e., $x_{jE} > W_{cap}$). As long as the emitter is sufficiently deep so that the emitter–base junction is located in the constant-Ge region, the SiGe-base transistor has a narrowed energy bandgap that is spatially constant across its entire quasineutral base layer. The emitter and base regions are as illustrated in Fig. 7.10. From Eq. (7.52), we have

$$\left.\frac{J_{C0}(\mathrm{SiGe}, x_{jE})}{J_{C0}(\mathrm{Si}, x_{jE})}\right|_{x_{jE} > W_{cap}} = \bar{\gamma}\bar{\eta} \exp(\Delta E_{g0}/kT). \quad (7.61)$$

From Eq. (7.56), we have

$$\left.\frac{V_A(\mathrm{SiGe}, x_{jE})}{V_A(\mathrm{Si}, x_{jE})}\right|_{x_{jE} > W_{cap}} = 1, \quad (7.62)$$

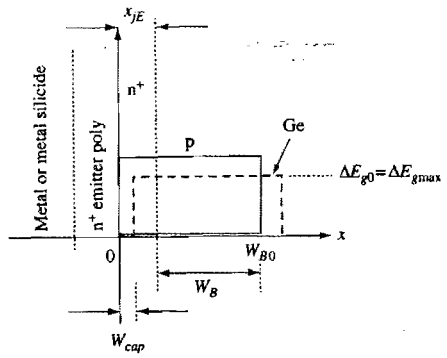


Figure 7.10. Schematic illustrating the emitter and base regions of a polysilicon-emitter SiGe-base bipolar transistor having a constant Ge distribution in the base. The emitter depth x_{jE} is assumed to be larger than the thickness W_{cap} of the starting Ge-free layer.

and from Eq. (7.59), we have

$$\left.\frac{t_B(\mathrm{SiGe}, x_{jE})}{t_B(\mathrm{Si}, x_{jE})}\right|_{x_{jE} > W_{cap}} = \frac{1}{\bar{\eta}}. \quad (7.63)$$

That is, compared to a Si-base transistor, the SiGe-base transistor has higher collector current and current gain, by about a factor of $\exp(\Delta E_{g0}/kT)$, but the same Early voltage.

The base transit time is reduced by only a factor of $1/\bar{\eta}$. Also, these ratios are independent of emitter depth as long as $x_{jE} > W_{cap}$.

- *Case of no Ge in the emitter* (i.e., $x_{jE} < W_{cap}$). When the emitter depth is smaller than the starting Ge-free cap thickness, the emitter and base regions are as illustrated in Fig. 7.11. The energy bandgap is no longer spatially constant across the entire quasineutral base. Instead, the bandgap is larger at the emitter end of the base where there is no Ge. The corresponding collector current ratio, Early voltage ratio, and base transit time ratio can be obtained from Eqs. (7.54), (7.57), and (7.60), respectively. They are

$$\left.\frac{J_{C0}(\mathrm{SiGe}, x_{jE})}{J_{C0}(\mathrm{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} = \bar{\gamma}\bar{\eta} \left\{\left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right) + \left(\frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}\right) \exp(-\Delta E_{g0}/kT)\right\}^{-1}, \quad (7.64)$$

$$\left.\frac{V_A(\mathrm{SiGe}, x_{jE})}{V_A(\mathrm{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} = \left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right) \exp\left(\frac{\Delta E_{g0}}{kT}\right) + \left(\frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}\right), \quad (7.65)$$

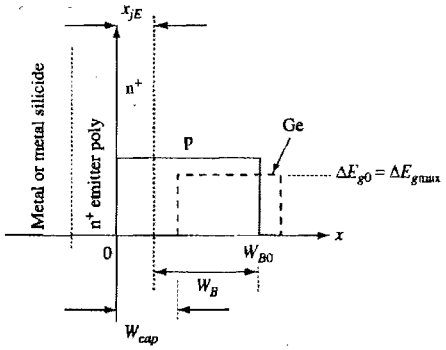


Figure 7.11. Schematic illustrating the emitter and base regions of a polysilicon-emitter SiGe-base bipolar transistor having a constant Ge distribution in the base. The emitter depth x_{jE} is assumed to be smaller than the thickness W_{cap} of the starting Ge-free layer, resulting in a small region of thickness $W_{cap} - x_{jE}$ near the emitter end without Ge.

and

$$\left.\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} = \frac{1}{\bar{\eta}}\left\{\left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right)^2 + \left(\frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}\right)^2 + 2\left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right)\left(\frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}\right)\exp(-\Delta E_{g0}/kT)\right\}. \tag{7.66}$$

Note that these ratios are functions of x_{jE}. $W_{B0} - x_{jE}$ is the base width and $(W_{cap} - x_{jE})/(W_{B0} - x_{jE})$ is the fraction of the quasineutral base with no Ge at all. Figure 7.12 is a plot of Eq. (7.66) as a function of $(W_{cap} - x_{jE})/(W_{B0} - x_{jE})$ for several values of $\Delta E_{g0}/kT$. It clearly shows that as long as there is a residual Ge-free region in the emitter end of the quasineutral base, there is improvement in base transit time over a Si-base transistor having the same base width and base dopant distribution. In fact, for a given value of $\Delta E_{g0}/kT$, which is equal to $\Delta E_{gmax}/kT$, the maximum improvement factor, which occurs at $(W_{cap} - x_{jE})/(W_{B0} - x_{jE}) \sim 0.5$, is quite comparable to that for a linearly graded Ge distribution (see Fig. 7.6). It is left as an exercise (Ex. 7.9) to the reader to show that there can also be significant improvement in current gain and Early voltage over a Si-base transistor.

It is of interest to note that as long as the Ge concentration is sufficiently large so that $\exp(-\Delta E_{g0}/kT)$ is negligible compared to $(W_{cap} - x_{jE})/(W_{B0} - x_{jE})$, these ratios approach the values of

$$\left.\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} \to \bar{\gamma}\bar{\eta}\left(\frac{W_{B0} - x_{jE}}{W_{cap} - x_{jE}}\right), \tag{7.67}$$

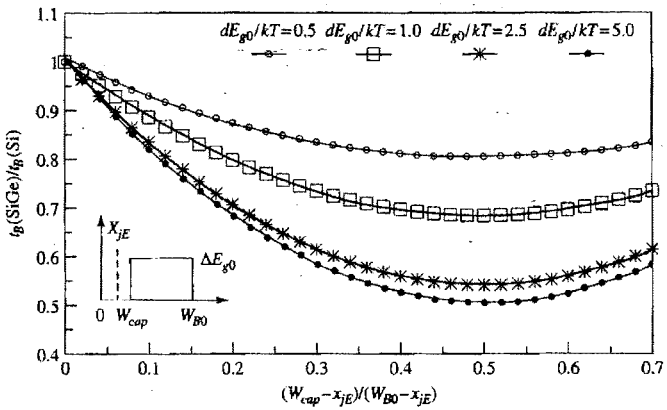


Figure 7.12. The base transit time ratio, Eq. (7.66), as a function of $(W_{cap} - x_{jE})/(W_{B0} - x_{jE})$ with $\Delta E_{g0}/kT$ as a parameter.

$$\left.\frac{V_A(\text{SiGe}, x_{jE})}{V_A(\text{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} \to \left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right)\exp\left(\frac{\Delta E_{g0}}{kT}\right), \tag{7.68}$$

and

$$\left.\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} \to \frac{1}{\bar{\eta}}\left\{\left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right)^2 + \left(\frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}\right)^2\right\}. \tag{7.69}$$

That is, both the collector current ratio, hence the current gain ratio, and the base transit time ratio become independent of ΔE_{g0} for $\Delta E_{g0}/kT \gg 1$, while the Early voltage increases exponentially with $\Delta E_{g0}/kT$. That the base transit time ratio becomes less sensitive to ΔE_{g0} for large $\Delta E_{g0}/kT$ is also evident in Fig. 7.12.

Equation (7.69) has a minimum value of $0.5/\bar{\eta}$ at $(W_{cap} - x_{jE})/(W_{B0} - x_{jE}) = 0.5$. The corresponding value for Eq. (7.67) is $2\bar{\gamma}\bar{\eta}$. For a typical Ge concentration of 20%, γ is about 0.4 (Prinz *et al.*, 1989), and η is about 1.4 (Kay and Tang, 1991). Since there is no Ge at the emitter end of the base, the corresponding values averaged over the entire quasineutral base layer should be somewhat larger than 0.4 for $\bar{\gamma}$ and somewhat less than 1.4 for $\bar{\eta}$.

So far, we have assumed that the Ge distribution drops abruptly to zero at $x = W_{cap}$. In practice this never happens, either by design or by the fact that it is impossible to realize a truly abrupt rise or fall in a Ge distribution. (See Washio *et al.*, 2002, for an example of a realistic Ge profile that is designed to ramp up and down abruptly.) Instead of ramping down at an infinite rate to zero at the emitter end, a Ge distribution can be ramped down only at some finite rate. A model for a SiGe-base transistor having a base-region Ge distribution that ramps up from zero concentration at the emitter end to some constant concentration some distance towards the collector end is developed in Ex. 7.10. Whether the Ge distribution ramps up at some finite rate as described in Ex. 7.10, or abruptly as

illustrated in Fig. 7.11, the device characteristics are qualitatively and quantitatively quite similar to those of a transistor having a simple triangular Ge distribution. Therefore, ***as long as there is some region of zero or relatively low Ge concentration at the emitter end of the quasineutral base, an otherwise constant-Ge SiGe-base transistor behaves like a SiGe-base transistor having a graded Ge distribution*** in that the transistor has larger current gain, larger Early voltage, but smaller base transit time compared to a Si-base transistor having the same polysilicon emitter, base width and base dopant distribution. This explains why "constant-Ge" SiGe-base transistors usually show higher speed, higher current gains and larger Early voltages than Si-base transistors.

7.4.5 Effect of Emitter Depth Variation on Device Characteristics

It is apparent from the previous discussions that for a given starting Ge distribution and starting base layer thickness, the final device characteristics depend on the depth of the single-crystalline n^+ emitter region. In a typical SiGe-base bipolar fabrication process, the transistors can have somewhat different emitter depths due to subtle or not so subtle process variations. The emitter depth variation within a wafer should be small, but the variation from wafer to wafer and from run to run can be appreciable. In this section, we use the models developed in the previous sections to examine the effect of emitter depth variation on device characteristics.

In modeling the effect of emitter depth variation, it is desirable to select a reference emitter depth and then compare the changes in device characteristics as the emitter depth is varied around the reference. In designing a SiGe-base process, often the goal is to choose a combination of starting Ge-free cap thickness and an emitter drive-in thermal cycle to obtain $x_{jE} = W_{cap}$, i.e., to result in no Ge in the emitter and no residue Ge-free region in the final quasineutral base. Due to process variation, there are always some transistors with $x_{jE} > W_{cap}$ and some with $x_{jE} < W_{cap}$. Therefore, it is of interest to examine how device characteristics vary around the reference emitter depth of $x_{jE} = W_{cap}$ (Ning, 2003a).

- *Effect on collector current and current gain.* The effects of emitter depth variation on collector current and current gain are the same. This is because any change in current gain is caused by a change in the collector current and not by a change in the base current, as discussed earlier in Section 7.4.2. Therefore, we shall refer to collector current variation and current gain variation interchangeably when there is no confusion.

  Let $J_{C0}(\text{SiGe}, x_{jE})$ and $J_{C0}(\text{SiGe}, W_{cap})$ denote the saturated collector current densities of a SiGe-base transistor when its emitter depth equals x_{jE} and when its emitter depth equals W_{cap}, respectively. A plot of the ratio $J_{C0}(\text{SiGe}, x_{jE})/J_{C0}(\text{SiGe}, W_{cap})$ as a function of $x_{jE} - W_{cap}$ gives the relative change of the collector current around the reference point of $x_{jE} = W_{cap}$. This current ratio can be written in the form

$$\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{SiGe}, W_{cap})} = \frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})} \frac{J_{C0}(\text{Si}, x_{jE})}{J_{C0}(\text{Si}, W_{cap})} \left(\frac{J_{C0}(\text{SiGe}, W_{cap})}{J_{C0}(\text{Si}, W_{cap})}\right)^{-1} \quad (7.70)$$

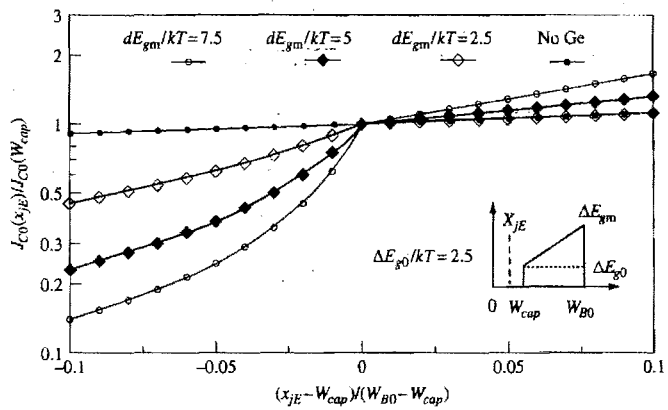


Figure 7.13 Relative collector current variation as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for a trapezoidal Ge profile with $\Delta E_{g0}/kT = 2.5$ and $\Delta E_{gmax}/kT$ as a parameter.

The ratios $J_{C0}(\text{SiGe}, x_{jE})/J_{C0}(\text{Si}, x_{jE})$ and $J_{C0}(\text{SiGe}, W_{cap})/J_{C0}(\text{Si}, W_{cap})$ can be obtained from Eqs. (7.52) and (7.54). Also, it can be inferred readily from Eq. (7.29) (also see Section 6.2.1) that the collector current ratio corresponding to Eq. (7.70) for a Si-base transistor with a boxlike base dopant distribution is

$$\frac{J_{C0}(\text{Si}, x_{jE})}{J_{C0}(\text{Si}, W_{cap})} = \frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}. \quad (7.71)$$

For our reference design point with $x_{jE} = W_{cap}$, the base width is $W_{B0} - x_{jE} = W_{B0} - W_{cap}$. Therefore, $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ is the emitter depth variation normalized to the reference base width. Figure 7.13 is a plot of Eq. (7.70) as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for a trapezoidal Ge distribution with $\Delta E_{g0}/kT = 2.5$, for several values of $\Delta E_{gmax}/kT$. $(x_{jE} - W_{cap}) > 0$ means that there is Ge in the emitter, and $(x_{jE} - W_{cap}) < 0$ means that there is a residual Ge-free layer in the quasineutral base. Figure 7.14 is a similar plot, but for $\Delta E_{g0}/kT = 0$. In this case, the reference with $x_{jE} = W_{cap}$ corresponds to a transistor having a simple triangular Ge distribution.

Figures 7.13 and 7.14 show that, in general, collector current variation increases with increase in ΔE_{gmax} and/or in ΔE_{g0}. For a given $\Delta E_{g0} > 0$, collector current variation is much larger when $x_{jE} < W_{cap}$ than when $x_{jE} > W_{cap}$. For $\Delta E_{g0} = 0$, collector current variation is about the same for $x_{jE} < W_{cap}$ and $x_{jE} > W_{cap}$. However, the collector current increases approximately as $\exp(\Delta E_{g0}/kT)$, as expected from Eqs. (7.52) and (7.54). Thus, reducing ΔE_{g0} will reduce current gain variation for $x_{jE} < W_{cap}$, but it will also reduce the magnitude of the current gain by a large amount. Optimizing the Ge profile to minimize current gain sensitivity to emitter depth variation will be discussed later in Section 7.4.6.

- *Effect on Early voltage.* The corresponding ratio for Early voltage is

$$\frac{V_A(\text{SiGe}, x_{jE})}{V_A(\text{SiGe}, W_{cap})} = \frac{V_A(\text{SiGe}, x_{jE})}{V_A(\text{Si}, x_{jE})} \frac{V_A(\text{Si}, x_{jE})}{V_A(\text{Si}, W_{cap})} \left(\frac{V_A(\text{SiGe}, W_{cap})}{V_A(\text{Si}, W_{cap})}\right)^{-1}, \quad (7.72)$$

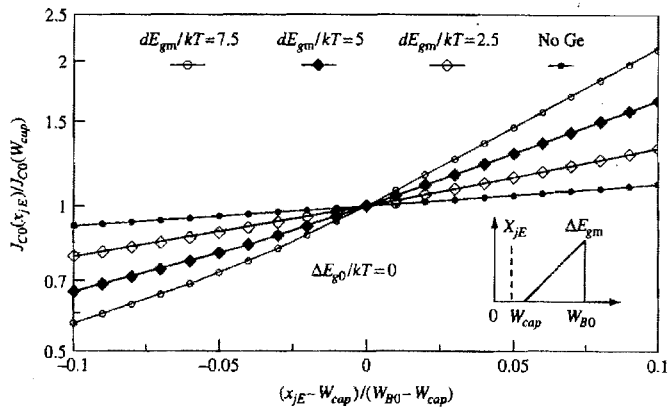


Figure 7.14. A similar plot as Fig. 7.13, but with $\Delta E_{g0}/kT = 0$.

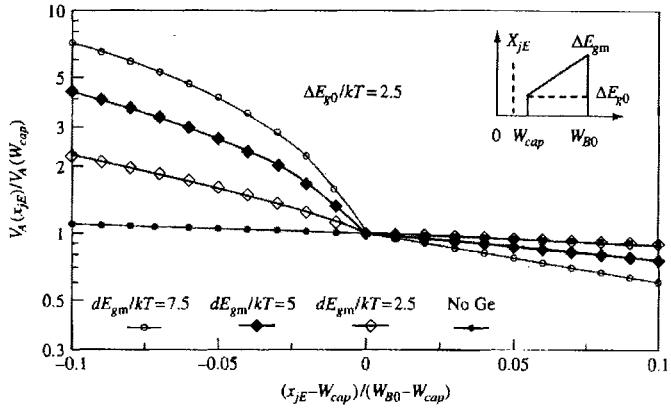


Figure 7.15 Relative Early voltage variation as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for a trapezoidal Ge profile with $\Delta E_{g0}/kT = 2.5$ and $\Delta E_{gmax}/kT$ as a parameter, the same as in Fig. 7.13.

where

$$\frac{V_A(\text{Si}, x_{jE})}{V_A(\text{Si}, W_{cap})} = \frac{W_{B0} - x_{jE}}{W_{B0} - W_{cap}} \tag{7.73}$$

is the Early voltage ratio for a Si-base bipolar transistor having a boxlike base dopant distribution (see Section 6.3.2). Figure 7.15 is a plot of Eq. (7.72) as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for the same trapezoidal Ge distribution as in Fig. 7.13. When there is Ge in the emitter, the Early voltage is not a sensitive function of emitter depth, decreasing only slowly as the emitter depth increases. However, when there is a residual Ge-free layer in the base, the Early voltage is a strong function of emitter

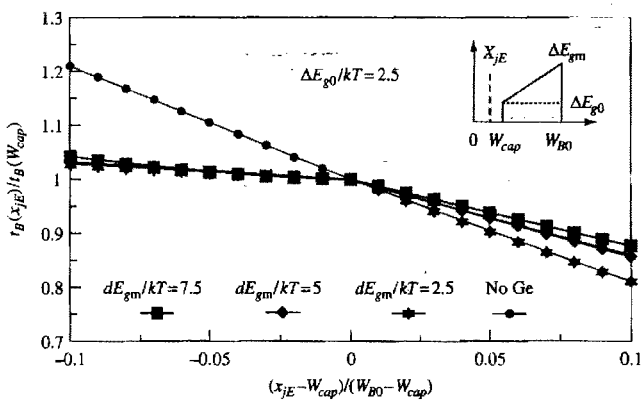


Figure 7.16 Relative base transit time variation as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for a trapezoidal Ge profile with $\Delta E_{g0}/kT = 2.5$ and $\Delta E_{gmax}/kT$ as a parameter, the same as in Fig. 7.13.

depth, due primarily to the first term in Eq. (7.57) which contains a large multiplying factor $\exp(\Delta E_{gmax}/kT)$.

- *Effect on base transit time.* In a similar manner, we can write the ratio of the base transit time as a function of x_{jE} to that at $x_{jE} = W_{cap}$ as

$$\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{SiGe}, W_{cap})} = \frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})} \frac{t_B(\text{Si}, x_{jE})}{t_B(\text{Si}, W_{cap})} \left(\frac{t_B(\text{SiGe}, W_{cap})}{t_B(\text{Si}, W_{cap})}\right)^{-1}, \tag{7.74}$$

where

$$\frac{t_B(\text{Si}, x_{jE})}{t_B(\text{Si}, W_{cap})} = \frac{(W_{B0} - x_{jE})^2}{(W_{B0} - W_{cap})^2} \tag{7.75}$$

is the base transit time ratio for a Si-base bipolar transistor with a boxlike base doping profile [see Eq. (7.24)]. Figure 7.16 is a plot of Eq. (7.74) as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for the same trapezoidal Ge profile as in Fig. 7.13.

For a Si-base transistor (the curve corresponding to no Ge in Fig. 7.16), the base transit time is proportional to W_B^2. A change of the base width by 10% results in about 20% change in base transit time.

Let us consider the region of $x_{jE} < W_{cap}$ in Fig. 7.16. This is the region where an increase in base width is caused by the emitter being shallower than intended (our reference design point is for $x_{jE} = W_{cap}$). The base width is increased by "adding" a thin Ge-free layer to the top part of the quasineutral base. Figure 7.16 shows that adding a thin p-type Ge-free layer to the top of the quasineutral base of a SiGe-base transistor increases the base transit time just a small amount, much less than anticipated from the W_B^2 dependence of the base transit time of a Si-base transistor.

Next let us consider the region of $x_{jE} > W_{cap}$ in Fig. 7.16, where a decrease in base width is caused by the emitter being deeper than intended. In this region, a decrease in

base width is accompanied by an increase in the Ge concentration at the emitter end of the quasineutral base layer, i.e., $\Delta E_{g0be}(x_{jE})$ increases as W_B decreases or as x_{jE} increases. In this case, the base transit time of a SiGe-base transistor still decreases more slowly with base width than a Si-base transistor.

The net is that the base transit time of a SiGe-base transistor is less sensitive to base width variation than a Si-base transistor. In particular, the base transit time of a SiGe-base transistor is relatively insensitive to increase in base width caused by the emitter depth being smaller than intended, particularly when the emitter depth is smaller than the starting Ge-free cap thickness. These results suggest that it is possible to optimize the Ge distribution to further reduce base transit time. This will be illustrated later in Section 7.4.6.2.

7.4.6 Some Optimal Ge Profiles

In this section, we apply the models developed in the previous sections to discuss tailoring the Ge profile in the quasineutral base for optimal or improved device characteristics. We first consider it from a current gain perspective, and then from a base transit time perspective.

7.4.6.1 Ge Profile from Current Gain Perspective

SiGe-base transistor designers often find current gains too large and/or varying too much among transistors. Compared to a Si-base transistor, the larger current gain in a SiGe-base transistor is caused by an increase in collector current, not by a decrease in base current. Also, as discussed in Section 7.4.2, it is preferable to reduce the current gain of a SiGe-base transistor by reducing its collector current than by increasing its base current.

- *Reducing current gain without degrading base transit time.* The most effective way to reduce collector current is to reduce the amount of bandgap narrowing at the emitter end of the quasineutral base layer. To see this, let us consider the case depicted in Fig. 7.7. The transistor has a trapezoidal Ge distribution with an emitter depth greater than its starting Ge-free cap thickness. The corresponding transit time ratio is given by Eq. (7.59), which shows that the base transit time is a function of the energy difference $[\Delta E_{gmax} - \Delta E_{g0be}(x_{jE})]$ across the quasineutral base layer. For this transistor the collector current ratio is given by Eq. (7.52), which shows that for a given energy difference $[\Delta E_{gmax} - \Delta E_{g0be}(x_{jE})]$, the collector current increases exponentially with increase in $\Delta E_{g0be}(x_{jE})$. Therefore, if we reduce $\Delta E_{g0be}(x_{jE})$ but keep $[\Delta E_{gmax} - \Delta E_{g0be}(x_{jE})]$ constant, we can reduce collector current, and hence current gain, significantly without affecting base transit time. Reducing $\Delta E_{g0be}(x_{jE})$ while keeping $[\Delta E_{gmax} - \Delta E_{g0be}(x_{jE})]$ constant means that ΔE_{gmax} is reduced by the same amount.
- *Minimizing sensitivity of current gain to emitter depth variation.* Designers often want to have a current gain that does not vary much with emitter depth. This can be accomplished by using the Ge distribution illustrated in Fig. 7.17 where the emitter–base junction is *confined* to a constant-Ge region (Oda *et al.*, 1997; Ansley *et al.*, 1998; Niu *et al.*, 2003). The model discussed in Section 7.4.3.1 for the case of no Ge in the

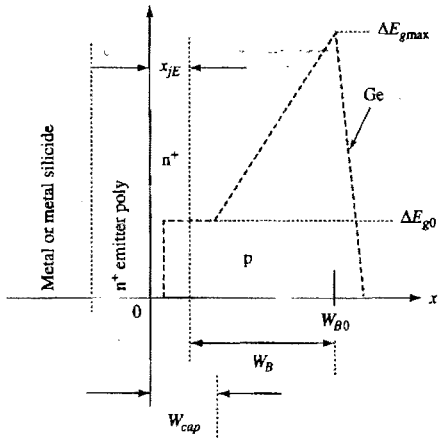


Figure 7.17. Schematic illustrating the emitter and base regions of a polysilicon-emitter SiGe-base bipolar transistor having a Ge distribution that makes the collector current less sensitive to emitter depth variation. For simplicity, the base dopant distribution is not shown. The emitter–base junction is confined to a region of finite but constant Ge concentration.

emitter can be readily extended to this case. Instead of a Ge-free cap, we have a constant-Ge region. In this case, the base bandgap narrowing parameter is given by

$$\Delta E_{gB,SiGe}(x) = \Delta E_{g0} + \left(\frac{x - W_{cap}}{W_{B0} - W_{cap}}\right)(\Delta E_{g\,max} - \Delta E_{g0}) \quad x > W_{cap}$$
$$= \Delta E_{g0} \quad x < W_{cap}. \qquad (7.76)$$

It is left as an exercise (Ex. 7.11) for the reader to show that the collector current ratio in this case is

$$\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})} = \frac{\overline{\gamma\eta}\,\exp(\Delta E_{g0}/kT)}{\left(\frac{W_{cap} - x_{jE}}{W_{B0} - x_{jE}}\right) + \left(\frac{W_{B0} - W_{cap}}{W_{B0} - x_{jE}}\right)\left(\frac{kT}{\Delta E_{g\,max} - \Delta E_{g0}}\right)\left[1 - \exp\left(\frac{\Delta E_{g0} - \Delta E_{g\,max}}{kT}\right)\right]} \qquad (7.77)$$

The collector current improvement factors of a SiGe-base transistor relative to a Si-base transistor for three Ge profiles are compared in Fig. 7.18 as a function of emitter depth variation, using $\overline{\gamma\eta} = 1$. Some insights into the dependence of collector current on Ge distribution can be inferred from Fig. 7.18. First, a high Ge concentration at or near the emitter–base junction leads to large collector current and current gain. Second, a Ge distribution that ramps down steeply or abruptly towards the emitter–base junction, as depicted by the "Ge-free cap" case in the figure, causes the collector current to be sensitive to emitter depth variation. This is probably the main reason why large current-gain variations are often observed in SiGe-base bipolar transistors.

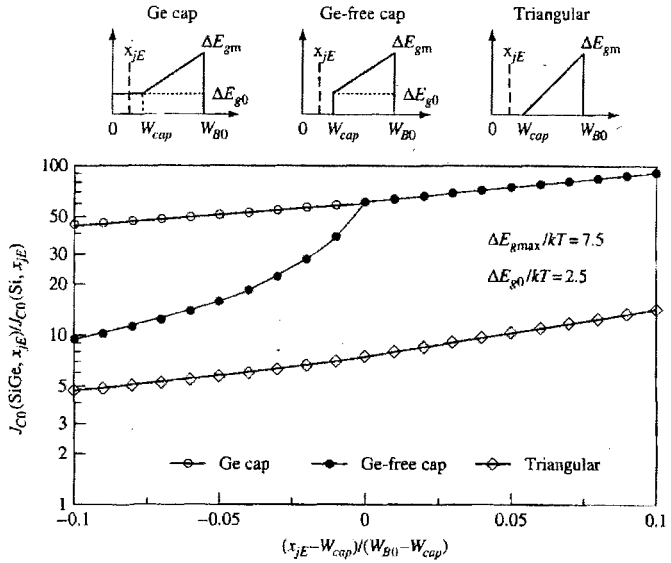


Figure 7.18. Relative improvement factors in collector current as a function of $(x_{jE} - W_{cap})/(W_{B0} - W_{cap})$ for three Ge profiles. $\overline{\gamma\eta} = 1$ assumed.

7.4.6.2 Ge Profile from Base Transit Time Perspective

The discussion in Section 7.4.5 suggests that it may be possible to tailor the Ge distribution in the quasineutral base to obtain a base transit time that is smaller than that given by the simple triangular Ge profile. Here we examine the dependence of base transit time on Ge distribution in greater detail, using the models developed for a trapezoidal Ge distribution shown in Fig. 7.7.

- *Dependence on* ΔE_{g0}. Let us consider the case of $x_{jE} = W_{cap}$. This is the simple trapezoidal Ge profile that has been studied quite extensively in the literature. In this case, the base transit time ratio is given by Eq. (7.59) by setting $\Delta E_{g0be}(x_{jE}) = \Delta E_{g0}$. The base transit time depends on the bandgap energy difference $(\Delta E_{gmax} - \Delta E_{g0})$. For a given ΔE_{gmax}, changing ΔE_{g0} changes this energy difference, and hence the base transit time. Figure 7.19 is a plot of the ratio $t_B(\Delta E_{g0}/kT)/t_B(\Delta E_{g0}/kT = 0)$ as a function of $\Delta E_{g0}/kT$ for $\Delta E_{gmax}/kT = 7.5$. It shows that simply increasing ΔE_{g0} at constant ΔE_{gmax} increases base transit time. Therefore, from a base transit time point of view, the simple triangular Ge profile (i.e. with $\Delta E_{g0} = 0$) is preferred.
- *Case of Ge not ramped down exactly as desired.* In forming a SiGe base by an epitaxial growth process, typically the Ge concentration is ramped up rapidly from zero to some peak concentration prior to the introduction of boron impurities, i.e., prior to starting the growth of the intrinsic-base layer. As the intrinsic-base layer is grown, the Ge concentration is ramped down at some rate to achieve the desired

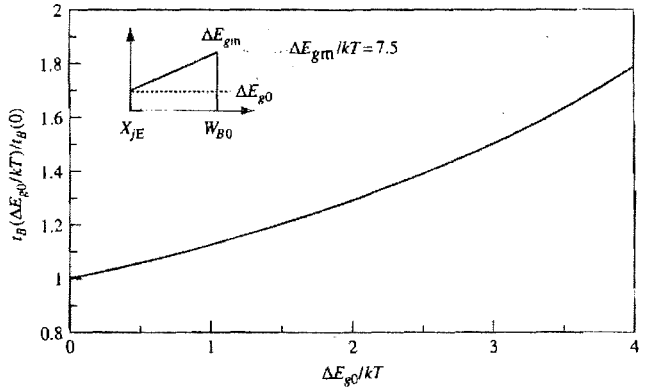


Figure 7.19. Base transit time for a trapezoidal Ge distribution as a function of $\Delta E_{g0}/kT$ for $\Delta E_{gmax}/kT = 7.5$, relative to the base transit time at $\Delta E_{g0}/kT = 0$. The base width is kept constant.

bandgap narrowing profile across the intrinsic base. To achieve a linearly graded base bandgap, the Ge concentration is ramped down at a constant rate as the intrinsic-base layer is grown such that the Ge concentration reaches zero at the emitter end of the quasineutral base.

In practice, it is difficult to achieve a truly linearly graded bandgap across the quasineutral base layer. If the Ge concentration is ramped down at a rate more slowly than intended, it will result in a finite, instead of zero, Ge concentration at the emitter end of the quasineutral base, i.e., it will result in $\Delta E_{g0be}(x_{jE}) > 0$. When that happens, the base transit time is degraded, or not improved over a Si-base transistor by as much as intended, as demonstrated in Fig. 7.19. And, as discussed in Section 7.4.6.1, when $\Delta E_{g0be}(x_{jE})$ is larger than intended, the collector current and current gain are also larger than intended.

Next, let us consider the case when the Ge concentration is ramped down at a rate faster than intended. Let us assume that the target design is to have a simple triangular Ge distribution with $x_{jE} = W_{cap}$, $\Delta E_{g0} = 0$, and some desired valued of ΔE_{gmax}. If the Ge concentration is ramped down at a rate faster than intended during growth of the base layer, there will be a finite region of the quasineutral base at the emitter end with no Ge at all. This is the case of "no Ge in emitter" described in Section 7.4.3.3. The base transit time ratio can be obtained from Eq. (7.60) by setting $\Delta E_{g0} = 0$. The quasineutral base width is $W_{B0} - x_{jE}$ and the thickness of portion of the quasineutral base having no Ge is $W_{cap} - x_{jE}$. Figure 7.20 is a plot of the relative change of base transit time as a function of $(W_{cap} - x_{jE})/(W_{B0} - x_{jE})$, using $\Delta E_{gmax}/kT = 7.5$. $t_B(W_{cap} - x_{jE})$ is the base transit time when there is a Ge-free layer of thickness $W_{cap} - x_{jE}$ at the emitter end of the quasineutral base. $t_B(0)$ is the base transit time for the intended Ge distribution where the thickness of the Ge-free layer is zero. The shape of the curve is caused by the balance of the various terms in Eq. (7.60). Figure 7.20 suggests that for a given ΔE_{gmax} and base width, as the thickness of the Ge-free layer at the emitter end of the base

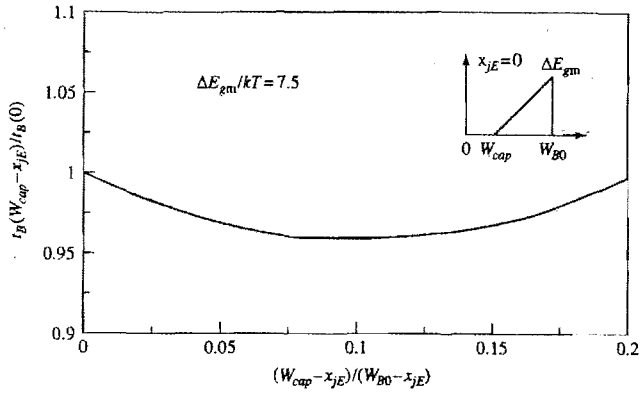


Figure 7.20 Relative change of base transit time at fixed quasineutral base width as a function of thickness of the Ge-free layer at the emitter end of the base.

increases from zero to some finite value, the base transit time goes through a minimum at a Ge-free layer thickness of about 10% of the base width. This result together with the dependence on ΔE_{g0} discussed above suggest that it is preferred to ramp down the Ge distribution more rapidly than intended instead of more slowly than intended. For a given quasineutral base width and ΔE_{gmax}, it is better to have a thin Ge-free layer at the emitter end of the base than to have $\Delta E_{g0be}(x_{jE}) > 0$.

7.4.6.3 Current Gain and Base Transit Time Tradeoff

The Ge distribution illustrated in Fig. 7.17 can be represented as the sum of a constant-Ge distribution and a graded-Ge distribution. That is, instead of Eq. (7.76), the base bandgap narrowing parameter can be written as

$$\Delta E_{gB,SiGe}(x) = \Delta E_{g0} + \Delta E'_{gB,SiGe}(x), \tag{7.78}$$

where

$$\begin{aligned} \Delta E'_{gB,SiGe}(x) &= \left(\frac{x - W_{cap}}{W_{B0} - W_{cap}}\right)(\Delta E_{g\,max} - \Delta E_{g0}) \quad x > W_{cap} \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad x < W_{cap}. \end{aligned} \tag{7.79}$$

Substituting Eq. (7.78) into Eq. (7.40), one can readily see that the part containing ΔE_{g0} drops out, and only the part containing $\Delta E'_{gB,SiGe}(x)$ remains in the integrals. That is, the constant-Ge part has no effect on base transit time. Only $\Delta E'_{gB,SiGe}(x)$ contributes to any base transit time enhancement. In other words, base transit time is a function of the energy difference $(\Delta E_{gmax} - \Delta E_{g0})$ only (also see Ex. 7.12). For a given ΔE_{gmax}, base transit time is smallest when $\Delta E_{g0} = 0$. The example illustrated in Fig. 7.20 thus corresponds to an optimal Ge distribution from a base transit time perspective.

However, the constant-Ge part causes the collector current and current gain to increase in proportion to $\exp(\Delta E_{g0}/kT)$, as can be inferred readily from Eq. (7.30). Therefore, for a given value of ΔE_{gmax}, there is a tradeoff between base transit time and current gain. Current gain can be increased readily by increasing ΔE_{g0}. But it will lead to an increase in base transit time as well.

7.4.7 Base-Width Modulation by V_{BE}

In the literature, the term base widening usually refers to widening of the quasineutral base layer at the collector end when the boundary between the quasineutral base layer and the base–collector space-charge layer extends towards and into the collector region. It is also known as Kirk effect, which is discussed in Section 6.3.3. Kirk effect causes an increase in base transit time, a reduction in base resistance, and a rolloff in collector current and current gain. The rolloff in collector current can also be caused by parasitic base and emitter resistances. Therefore, it may be difficult to recognize Kirk effect from the observed collector current rolloff. However, Kirk effect is readily recognizable from the rolloff in current gain because the parasitic resistances affect both the base and collector currents equally, while Kirk effect does not affect the base current. Kirk effect usually sets in at forward V_{BE} larger than 0.8 V when the mobile electron density becomes larger than the collector doping concentration. The larger the collector doping concentration, the larger the V_{BE} before significant Kirk effect sets in.

The space-charge layer width of an emitter–base diode is a function of V_{BE}. That means the width of a quasineutral base layer is modulated by V_{BE}. In Section 6.3.2.1, the modulation of the width of a quasineutral base layer by base–collector voltage V_{BC} was discussed. The degree of base-width modulation in a transistor by V_{BC} is indicated by its Early voltage. As a result, base-width modulation by V_{BE} is sometimes referred to in the literature as *reverse Early effect* (Crabbé *et al.*, 1993b; Salmon *et al.*, 1997; Deixler *et al.*, 2001). (This should not be confused with the Early voltage for a transistor operated in the reverse-active mode. The Early voltage in the reverse-active mode can be much smaller than that in the forward-active mode. See Section 7.4.8 below.)

Since the emitter is much more heavily doped than the base, the emitter–base diode can be treated as a one-sided diode. In this approximation, the emitter–base diode depletion layer resides in the base side only. From Eq. (2.80), this width is

$$W_{dBE}(V_{BE}) = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} - V_{BE})}{qN_B}}, \tag{7.80}$$

where ψ_{bi} is the built-in potential of the emitter–base diode, ε_{si} is the permittivity of silicon, and N_B is the doping concentration in the base. The base-side boundary of the emitter–base diode space-charge layer is also the emitter-side boundary of the quasineutral base layer. Therefore, as a bipolar transistor is turned on by increasing V_{BE} from zero to some positive value, W_{dBE} is reduced by an amount $W_{dBE}(0) - W_{dBE}(V_{BE})$, causing the quasineutral base width to increase by the same amount. As we shall show below, such widening of the base at the emitter end has negligible effect on a typical Si-base bipolar transistor. However, in a SiGe-base bipolar transistor, the effect can be readily observable because it is amplified by the base bandgap profile near the emitter end. It explains why

base widening at the emitter end has rarely been discussed in the literature until recently when SiGe-base bipolar transistors are studied in detail (Crabbé *et al.*, 1993b; Cressler *et al.*, 1993a,b; Paasschens *et al.*, 2001).

7.4.7.1 Model for Base-Width Modulation Caused By V_{BE}

It should be noted that in all of the models developed and discussed so far, we have implicitly ignored the emitter–base diode space-charge layer altogether by assuming that the quasineutral base extends from $x = x_{jE}$ at the emitter end to $x = W_{B0}$ at the collector end. This assumption is valid as long as effects caused by variation of W_{dBE} are negligible. Ignoring W_{dBE} greatly simplifies the schematics illustrating the emitter–base diode and depicting the Ge distribution within the quasineutral base layer.

However, when variation of W_{dBE} cannot be ignored, as is the case being considered here, the models and equations are still valid provided that we treat the quasineutral base as extending from $x = x_{jE}$ at the emitter end to $x = W_{B0}$ at the collector end. That is, x_{jE} is now used to denote the sum of the depth of the single-crystalline emitter region and the width of the emitter–base diode space-charge layer. In doing so, x_{jE} becomes a function of V_{BE}. When W_{dBE} varies with V_{BE}, the parameter x_{jE} varies with V_{BE} by the same amount. Equation (7.80) can be used to calculate the variation of x_{jE} as a function of V_{BE}. Widening of the quasineutral base at the emitter end can then be treated as "emitter depth variation". The models developed in Section 7.4.5 can be readily adopted to describe the effects of base-width modulation caused by V_{BE}.

7.4.7.2 J_{C0} Rolloff at Low Currents

It can readily be inferred from Eq. (7.29) that any widening of the quasineutral base layer will cause the saturated collector current density to decrease. Thus, a certain degree of J_{C0} rolloff caused by base widening at the emitter end is inherent in the operation of a bipolar transistor. For a Si-base transistor, this effect is quite small and usually ignored. As an illustration, consider a Si-base bipolar transistor having a boxlike base dopant distribution with an average concentration of $N_B = 5 \times 10^{18}$ cm^{-3}. The change in base width between $V_{BE} = 0.3$ V and $V_{BE} = 0.7$ V, estimated from Eq. (7.80), is about 4 nm. If the transistor has a quasineutral base width of 70 nm at $V_{BE} = 0.3$, this change is about 6%. That is J_{C0} at $V_{BE} = 0.7$ V is reduced by only about 6% compared to J_{C0} at $V_{BE} = 0.3$ V. If the transistor has a larger N_B, or if the base dopant distribution has a higher concentration at the emitter end, which is typically the case for an implanted base, the amount of rolloff in J_{C0} is even smaller.

However, in the case of a SiGe-base transistor, as V_{BE} is increased, the base bandgap at the emitter end of the quasineutral base may change. This is demonstrated in Fig. 7.21 where the quasineutral base widths at two values of V_{BE} are illustrated for two SiGe-base transistors. For simplicity, the transistors are assumed to have a linearly graded Ge distribution in the base. For the transistor in Fig. 7.21(a), $\Delta E_{g0be}(V_{BE} = 0)$ is larger than $\Delta E_{g0be}(V_{BE} > 0)$. For the transistor in Fig. 7.21(b), the Ge distribution is such that $\Delta E_{g0be} = 0$ for all positive values of V_{BE}, and the Ge-free region within the quasineutral base is thinner at $V_{BE} = 0$ than at $V_{BE} > 0$. The rolloff in J_{C0} for both transistors can be inferred from Fig. 7.14, which can be interpreted as a plot of variation of the saturated

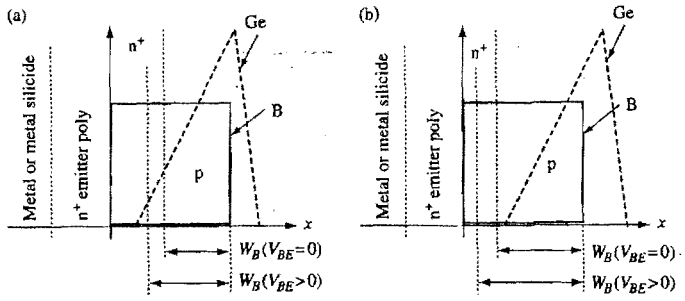


Figure 7.21. Schematics illustrating base widening at the emitter end for two SiGe-base transistors. (a) For this transistor, the Ge concentration ramps down to zero at a point beyond the emitter end of the quasineutral base layer when the transistor is biased at $V_{BE} = 0$. As V_{BE} is increased, $\Delta E_{g0be}(V_{BE})$ decreases. (b) For this transistor, the Ge concentration ramps down to zero before reaching the emitter end of the quasineutral base layer when the transistor is biased at $V_{BE} = 0$. As V_{BE} is increased, $\Delta E_{g0be}(V_{BE})$ remains zero.

collector current density as a function of base-width modulation caused by V_{BE}. The situation in Fig. 7.21(a) corresponds to the $(x_{jE} - W_{cup}) > 0$ part of Fig. 7.14. As V_{BE} is increased, the emitter end of the boundary of the quasineutral base, i.e., the location of x_{jE}, moves from right to left in the right half of Fig. 7.14. The situation in Fig. 7.21(b) corresponds to the $(x_{jE} - W_{cup}) < 0$ part of Fig. 7.14. As V_{BE} is increased, the emitter end of the boundary of the quasineutral base, i.e., the location of x_{jE}, moves from right to left in the left half of Fig. 7.14. For the same amount of base widening, i.e., for the same change in $|(x_{jE} - W_{cup})|$, Fig. 7.14 shows that the rolloff in J_{C0} is larger for the transistor in Fig. 7.21(a) than for the transistor in Fig. 7.21(b). Both transistors show significantly larger rolloff in J_{C0} than a Si-base transistor which is represented by the no-Ge curve in Fig. 7.14.

The magnitude of J_{C0} rolloff due to base widening at the emitter end is a strong function of the details of the Ge distribution near the emitter end, particularly if the Ge profile is more like a trapezoid than a triangle. This can be seen by comparing Figs 7.13 and 7.14. Very large rolloff in J_{C0} can be expected from a trapezoidal-like Ge distribution.

A rolloff in J_{C0} should lead to a rolloff in current gain. Figure 7.22 is a plot of observed current gain rolloff in a SiGe-base bipolar transistor (Crabbé *et al.*, 1993b). The initial rise, instead of falloff, in current gain at very low currents is caused by the nonideal nature of base current. (See Section 6.3.4 for a discussion on the ideality of base current in practical transistors.) The nonideal nature of base current causes J_{B0} to decrease with increasing V_{BE}. When J_{B0} decreases more rapidly with V_{BE} than J_{C0}, current gain will rise with increasing V_{BE} or with increasing collector current. This causes the measured current gain to increase at low collector currents.

Consider the data at 300 K in Fig. 7.22. The rapid rolloff in current gain at current densities greater than about 1.5 mA/μm^2 is caused by Kirk effect, which is base widening at the collector end. The slow current gain rolloff at current densities less than 1.5 mA/μm^2 is caused by base-width modulation by V_{BE}, which is base widening at the emitter end.

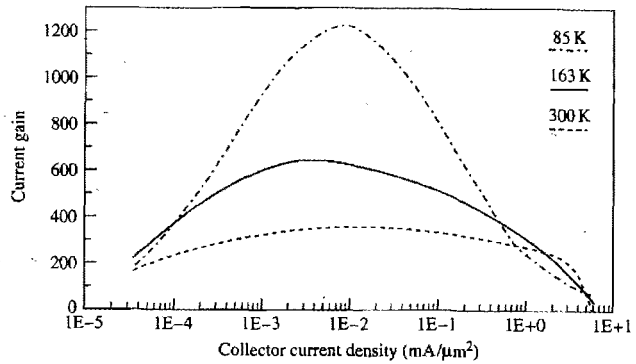


Figure 7.22. Measured current gain rolloff in a typical SiGe-base transistor as a function of collector current density. (After Crabbé *et al.*, 1993b.)

Figure 7.22 also shows that current gain rolloff due to V_{BE}-induced base widening increases rapidly as temperature decreases. This is to be expected because the values of $\Delta E_{g\max}$ and ΔE_{g0} are fixed for a given SiGe-base transistor, but the values of $\Delta E_{g\max}/kT$ and $\Delta E_{g0}/kT$ increase as temperature decreases, causing both current gain and "emitter depth variation" effect to increase.

7.4.7.3 Ideality of the Collector Current

As shown in Fig. 6.5, the collector current of a Si-base bipolar transistor is quite ideal, i.e., it is proportional to $\exp(qV_{BE}/mkT)$ with m very close to unity at V_{BE} less than about 0.9 V (before Kirk effect sets in and/or before emitter series resistance effect becomes significant). This fact is often used to determine the operating temperature of a Si-base transistor. As discussed in the subsection above, base widening at the emitter end can cause appreciable J_{C0} rolloff in a SiGe-base bipolar transistor even before Kirk effect sets in. That is, the measured collector current of a SiGe-base transistor often has an $\exp(qV_{BE}/mkT)$ dependence with m greater than unity. Therefore, unless the ideality factor m can be determined separately, there can be an appreciable error in using the collector current of a SiGe-base bipolar transistor to determine the device operating temperature.

7.4.7.4 V_{BE} as a Reference

Circuit designers often use V_{BE} as a reference. V_{BE}-refcrenced circuits are based on the assumption that the collector current I_C of a bipolar transistor is determined by its emitter area and V_{BE}. Circuit dcsigners often refer to "V_{BE}" of a transistor as the V_{BE} value needed to achieve a target I_C value. For a Si-base bipolar transistor, the relationship between I_C and V_{BE} is simply $I_C = A_E J_{C0}\exp(qV_{BE}/kT)$ [see Eq. (6.32)], and J_{C0} is independent of V_{BE} for V_{BE} less than about 0.9 V. However, as mentioned above, the dependence of J_{C0} on V_{BE} may not be negligible for a SiGe-base bipolar transistor. Furthermore, J_{C0} of a SiGe-base bipolar transistor is a strong function of its Ge profile near the emitter end of the quasineutral base. It has been shown that a small change in the Ge profile shape can cause an appreciable change in "V_{BE}" (Salmon *et al.*, 1997; Deixler *et al.*, 2001). Therefore, special attention should be paid to the nonideality nature of the collector current when designing V_{BE}-referenced circuits using SiGe-base bipolar transistors.

7.4.7.5 Minimizing V_{BE}-Induced Base-Width Modulation Effects

Base widening at the emitter end can be minimized by minimizing the dependence of W_{dBE} on V_{BE}. This can be accomplished easily by increasing base doping concentration N_B, as can be inferred from Eq. (7.80) and discussed in Section 7.4.7.2. Additionally, the base and emitter fabrication processes should also be designed so that x_{jE} is located in a region of constant or slowly varying Ge concentration (x_{jE} is the sum of the depth of the single-crystalline emitter region and the width of the emitter–base diode space-charge layer). Preferably, x_{jE} should be located in a region of relatively low Ge concentration, or in a Ge-free region such as the case illustrated in Fig. 7.21(b). As discussed in Section 7.4.6.2, such a design is also preferred from base transit time consideration.

7.4.8 Reverse-Mode I–V Characteristics

A bipolar transistor is normally operated in the forward-active mode. Occasionally, a transistor is operated in the reverse-active mode either unintentionally or by design. For instance, when a transistor goes into saturation in a circuit, its collector–base junction becomes forward biased and its collector current is the difference between a forward component and a reverse component. In this section we compare the reverse-mode currents to the forward-mode currents.

The normal (or forward) and reverse modes of operation of a SiGe-base bipolar transistor are depicted schematically in Fig. 7.23. Here, for simplicity of illustration, we assume a simple triangular Ge distribution in the quasineutral base. As discussed in Section 6.4.1, the reciprocity relationship between emitter and collector implies that the magnitude of the collector current in normal mode is equal to the magnitude of the collector current in reverse mode. That is, we have *in theory*

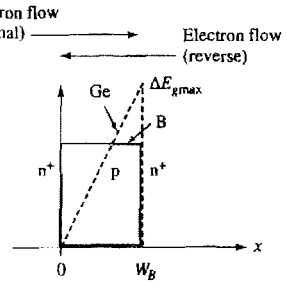


Figure 7.23. Schematic illustrating a SiGe-base bipolar transistor operated in the forward and reverse modes. In the forward mode, the left n^+ region is the emitter and the right n^+ region is the collector. In the reverse mode, the right n^+ region is the emitter and the left n^+ region is the collector.

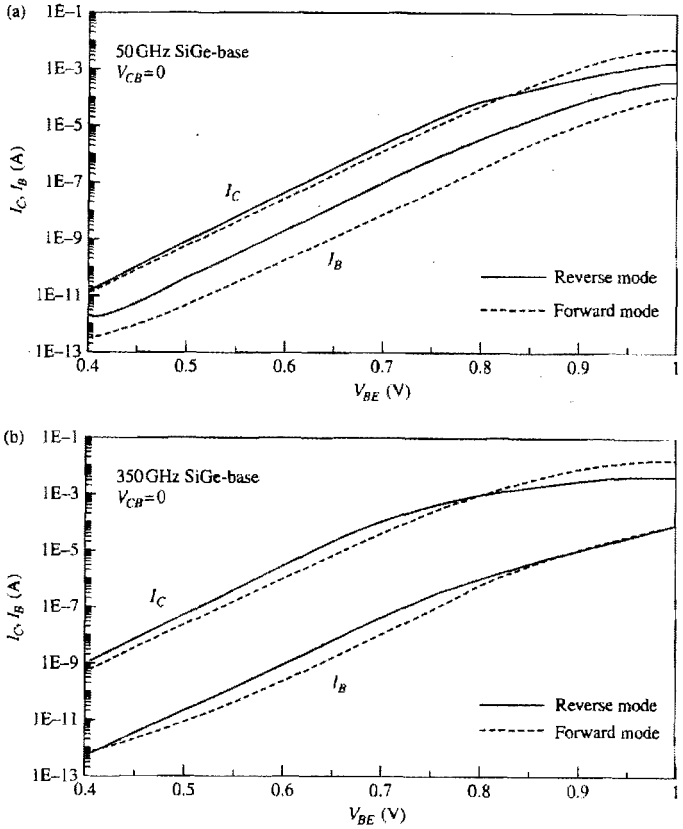


Figure 7.24. Measured Gummel characteristics of two SiGe-base bipolar transistors in the normal forward-active mode and in the reverse-active mode. (a) A transistor having f_T = 50 GHz, V_A(forward) = 55.7 V and V_A(reverse) = 1.34 V. (b) A transistor having f_T = 350 GHz, V_A(forward) = 22.2 V and V_A(reverse) = 0.16 V. (After Rieh *et al.*, 2005.)

$$I_C(\text{forward}) = I_C(\text{reverse}) \tag{7.81}$$

for Si-base transistors as well as for SiGe-base transistors with an arbitrary Ge distribution in the quasineutral base. Figure 7.24 shows the measured forward and reverse Gummel characteristics for two SiGe-base bipolar transistors. The transistor in Fig. 7.24(a) has I_C(reverse) about the same as I_C(forward), but its I_B(reverse) is significantly larger than its I_B(forward). The transistor in Fig. 7.24(b) has both I_C and I_B appreciably larger in the reverse mode than in the forward mode. The physical mechanisms governing these subtle differences are discussed next.

Let us first examine the base currents. The base currents of a transistor in forward and reverse modes are quite different because the emitter parameters in forward mode are quite different from the emitter parameters in reverse mode. In forward mode, the transistor has a polysilicon emitter. In reverse mode, the transistor has a complex n-type region (the pedestal collector region and the heavily doped subcollector region) as emitter. Also, the emitter–base diode area in reverse mode (intrinsic-base area plus extrinsic-base area) is much larger than the emitter–base diode area in forward mode (intrinsic-base area only). The net is that the base current should be larger in reverse mode than in forward mode.

Regarding the collector currents, the two collector currents for the transistor in Fig. 7.24(a) are about the same for V_{BE} values less than about 0.8 V where the currents are reasonably ideal. This is consistent with Eq. (7.81). At V_{BE} values larger than about 0.8 V, the two collector currents are no longer ideal and are significantly below their ideal values because of a combination of series resistance effect and Kirk effect. The emitter series resistance in the forward mode is smaller than that in the reverse mode. The emitter series resistance in the reverse mode includes the resistances in the pedestal collector, the subcollector layer and the reach-through region (see Fig. 6.7). On the other hand, Kirk effect is significantly less in the reverse mode because of the absence of a lightly doped region in the "collector". That is, that I_C (reverse) drops below I_C (forward) at high V_{BE} is primarily due to the large emitter series resistance in the reverse mode.

Even in the voltage range where series resistance and Kirk effects are negligible, there are two other effects that can cause the measured collector currents to be different than predicted by Eq. (7.81). These are the effect of base-width modulation due to V_{BE} (see Section 7.4.7) and base-width modulation effect due to V_{BC} (see Section 6.3.2.1). We shall examine these two effects separately.

As explained in Section 7.4.7, base-width modulation due to V_{BE} causes the collector current to differ from its ideal value. In forward mode, the emitter–base diode space-charge layer resides primarily in the base side because the doping concentration in the base is much smaller than that in the emitter. As V_{BE} is increased, the emitter–base diode space-charge layer thickness, W_{dBE}, is reduced [see Eq. (7.80)], causing the quasineutral base layer to widen by the same amount. Widening of the quasineutral base causes a decrease in J_{C0}(forward). In the case of a SiGe-base transistor, this reduction in J_{C0}(forward) is amplified by the graded Ge profile (see Section 7.4.7.2). However, in reverse mode, the "emitter" doping concentration is much smaller than the base doping concentration. As a result, the emitter–base diode space-charge layer resides primarily in the "emitter" side instead of in the base side. As "V_{BE}" is increased in reverse mode, the change in "W_{dBE}" of the emitter–base diode is absorbed in the "emitter" side instead of in the base side. That is, there should be negligible "base widening at the emitter end" when a transistor is operated in reverse mode. Therefore, on this basis alone, we should expect the measured J_{C0}(reverse) to be somewhat larger than the measured J_{C0}(forward).

The effect of base-width modulation by V_{BC} in a transistor is characterized by its Early voltage (see Section 6.3.2.1). It can be inferred readily from Fig. 6.8 that, for the same base current or same V_{BE}, the smaller the Early voltage the larger the collector current. Also, it can be shown that, for a SiGe-base transistor having a graded base bandgap, the Early voltage in reverse mode is much smaller than that in forward mode (see Ex. 7.14). The smaller Early voltage in reverse mode definitely contributes to the observed I_C (reverse) being larger than I_C (forward) in a typical SiGe-base transistor.

Returning to Fig. 7.24, it is apparent that a SiGe-base transistor having a higher f_T has greater asymmetry between I_C (forward) and I_C (reverse) than a transistor having a lower f_T. This observation is consistent with the two base-width-modulation effects discussed above. The important point is that the subtle differences in the forward and reverse I–V characteristics of a SiGe-base transistor should be included in modeling bipolar circuits which involve transistors operated in reverse and/or saturation modes.

7.4.9 Heterojunction Nature of a SiGe-Base Bipolar Transistor

In a wide-gap-emitter HBT, the small base current, and hence the large current gain, is due to the large energy barrier for base current injection at the emitter–base junction (Kroemer, 1957). A wide-gap-emitter HBT usually has a large Early voltage as well because the small base current allows the intrinsic-base layer to be doped very heavily and still maintain sufficient current gain. A heavily doped intrinsic base in turn leads to large Early voltage [cf. Eq. (6.72)]. Both adequate current gain and large Early voltage can be obtained simultaneously in a wide-gap-emitter HBT.

In the literature, most SiGe-base transistors have, by design, either a graded base bandgap or a constant but narrowed base bandgap. Both the graded-bandgap and the constant-bandgap SiGe-base transistors are often referred to as HBTs as well. In this section, we want to examine the heterojunction nature of these two types of SiGe-base transistors, and contrast their properties with wide-gap-emitter HBTs.

7.4.9.1 Heterojunction Nature of a Graded-Bandgap SiGe-Base Transistor

Consider a SiGe-base transistor having a simple linearly graded base bandgap, i.e., a transistor having a simple triangular Ge profile in the base. Compared to a Si-base transistor, the linearly graded base bandgap leads to a larger J_{C0}, a larger Early voltage and a smaller base transit time. There is no change in base current. Referring to Fig. 7.6, we see that the increase in Early voltage can be very large, while the increase in J_{C0}, and hence the increase in current gain, and reduction in base transit time are moderate in comparison.

There is a tradeoff between Early voltage and current gain in a SiGe-base transistor [see Eq. (7.37)]. Of course, additional tradeoffs can be made by modifying the intrinsic-base sheet resistance, just as in any transistor [see Eq. (6.74)]. The key difference between a graded-base-bandgap transistor and a wide-gap-emitter transistor is that the graded base bandgap leads to improvements in current gain, Early voltage and transit time simultaneously and naturally, while a wide-gap-emitter HBT, without base-bandgap grading, inherently improves only current gain.

It is interesting to note that Eq. (7.81) implies that the two SiGe-base bipolar transistors illustrated in Fig. 7.25 should have the same basic current-voltage characteristics (ignoring Early voltage effect). Transistor (a) is a usual SiGe-base transistor with a simple graded Ge distribution in its quasineutral base, while transistor (b) is the same transistor as (a) in every respect except that its Ge distribution is retrograded. That is, for transistor (b), ΔE_{gmax} is located at the emitter end of the base layer. The energy-band diagrams corresponding to the emitter–base diodes of these two transistors are illustrated in

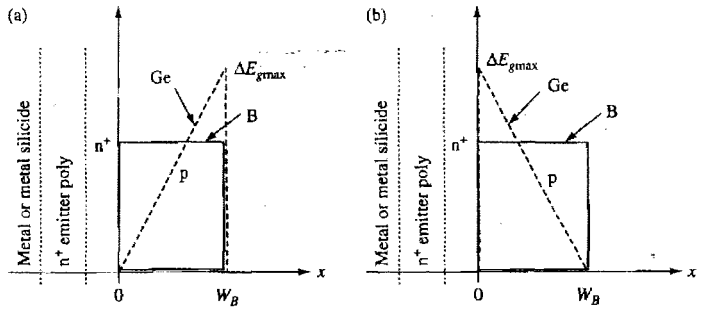


Figure 7.25. Schematics illustrating the emitter-base diodes of two SiGe-base bipolar transistors. Transistor (a) is a usual SiGe-base bipolar transistor having a graded Ge distribution, with ΔE_{gmax} at the collector end of its quasineutral base. Transistor (b) is the same transistor except that the Ge distribution is retrograded, with ΔE_{gmax} at the emitter end of its quasineutral base.

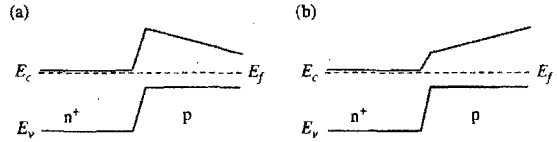


Figure 7.26. Energy-band diagrams corresponding to the emitter–base diodes illustrated in Fig. 7.25, at $V_{BE} = 0$.

Fig. 7.26. It shows that the two transistors have very different heterojunction features at the emitter–base junction. The energy barrier for electron injection at the junction interface appears larger for transistor (a) than for transistor (b), although the maximum electron barriers across the entire quasineutral base are the same.

To first order, the two transistors have the same collector current. However, once we include second order effects, such as base-width modulation caused by V_{BE} and V_{BC}, the two transistors in Fig. 7.25 have readily distinguishable characteristics, as discussed in Section 7.4.8. In addition, transistor (b) has a retarding field in its base region which gives it a larger base transit time than transistor (a) (see Ex. 7.15). In other words, the characteristics of a SiGe-base transistor having a nonuniform base bandgap are determined more by the direction of the base bandgap grading and the amount of bandgap difference across the base, and less by the electron and hole injection barriers at the emitter–base junction.

7.4.9.2 Heterojunction Nature of a Constant-Bandgap SiGe-Base Transistor

To clearly distinguish a constant-Ge SiGe-base transistor from a graded-Ge SiGe-base transistor, we assume the Ge distribution in the constant-Ge transistor to ramp down abruptly near the emitter end, such as those illustrated in Figs 7.10 and 7.11, instead of at some finite rate.

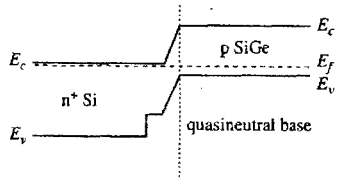


Figure 7.27. Energy-band diagram corresponding to the emitter–base diodes illustrated in Fig. 7.10, at $V_{BE} = 0$.

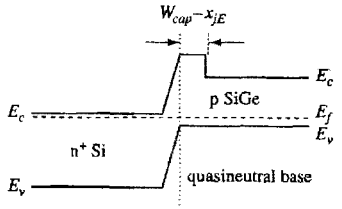


Figure 7.28. Energy-band diagram corresponding to the emitter–base diodes illustrated in Fig. 7.11, at $V_{BE} = 0$.

- *Case of Ge in the emitter* (i.e., $x_{jE} > W_{cap}$). The Ge distribution and the dopant distributions in the emitter and base regions are illustrated schematically in Fig. 7.10. The corresponding energy-band diagram is illustrated in Fig. 7.27. The emitter bandgap is large compared to the base bandgap. Indeed, one can think of this transistor as a wide-gap-emitter HBT. However, as discussed in Section 7.4.2.2, compared with a Si-base transistor having the same polysilicon emitter and base dopant distribution, this transistor has the same, instead of smaller, base current. It has larger current gain, consistent with a wide-gap-emitter transistor, but the larger current gain is due to an increase in collector current, not a reduction in base current. Also, just like a wide-gap-emitter HBT without base-bandgap grading, there is little improvement in Early voltage or in base transit time (other than from the correction factors η and γ).
- *Case of no Ge in the emitter* (i.e., $x_{jE} < W_{cap}$). In this case, the Ge distribution and the dopant distributions in the emitter and base regions are as illustrated schematically in Fig. 7.11. The corresponding energy-band diagram is illustrated in Fig. 7.28. If we focus at the region near the emitter–base junction, the transistor appears not to be a wide-gap-emitter device at all because the emitter and the emitter end of the quasi-neutral base have the same energy bandgap. Indeed, compared to a Si-base transistor having the same polysilicon emitter and base dopant distribution, this transistor has the same base current. However, as discussed in Section 7.4.4 and shown in Ex. 7.10, this transistor not only has higher current gain, due to its larger collector current, but also larger Early voltage and smaller base transit time. The high-low energy gap in the base makes this transistor behave more like a graded-base-bandgap transistor than a constant-base-bandgap transistor.

The net of all these is that ***it is best to think of a SiGe-base device as a graded-base-bandgap bipolar transistor instead of an HBT.*** By focusing on the base-bandgap grading characteristics instead of the emitter–base junction parameters, we focus on the device properties that depend on electron injection into and transport across the base. It is the electron injection and transport across the base region that makes a SiGe-base transistor superior to a Si-base transistor.

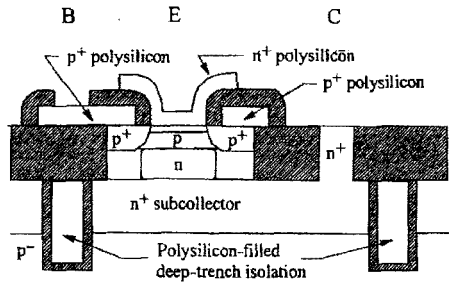


Figure 7.29. Schematic illustrating the structure of the commonly used bipolar transistor and its salient features.

7.5 Modern Bipolar Transistor Structures

After a relatively large research and development effort worldwide in the 1970s and 1980s, bipolar technology has become fairly mature. Since the mid 1990s, the growth in wireless and RF applications has revived the interest in bipolar technology research and development. This time, the focus is on optimizing the SiGe-base bipolar transistor. Techniques for implementing the device design concepts discussed in this chapter have been developed, and in most cases implemented. The most widely used bipolar technology today is probably the deep-trench-isolated, double-polysilicon, self-aligned bipolar technology (Ning *et al.*, 1981; Chen *et al.*, 1989) and variations of it. This device structure is illustrated schematically in Fig. 7.29. The process flow for fabricating this transistor structure is outlined in Appendix 2. The salient features of this device are: (i) deep-trench isolation between adjacent transistors, (ii) polysilicon emitter, (iii) polysilicon base contact, which is self-aligned to the emitter, and (iv) pedestal collector, i.e., the collector region directly underneath the emitter is more heavily doped than its surrounding regions. Also, for analog circuit applications, a SiGe-base transistor is particularly advantageous. The basic concept for each of these features is discussed below.

More recently, vertical bipolar transistors employing silicon-on-insulator (SOI) substrate with silicon thickness that are compatible with SOI CMOS have been demonstrated (Cai *et al.*, 2002a; 2003). The idea is to develop high-speed and low-power SOI BiCMOS for mixed-signal applications. The subject of SOI bipolar is discussed in Section 10.2.

7.5.1 Deep-Trench Isolation

The isolation region must be deep enough to isolate the subcollectors of adjacent transistors. Prior to the advent of deep-trench isolation, p-type diffusion regions are used to isolate subcollectors, as illustrated in Fig. 7.4. A p-type diffusion-isolation region is typically as wide as it is deep. This is because of the fact that, as the p-type impurities diffuse

downward, they also diffuse laterally. Furthermore, to minimize junction capacitance and to avoid excessively low collector–substrate junction breakdown voltage, the p-type diffusion-isolation regions should not butt against the heavily doped n-type subcollector regions. There is usually an n^- region between a p-type isolation and a subcollector or reach-through, as illustrated in Fig. 7.4. As a result, the total silicon area taken up by the diffusion-isolation regions is very large. The area of a diffusion-isolated bipolar transistor is completely dominated by its isolation.

Replacing diffusion isolation by deep-trench isolation reduces very significantly the area taken up by isolation. The horizontal dimension of the deep trenches is usually defined by lithography. With deep-trench isolation, the trenches can cut right through the subcollector layer, resulting in much smaller collector–substrate area and capacitance than with diffusion isolation.

The diffusion-isolation process is less complex, and hence may cost less, than the trench-isolation process. For reason of cost, diffusion isolation is still used in some bipolar products. However, the area taken up by isolation translates into cost as well. Also, as power dissipation is a factor of growing importance in the choice of a technology, product designers often favor trench isolation over diffusion isolation because of its smaller parasitic capacitance, which leads to systems with lower power dissipation.

7.5.2 Polysilicon Emitter

The benefit of polysilicon emitter has already been discussed in Section 7.1. Polysilicon emitter allows extremely small emitter-junction depths to be achieved without the large base current associated with a metal-contacted shallow emitter. Small emitter-junction depths are needed for making thin-base transistors with reproducible base-region parameters, and hence reproducible collector current characteristics. Thus, in many respects, polysilicon emitter is the enabling technology for scaling bipolar transistors to small dimensions. All modern bipolar transistors, including SiGe-base transistors, employ polysilicon-emitter technology.

The very low thermal cycle associated with the formation of a polysilicon emitter, compared with that associated with a diffused-emitter process, has resulted in a drastic reduction in the density of defects called *pipes*, which are localized emitter–collector shorts formed in the intrinsic-base layer. In general, the emitter thermal cycle can be further minimized by using phosphorus instead of arsenic to dope the polysilicon layer, particularly if the polysilicon layer is *in-situ* doped. Furthermore, rapid thermal annealing, instead of furnace annealing, leads to the shallowest emitters with low series resistance. The reader is referred to the literature for reports on using phosphorus-doped polysilicon for emitter (Nanba *et al.*, 1991; Crabbé *et al.*, 1992; Shiba *et al.*, 1996).

7.5.3 Self-Aligned Polysilicon Base Contact

As illustrated in Fig. 7.29, instead of being contacted directly by metal, the extrinsic base is contacted indirectly via a layer of p-type polysilicon. Metal contact to the p-type polysilicon is made on top of the field-oxide region. In this way, the extrinsic-base area

does not have to accommodate the base metal contact, and hence can be made quite small, resulting in very small extrinsic-base–collector junction area and capacitance.

Furthermore, by separating the extrinsic-base polysilicon layer and the emitter polysilicon layer by only a thin vertical insulator layer, the extrinsic-base area is further reduced. This thin vertical insulator layer is often referred to as a *sidewall* insulator layer. It is typically formed by the deposition of a thin insulator layer of the desired thickness followed by anisotropic reactive-ion etching, removing the insulator completely everywhere except where the insulator covers a vertical surface. The ratio of the total collector–base (extrinsic base + intrinsic base) junction area to the emitter–base junction area is typically 3:1 or less (Ning *et al.*, 1981).

Perhaps more important is the fact that polysilicon-base technology allows the extrinsic base to be formed independently of the intrinsic base. This decoupling of the intrinsic and extrinsic base greatly enlarges the intrinsic-base design and process window. It allows thin-base transistors to be made readily. Practically all modern bipolar transistors employ polysilicon-base technology, although not necessarily with self-alignment to get the smallest possible extrinsic-base area.

In a polysilicon-base-contact technology, the extrinsic-base resistance is limited primarily by the sheet resistance of the polysilicon layer. This sheet resistance can be minimized by forming a metal silicide layer on the polysilicon layer for as large an area as possible. One way is to replace the polysilicon layer by a composite layer of metal silicide and polysilicon. Another way is to form a silicide layer on the polysilicon layer practically everywhere (Chiu *et al.*, 1987; Iinuma *et al.*, 1995), or to form a sidewall silicide layer on the vertical edges of the polysilicon layer (Shiba *et al.*, 1991).

7.5.4 Pedestal Collector

A pedestal collector has a higher doping concentration in the active collector directly underneath the intrinsic base than its surrounding regions (Yu, 1971), as illustrated in Fig. 7.29. The high collector doping concentration minimizes base-widening effect, while the low parasitic-collector doping concentration reduces the total base-collector junction capacitance. A pedestal collector can be achieved quite simply by ion implantation when the emitter opening is defined in the device fabrication process (see Appendix 2).

7.5.5 SiGe-Base

The benefit of incorporating Ge into the intrinsic base layer has been discussed in Section 7.4. All SiGe-base transistors used in products employ polysilicon emitter, which is required to achieve thin base widths. For maximum device performance, the other device features, such as polysilicon base contact, trench isolation and pedestal collector are also employed, just like in a regular Si-base transistor. Recently, it was shown that the incorporation of carbon into silicon (Stolk *et al.*, 1995) and into SiGe (Lanzerotti *et al.*, 1996) can greatly suppress the diffusion of boron. As a result, designers have been incorporating carbon into the base SiGe layer to obtain ultra-thin-base bipolar transistors (Rücker *et al.*, 1999).

Exercises

7.1 This exercise is designed to show the sensitivity of the current gain to the polysilicon thickness of a polysilicon-emitter transistor. For the polysilicon-emitter model described in Exercise 6.3, the emitter Gummel number is a function of the emitter junction depth W_E and the polysilicon thickness W_{E1}, i.e.,

$$G_E(W_E,\ W_{E1}) = N_E\left(\frac{n_i^2}{n_{ieE}^2}\right)\left(\frac{W_E}{D_{pE}} + \frac{L_{pE1}\tanh\left(W_{E1}/L_{pE1}\right)}{D_{pE1}}\right).$$

It is reasonable to assume the lifetimes in heavily doped silicon and heavily doped polysilicon to be the same for the same doping concentration, since both are determined by Auger recombination. It is also reasonable to assume the mobility in polysilicon to be smaller than that in silicon, since there is additional grain-boundary scattering in polysilicon.

(a) A typical nonpolysilicon emitter has $N_E = 10^{20}$ cm^{-3} and $W_E = 300$ nm, and $W_{E1} = 0$. Estimate G_E (300 nm, 0) using the hole mobility and lifetime values in Fig. 2.24(a) and (b).

(b) A typical polysilicon emitter has $N_E = 10^{20}$ cm^{-3} and $W_E = 30$ nm. Let us assume the hole mobility in polysilicon to be $\frac{1}{3}$ that in silicon, and assume the hole lifetimes in silicon and polysilicon to be the same. Estimate G_E (30 nm, W_{E1}) for $W_{E1} = 50$, 100, 200, and 300 nm.

(c) Graph the ratio G_E (30 nm, W_{E1})/G_E (300 nm, 0) as a function of W_{E1}.

7.2 The intrinsic-base sheet resistivity is given by Eq. (7.5), namely

$$R_{Sbi} = \left(q\int_0^{W_B} p_p(x)\mu_p(x)dx\right)^{-1}.$$

Most n–p–n transistors have a low-current R_{Sbi} value of about 10^4 Ω/□. Assuming a boxlike base doping profile, graph N_B as a function of W_B for W_B between 50 and 300 nm for $R_{Sbi} = 10^4$ Ω/□. This graph illustrates how the base doping concentration varies with the intrinsic-base width in scaling in many practical device designs.

7.3 The base transit time for a box profile is given by

$$t_B \approx \frac{W_B^2}{2D_{nB}}.$$

For a fixed intrinsic-base sheet resistivity of 10^4 Ω/□, calculate and plot t_B as a function of W_B for W_B between 50 and 300 nm. Use the mobility and lifetime values in Fig. 2.24(a) and (b) to estimate D_{nB}. (See Exercise 7.2 for the relation between W_B and N_B.)

7.4 This exercise illustrates the tradeoff between the collector current density and the Early voltage in an optimized bipolar device design. The Early voltage for a boxlike base doping profile is given by $V_A = qN_BW_B/C_{dBC}$, where C_{dBC} is the base–collector junction depletion-layer capacitance per unit area [cf. Eq. (6.72)]. To maintain negligible base widening in scaling, we assume the collector current density is maintained at $J_C = 0.3qv_{sat}N_C$, where $v_{sat} = 10^7$ cm/s is the electron saturated velocity. Thus, as N_C is increased to increase J_C, C_{dBC} is increased and V_A is decreased.

(a) Use the one-sided junction approximation for the base–collector diode, and assume $V_{CB} = 2$ V (for purposes of calculating C_{dBC}). Plot C_{dBC} as a function of J_C for J_C between 0.1 and 5 mA/μm^2.

(b) For a base design with $qN_BW_B = 1.6 \times 10^{-6}$ C/cm^2 (e.g., $N_B = 10^{18}$ cm^{-3} and $W_B = 100$ nm), estimate and plot V_A as a function of J_C for J_C between 0.1 and 5 mA/μm^2.

7.5 Consider an n–p–n transistor with a wide base of $W_B = 500$ nm. Suppose the base doping concentration is linearly graded, i.e., N_B has the form $N_B(x) = A - \alpha x$, with $N_B(0) = 2 \times 10^{17}$ cm^{-3} and $N_B(W_B) = 2 \times 10^{16}$ cm^{-3}. For such light doping concentrations, the effect of heavy doping is negligible. Plot the built-in electric field due to the dopant distribution as a function of distance between $x = 0$ and $x = W_B$.

7.6 Consider an n–p–n transistor with a linearly graded base doping profile (cf. Exercise 7.5). The doping concentration at the emitter–base junction is $N_B(0) = 5 \times 10^{18}$ cm^{-3}. The doping concentration at the base–collector junction is $N_B(W_B) = 5 \times 10^{17}$ cm^{-3}, and $W_B = 100$ nm. For such high doping concentrations, the heavy-doping effect cannot be ignored. Plot the electric fields due to the dopant distribution and due to the heavy-doping effect, as well as the total electric field, as a function of distance from $x = 0$ to $x = W_B$. [Use the bandgap-narrowing parameter in Eq. (6.17).]

7.7 The emitter series resistance r_e of a polysilicon-emitter n–p–n transistor, with negligible polysilicon–silicon interface oxide, has three components, namely, the resistance due to the single-crystal emitter region, the resistance due to the polysilicon layer, and the resistance due to the metal–polysilicon contact. Consider a polysilicon emitter with $N_E = 10^{20}$ cm^{-3}, a single-crystal region of width $W_E = 30$ nm, a polysilicon layer of thickness $W_{E1} = 200$ nm, and a metal–polysilicon contact resistivity of 2×10^{-7} Ω-cm^2. Assume that, for the same doping concentration, the resistivity of polysilicon is 3 times that of single-crystal silicon. Calculate the series resistance components, as well as the total series resistance r_e, for an emitter 1 μm^2 in area. (Use the resistivity for n-type silicon shown in Fig. 2.9.)

7.8 In the literature, the heavy-doping effect in the emitter is well recognized, but in the base it is often ignored. The saturated collector current density for an n–p–n transistor is [see Eq. (6.33)]

$$J_{C0} = \frac{q}{\displaystyle\int_0^{W_B} \frac{p_p(x)}{D_{nB}(x)n_{ieB}^2(x)}dx}.$$

Assume a uniformly doped base with $N_B = 10^{18}$ cm^{-3} and $W_B = 100$ nm. Also assume low current levels so that $p_p = N_B$. Estimate J_{C0} for the following two cases: (a) the heavy-doping effect in the base is neglected, i.e., $n_{ieB} = n_i$, and (b) the heavy-doping effect in the base is included. (This exercise demonstrates that heavy doping in the intrinsic base of modern bipolar transistors cannot be ignored.)

7.9 Plot the collector current ratio given by Eq. (7.64) and the Early voltage ratio given by Eq. (7.65) as a function of the ratio $(W_{cap} - x_{jE})/(W_{B0} - x_{jE})$ from $(W_{cap} - x_{jE})/(W_{B0} - x_{jE}) = 0$ to $(W_{cap} - x_{jE})/(W_{B0} - x_{jE}) = 1$, for $\Delta E_{g0}/kT = 0.5$, 1, 2.5, and 5.

7.10 In practice, the Ge concentration in a constant-Ge SiGe-base transistor does not ramp down abruptly at the emitter end of the base. Instead it is ramped down at some finite rate. If the emitter depth is not deep enough to extend beyond the ramp part, the Ge ramp has to be included in modeling an otherwise constant-Ge SiGe-base transistor. Figure 7.30 illustrates the Ge distribution, with the emitter–base junction right at the foot of the Ge ramp. The quasineutral base width is W_B, which includes the Ge ramp region of width W_{B1}. Show that, compared to a Si-base transistor having the same polysilicon emitter, base width and base dopant distribution,

$$\frac{J_{C0}(\text{SiGe})}{J_{C0}(\text{Si})} = \overline{\gamma\eta}\left\{\frac{kT}{\Delta E_{g0}}\frac{W_{B1}}{W_B}\left[1-\exp(-\Delta E_{g0}/kT)\right] + \left(1-\frac{W_{B1}}{W_B}\right)\exp(-\Delta E_{g0}/kT)\right\}^{-1},$$

$$\frac{V_A(\text{SiGe})}{V_A(\text{Si})} = \frac{kT}{\Delta E_{g0}}\frac{W_{B1}}{W_B}\left[\exp(\Delta E_{g0}/kT) - 1\right] + \left(1-\frac{W_{B1}}{W_B}\right),$$

and

$$\frac{t_B(\text{SiGe})}{t_B(\text{Si})} = \frac{1}{\bar{\eta}}\left\{\left(1-\frac{W_{B1}}{W_B}\right)^2 + \frac{2kT}{\Delta E_{g0}}\left(\frac{W_{B1}}{W_B}\right)^2\right\} + \frac{1}{\bar{\eta}}\left\{\frac{2kT}{\Delta E_{g0}}\frac{W_{B1}}{W_B}\left(1-\frac{W_{B1}}{W_B}-\frac{kT}{\Delta E_{g0}}\frac{W_{B1}}{W_B}\right)\left[1-\exp(-\Delta E_{g0}/kT)\right]\right\}.$$

Note that the base transit time ratio reduces to $1/\bar{\eta}$ when W_{B1} is reduced to zero, as expected. Our transit time ratio is different from that in Eq. (4) of Cressler *et al.* (1993b) which does not reduce to the expected value as W_{B1} is reduced to zero.

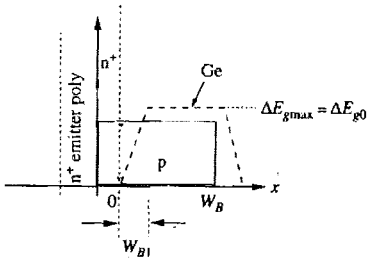


Figure 7.30.

7.11 Show that the collector current ratio for the SiGe-base bipolar transistor illustrated in Fig. 7.17 is

$$\frac{J_{C0}(\text{SiGe}, x_{jE})}{J_{C0}(\text{Si}, x_{jE})} = \frac{\overline{\gamma\eta}\exp(\Delta E_{g0}/kT)}{\left(\frac{W_{cap}-x_{jE}}{W_{B0}-x_{jE}}\right) + \left(\frac{W_{B0}-W_{cap}}{W_{B0}-x_{jE}}\right)\left(\frac{kT}{\Delta E_{g\,\max}-\Delta E_{g0}}\right)\left[1-\exp\left(\frac{\Delta E_{g0}-\Delta E_{g\,\max}}{kT}\right)\right]}.$$

7.12 Show that the base transit time ratio for the Ge distribution illustrated in Fig. 7.17 is

$$\left.\frac{t_B(\text{SiGe}, x_{jE})}{t_B(\text{Si}, x_{jE})}\right|_{x_{jE} < W_{cap}} = \frac{1}{\bar{\eta}}\frac{(W_{cap}-x_{jE})^2}{(W_{B0}-x_{jE})^2} + \frac{2}{\bar{\eta}}\frac{(W_{cap}-x_{jE})}{(W_{B0}-x_{jE})}\frac{(W_{B0}-W_{cap})}{(W_{B0}-x_{jE})}\left(\frac{kT}{\Delta E_{g\,\max}-\Delta E_{g0}}\right) \times\left[1-\exp\left(\frac{\Delta E_{g0}-\Delta E_{g\,\max}}{kT}\right)\right] + \frac{2}{\bar{\eta}}\frac{(W_{B0}-W_{cap})^2}{(W_{B0}-x_{jE})^2}\left(\frac{kT}{\Delta E_{g\,\max}-\Delta E_{g0}}\right) \times\left\{1-\left(\frac{kT}{\Delta E_{g\,\max}-\Delta E_{g0}}\right)\left[1-\exp\left(\frac{\Delta E_{g0}-\Delta E_{g\,\max}}{kT}\right)\right]\right\}.$$

Note that the transit time is a function of the energy difference $(\Delta E_{g\max} - \Delta E_{g0})$ only, and does not depend on $\Delta E_{g\max}$ or ΔE_{g0} individually. Changing $\Delta E_{g\max}$ and ΔE_{g0} together such that the difference remains the same has no effect on the base transit time.

7.13 It is instructive to use the equations obtained in Ex. 7.10 to study a "constant-Ge" SiGe-base transistor having a finite Ge-ramp region that starts at the emitter end of its quasineutral base, i.e., Fig. 7.30. Plot the ratios for collector current, Early voltage and base transit time as a function of W_{B1}/W_B from $W_{B1}/W_B = 0$ to $W_{B1}/W_B = 1$, for $\Delta E_{g0}/kT = 2.5$, 5 and 10.

7.14 Referring to Fig. 7.23, show that

$$\frac{J_{C0}(\text{SiGe, reverse})}{J_{C0}(\text{Si, reverse})} = \frac{\overline{\gamma\eta}\Delta E_{g\,\max}/kT}{1-\exp(-\Delta E_{g\,\max}/kT)}$$

by carrying out the integration in Eq. (7.30). Comparison of this result with Eq. (7.32) shows that $J_{C0}(\text{forward}) = J_{C0}(\text{reverse})$ for a SiGe-base bipolar transistor. This result is expected from the reciprocity relationship between the emitter and collector of an ideal bipolar transistor.

7.15 Referring to Fig. 7.25, show that
(a)

$$\frac{V_A(\text{transistor b})}{V_A(\text{transistor a})} = \exp\left(\frac{-\Delta E_{g\,\max}}{kT}\right),$$

and
(b)

$$\frac{t_B(\text{transistor b})}{t_B(\text{transistor a})} = \frac{\dfrac{kT}{\Delta E_{g\,\max}}\exp\left(\dfrac{\Delta E_{g\,\max}}{kT}\right)\left[1-\exp\left(\dfrac{-\Delta E_{g\,\max}}{kT}\right)\right]-1}{1-\dfrac{kT}{\Delta E_{g\,\max}}\left[1-\exp\left(\dfrac{-\Delta E_{g\,\max}}{kT}\right)\right]}.$$

8 Bipolar Performance Factors

In Chapter 7, the design of the individual regions and parameters of a bipolar transistor was discussed. It was noted that, during device operation, an individual device region is not isolated from and independent of the other device regions. Optimization of one device parameter often adversely affects the other device parameters. Thus, ***optimization of the design of a bipolar transistor is a tradeoff process***. This design tradeoff should be done at the circuit and/or chip level, for the optimum design of a transistor is a function of its application and environment. In this Chapter, we will first discuss some figures of merit for evaluating a bipolar transistor for typical digital and analog circuit applications, and then discuss the tradeoffs in the design of a bipolar transistor for these applications.

When we consider the performance of a circuit, the wires connecting the transistors and elements that make up the circuit and connecting the output of the circuit to the input of another circuit must be included. The resistance and capacitance as well as the signal propagation delays associated with the interconnect wires have been discussed in Section 5.2.4 in connection with CMOS circuits. The reader is referred to that subsection for details. In this Chapter, the wire capacitance which acts as a load on a bipolar circuit is included when we consider the performance and optimization of bipolar transistors and circuits.

In practice, the choice of a particular device design point is often dictated by many nontechnical factors. These factors include cost, time to market, production volume, etc. They will not be considered here.

8.1 Figures of Merit of a Bipolar Transistor

It is often desirable to consider the merit of a transistor in terms of some simple, and preferably readily measurable, parameters. However, it is important to note that the relevance or significance of a particular figure of merit depends on the application. Some of the commonly used figures of merit are discussed here.

8.1.1 Cutoff Frequency

For small-signal applications, the *cutoff frequency, or transition frequency, or unity-current-gain frequency*, f_T, is probably the most often used figure-of-merit of a bipolar transistor. It is defined as the transition frequency at which the common-emitter,

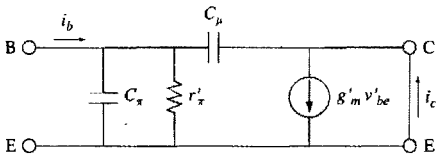


Figure 8.1. Small-signal equivalent circuit for determining the cutoff frequency of a bipolar transistor. Parasitic resistances are neglected.

short-circuit load, small-signal current gain drops to unity. It is a measure of the maximum useful frequency of a transistor when it is used as an amplifier. In Appendix 18, the cutoff frequency is derived using a two-port network analysis. Here we use a physically intuitive approach to obtain commonly used approximations for the cutoff frequency.

For simplicity, we shall first neglect the internal parasitic resistances and use the equivalent circuit shown in Fig. 6.19 to determine the intrinsic cutoff frequency f'_T. Here, the convention is such that the primed parameters refer to an intrinsic device, while the unprimed parameters are for an extrinsic device. With the output shorted, r_0 and $C_{dCS,tot}$ have no influence, and the resulting equivalent circuit is shown in Fig. 8.1. From this equivalent circuit, the small-signal collector and base currents can be written as

$$i_c = g'_m v'_{be} - j\omega C_\mu v'_{be} \tag{8.1}$$

and

$$i_b = \left(\frac{1}{r'_\pi} + j\omega C_\pi + j\omega C_\mu\right) v'_{be}, \tag{8.2}$$

where g'_m is the intrinsic transconductance, r'_π is the intrinsic input resistance, $C_\mu = C_{dBC,tot}$ is the base-collector junction depletion-layer capacitance, and $C_\pi = C_{dBE,tot} + C_{DE}$ is the sum of the base–emitter junction depletion-layer capacitance and the emitter diffusion capacitance, v'_{be} is the applied small-signal input voltage, and i_b and i_c are the small-signal base and collector currents due to v'_{be}. (The reader is referred to Section 6.4.3 for the derivation of the small-signal equivalent-circuit model of a bipolar transistor.) The small-signal frequency-dependent common-emitter current gain is

$$\beta'(\omega) = \frac{i_c}{i_b} = \frac{g'_m - j\omega C_\mu}{(1/r'_\pi) + j\omega(C_\pi + C_\mu)}. \tag{8.3}$$

In the low-frequency limit, Eq. (8.3) gives

$$\beta'(\omega = 0) = g'_m r'_\pi, \tag{8.4}$$

which, according to Eq. (6.100) is simply the static common-emitter current gain β_0.

The cutoff frequency f'_T of the intrinsic transistor is given by setting $\beta'(\omega = \omega'_T) = 1$. That is, from Eq. (8.3), we have

$$1 = \sqrt{\frac{g'^2_m + \omega'^2_T C^2_\mu}{\frac{g'^2_m}{\beta_0^2} + \omega'^2_T (C_\pi + C_\mu)^2}}, \tag{8.5}$$

which can be rearranged to give

$$\omega'_T = 2\pi f'_T = \sqrt{\frac{g'^2_m - \frac{g'^2_m}{\beta_0^2}}{(C_\pi + C_\mu)^2 - C^2_\mu}} \approx \frac{g'_m}{\sqrt{(C_\pi + C_\mu)^2 - C^2_\mu}}. \tag{8.6}$$

For a typical modern bipolar transistor which has a relatively small $C_\mu (= C_{dBC,tot})$, we have $C_\pi (= C_{dBE,tot} + C_{DE}) \gg C_\mu$. Therefore, Eq. (8.6) can be simplified to give the commonly used approximation of

$$\omega'_T = 2\pi f'_T \approx \frac{g'_m}{C_\pi + C_\mu} \tag{8.7}$$

or

$$\frac{1}{2\pi f'_T} = \tau_F + \frac{kT}{qI_C}(C_{dBE,tot} + C_{dBC,tot}) \tag{8.8}$$

where τ_F is the forward transit time given by Eq. (6.116), and I_C is the collector current.

If the internal parasitic resistances are included, then the equivalent circuit shown in Fig. 6.20 should be used. In this case, a similar analysis can be followed to obtain the extrinsic cutoff frequency f_T. A commonly used approximation is given by (see Exercise 8.1)

$$\frac{1}{2\pi f_T} = \tau_F + \frac{kT}{qI_C}(C_{dBE,tot} + C_{dBC,tot}) + C_{dBC,tot}(r_e + r_c), \tag{8.9}$$

where r_c is the collector series resistance and r_e is the emitter series resistance. The same result is derived in Appendix 18 through a two-port network analysis of an extrinsic transistor. Equation (8.9) is often used to determine the low-current value of τ_F. This is done by plotting the measured values of $1/f_T$ as a function of $1/I_C$. Figure 8.2 is an

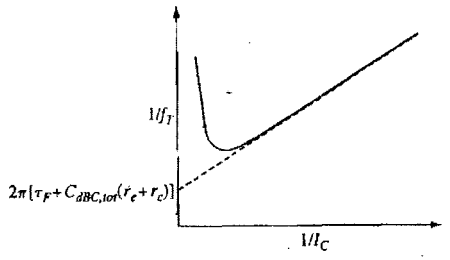


Figure 8.2. Schematic illustration of a $1/f_T$ versus $1/I_C$ plot. The extrapolated intercept at $(1/I_C) = 0$ can be used to determine $\tau_F + C_{dBC,tot}(r_e + r_c)$.

illustration of such a plot. At low currents, $1/f_T$ varies linearly with $1/I_C$. Equation (8.9) suggests that extrapolation of the linear portion of $1/f_T$ to $(1/I_C)=0$ gives $2\pi[\tau_F + C_{dBC,tot}(r_e + r_c)]$. However, at large currents, the measured $1/f_T$ increases very rapidly as current is increased. This is also illustrated in Fig. 8.2. This rapid rise is due to base-widening effects, and will be discussed in Section 8.5.1.

In the literature, the peak f_T is often quoted for a transistor designed for large-signal digital circuit applications. The sensitivity of the performance of a digital bipolar circuit to the peak f_T value of its transistors will be covered in Section 8.3.3.

8.1.2 Maximum Oscillation Frequency

The cutoff frequency is certainly a good indicator of the low-current forward transit time. However, as a figure of merit, it does not include the effects of base resistance, which are very important in determining the transient response of a bipolar transistor. Consequently, other figures of merit have been proposed and discussed in the literature (Taylor and Simmons, 1986; Hurkx, 1994; Hurkx, 1996). One that is relatively simple and commonly used is the *maximum oscillation frequency*, f_{max}, which is the frequency at which the unilateral power gain becomes unity. A commonly used approximation is given by (Pritchard, 1955; Thornton *et al.*, 1966; Roulston, 1990)

$$f_{max} = \left(\frac{f_T}{8\pi r_b C_{dBC,tot}}\right)^{1/2}, \tag{8.10}$$

where r_b is the base resistance. The reader is referred to Appendix 18 for a derivation of f_{max} using two-port network analysis of an extrinsic transistor.

The important point is that both f_T and f_{max} should be considered only as qualitative indicators of the frequency response of a transistor. There are many other elements that can impact the performance of a transistor, and the magnitude of the impact depends on the circuit application and on the design point of the transistor, which will be discussed further later.

8.1.3 Ring Oscillator and Gate Delay

For large-signal digital- or logic-circuit applications, neither f_T nor f_{max} is really a good indicator of device performance (Taylor and Simmons, 1986). For a digital circuit, the gate delay itself is often used as a figure of merit for the transistors in the circuit. (For digital circuits, the terms "circuit" and "gate" are used interchangeably.)

Since the merit of a transistor is reflected in the switching speed of the circuit in which the transistor is used, the merit of a transistor therefore depends on the circuit and its design point. That is, ***a transistor optimized for one circuit and its design point may not be optimum for another design point, and certainly not for another circuit.*** For high-performance logic applications, the most commonly used bipolar circuit is *the emitter-coupled logic* or *ECL* circuit. Most publications on digital bipolar transistor technology quote the measured or modeled ECL gate delay, often with negligible external

capacitance loading, as a figure of merit. In this chapter, an ECL gate is used to illustrate the optimization of a bipolar transistor for digital-circuit applications.

As explained in Section 5.3.1, the switching speed of a logic gate can be measured very easily from a ring-oscillator arrangement of the circuit. For almost all logic circuits, a ring oscillator consists of an odd number of stages of the logic gate connected with the output of one stage feeding the input of the next stage, and the output of the last stage feeding the input of the first stage, thus forming a ring configuration. The average switching delay of a circuit can be measured directly by measuring the period of its ring-oscillator waveform. This average delay is equal to $P/2n$, where P is the period, and n the number of stages, of the ring oscillator. However, for some circuits, such as a bipolar ECL circuit (see next section), which have both an inverted output and a noninverted output, a ring oscillator can be formed by using an even number of stages. In this case, the inverted outputs from one half of the stages and the noninverted outputs from the other half of the stages are used to form the ring. The average stage delay is still given by $P/2n$.

8.2 Digital Bipolar Circuits

An ECL gate with fan-in of 1 and fan-out of 1 is shown in Fig. 8.3. Both the inverted output V_{out} and the noninverted output $\bar{V}_{out}$ are shown. In this circuit configuration, the voltage V_S and the resistor R_S together set the *switch current* I_S of the ECL gate. This current is constant, i.e., it does not change when the circuit switches. The two resistors R_L are the load resistors of the gate. The capacitor C_L represents the total external load capacitance connected to the output of the gate. The two resistors R_E together with transistors Q_3 and Q_5 form the two emitter followers. (In an emitter follower, the emitter voltage follows the base voltage. Thus, if the base voltage of Q_3 goes up, the emitter voltage of Q_3 also goes up by the same amount.) The input voltage V_{in} and the output voltages V_{out} and $\bar{V}_{out}$ swing above and below a fixed reference voltage V_{ref}, usually approximately symmetrically, by one-half of the logic swing ΔV.

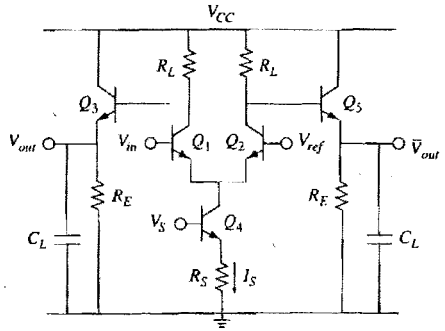


Figure 8.3. Schematic of an emitter-coupled logic gate for fan-in = fan-out = 1 and an output capacitance loading of C_L. Both the inverting and the noninverting outputs are shown.

When V_{in} is high, i.e., when $V_{in} = V_{ref} + \Delta V/2$, transistor Q_1 is turned on much harder than the reference transistor Q_2. As a result, the switch current I_S flows mainly through transistor Q_1 and its load resistor R_L. The IR drop across this load resistor in turn lowers the base voltage of transistor Q_3. The output voltage V_{out} follows the base voltage of Q_3 and hence becomes low. At the same time, with negligible current flowing through transistor Q_2 and its load resistor R_L, the base voltage of transistor Q_5 is pulled up to a high voltage. The output voltage $\bar{V}_{out}$ follows the base voltage of Q_5 to high. Thus, V_{out} is inverting, while $\bar{V}_{out}$ is noninverting. A similar analysis shows that when V_{in} is switched to low, i.e., when $V_{in} = V_{ref} - \Delta V/2$, V_{out} is switched to high and $\bar{V}_{out}$ is switched to low.

When the gate switches, the switch current I_S is steered, or switched, from one load resistor to the other. (For its current-switching characteristics, an ECL gate is sometimes called a *current-switch emitter-follower circuit*.) The logic swing ΔV is equal to the IR drop in one of the load resistors, i.e., $\Delta V = I_S R_L$.

Instead of using a fixed voltage as the reference for the input signal, the inversion of the input signal can be used as the reference. That is, in Fig. 8.3, V_{ref} can be replaced by $\bar{V}_{in}$. For instance, if V_{in} changes from 0 to say +200 mV, then $\bar{V}_{in}$ changes from 0 to −200 mV. An ECL circuit having an inverted input signal as the reference voltage is called a *differential-ECL* or *differential-current-switch circuit*. With V_{in} and $\bar{V}_{in}$ moving in opposite directions, V_{in} in a different-ECL circuit needs to swing only 200 mV to result in the same change in transistor current as a 400 mV swing in a regular ECL circuit. That is, a differential-ECL circuit can have a logic swing that is one half that of a regular ECL circuit. Compared to a regular ECL circuit, the relatively small signal swing of a differential-ECL leads to superior speed and lower power dissipation (Eichelberger and Bello, 1991). However, the connection for V_{in} and that for $\bar{V}_{in}$ must be routed together on a chip. As a result, it takes more wiring channels and/or wiring levels to wire up a chip using differential ECL circuits than ECL circuits.

8.2.1 Delay Components of a Logic Gate

It has been shown (Tang and Solomon, 1979; Chor *et al.*, 1988) that the switching delay of a bipolar logic gate can be expressed as a linear combination of all the time constants of the circuit, with each time constant weighed by a factor that is determined by the detailed arrangement of the circuit. For the ECL gate depicted in Fig. 8.3, the switching delay can therefore be written as

$$T_{delay} = \sum_i K_i R_i C_i + \sum_j K_j \tau_j, \tag{8.11}$$

where the first sum is over all the resistances and capacitances of the transistors in the circuit, the second sum includes the forward and reverse transit times of the transistors, and K_i and K_j are the corresponding weighing factors. Since the transistors in an ECL circuit are all biased in the forward-active mode (i.e., the emitter–base diodes are zero-biased or forward biased, and the base–collector diodes are zero-biased or reverse biased), the reverse transit times, which are associated with forward-biased base–collector diodes, are zero. Only the forward transit times need to be included in Eq. (8.11).

For simplicity of discussion, it is often assumed that all the transistors in the circuit are the same. In this case, Eq. (8.11) is reduced to (Chor *et al.*, 1988)

$$\begin{aligned} T_{delay} = {} & K_1\tau_F + r_{bi}\left(K_2 C_{dBCi,tot} + K_3 C_{dBCx,tot} + K_4 C_{dBE,tot}\right) \\ & + r_{bx}\left(K_6 C_{dBCi,tot} + K_7 C_{dBCx,tot} + K_8 C_{dBE,tot}\right) \\ & + R_L\left(K_{10} C_{dBCi,tot} + K_{11} C_{dBCx,tot} + K_{12} C_{dBE,tot} + K_{13} C_{dCS,tot} + K_{14} C_L\right) \\ & + r_c\left(K_{15} C_{dBCi,tot} + K_{16} C_{dBCx,tot} + K_{18} C_{dCS,tot}\right) \\ & + r_e\left(K_{19} C_{dBCi,tot} + K_{20} C_{dBCx,tot} + K_{21} C_{dBE,tot} + K_{23} C_{dCS,tot} + K_{24} C_L\right) \\ & + C_{DE}\left(K_5 r_{bi} + K_9 r_{bx} + K_{17} r_c + K_{22} r_e\right), \end{aligned} \tag{8.12}$$

where the same numbering system for the K-factors as in Chor *et al.* is followed. The internal resistances and capacitances of the transistors are illustrated in Fig. 6.18. The circuit resistances and capacitances are shown in Fig. 8.3. It should be noted that transistor Q_4 and resistor R_S, functioning only to set the switch current, are not involved in the switching of the circuit and hence do not enter into Eq. (8.12).

In practice, the performance of a bipolar logic gate is often characterized as a function of the operating current of its transistors. Since the power dissipation of a logic gate is proportional to the total current passing through the transistors, the performance of a logic gate can also be characterized as a function of its power dissipation. That is, in principle, the delay-versus-current and the delay-versus-power dissipation characteristics contain the same information, and either one can be used to describe the behavior of the transistors in the circuit. The circuit delay-versus-current or delay-versus-power dissipation characteristics are usually obtained by varying the resistor values in the circuit, keeping the transistor geometries and parameters fixed. In so doing, the collector current density becomes proportional to the collector current. In the published literature, sometimes the circuit delay is plotted as a function of the collector current density, and sometimes simply as a function of the collector current. In any event, the delay-versus-current or delay-versus-power characteristics reflects the performance of a *fixed* transistor design as a function of its collector current density.

The relative magnitudes of the delay components represented in Eq. (8.12) have been evaluated for the modern bipolar transistor shown in Fig. 7.29 (Tang and Solomon, 1979; Chor *et al.*, 1988). The results are illustrated schematically in Fig. 8.4. Each delay component depends on a key device or circuit parameter. In the remainder of this subsection, we analyze each of the delay components qualitatively. Such an analysis is very helpful as a guide to optimizing the device design.

8.2.1.1 Transit-Time Delay Component

The first term in Eq. (8.12) is proportional to the forward transit time τ_F. As shown in Section 6.4.4, at low collector currents (hence low collector current densities), where base widening is negligible, τ_F is a constant, independent of I_C. However, once base widening becomes appreciable, τ_F increases with I_C. Therefore, the transit-time delay component of an ECL gate is expected to be independent of current at low collector currents, but to increase with current once the current density exceeds the base-widening

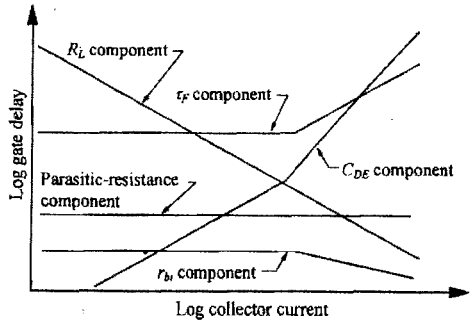


Figure 8.4. Schematic illustration of the relative magnitudes of the gate delay components of an ECL gate and their dependence on collector current or power dissipation.

threshold. This is illustrated in Fig. 8.4. Most high-speed digital circuits are designed to have the transit time as one of the dominant delay components (Tang and Solomon, 1979; Chor *et al.*, 1988).

8.2.1.2 Intrinsic-Base-Resistance Delay Component

The second term in Eq. (8.12) is due to the RC time constants associated with the intrinsic-base resistance. At low collector current densities, the intrinsic-base resistance is a constant, independent of current. However, as can be seen from Eq. (7.5) and Appendix 15, once base widening occurs at high collector current densities, the intrinsic-base resistance decreases rapidly with further increase in collector current density. Therefore, the intrinsic-base-resistance delay component of an ECL gate is independent of current as long as base widening is negligible. Once base widening becomes appreciable, this delay component decreases with further increase in current, as illustrated in Fig. 8.4. The base-resistance delay component is usually quite small (Tang and Solomon, 1979; Chor *et al.*, 1988).

8.2.1.3 Parasitic-Resistance Delay Components

The third, fifth, and sixth terms in Eq. (8.12) are due to the RC time constants associated with the extrinsic-base resistance, the collector resistance, and the emitter resistance, respectively. Since these parasitic resistors, to first order, are all independent of the operating current of a transistor, these delay components are independent of current, as illustrated in Fig. 8.4. The parasitic-resistance delay components are also quite small (Tang and Solomon, 1979; Chor *et al.*, 1988).

8.2.1.4 Load-Resistance Delay Component

The fourth term in Eq. (8.12) is due to all the RC time constants associated with the load resistors R_L. For circuits with a large load capacitance C_L, the load-resistance delay component is often dominated by the R_LC_L term. For this reason, the load-resistance delay component is also referred to as the load-capacitance delay component.

Referring to Fig. 8.3, it can be seen that the logic swing ΔV, the switch current I_S, and the load resistor R_L are interrelated by

$$R_L = \frac{\Delta V}{I_S}. \tag{8.13}$$

Since the logic swing is fixed, the load-resistance delay component of an ECL circuit is inversely proportional to the switch current. This is illustrated in Fig. 8.4. Most ECL circuits are designed to operate at large currents in order to minimize the load-resistance delay component.

8.2.1.5 Diffusion-Capacitance Delay Component

The last term in Eq. (8.12) is associated with the emitter diffusion capacitance C_{DE}. As shown in Eq. (6.114), the stored charge associated with forward-biasing the emitter–base diode can be written as $Q_{DE} = \tau_F I_C = \tau_F I_S$. For modeling purposes, the emitter diffusion capacitance is often approximated by (stored minority-carrier charge)/(average change in input voltage when the gate changes state) $= 2Q_{DE}/\Delta V$ (Tang and Solomon, 1979; Chor *et al.*, 1988), i.e.,

$$C_{DE} \approx \frac{2\tau_F I_S}{\Delta V}. \tag{8.14}$$

Thus, the diffusion-capacitance delay component of an ECL gate is proportional to the switch current at low currents, where base widening is negligible and τ_F is independent of current. At high currents where base widening is appreciable, however, τ_F itself increases with I_S, and the diffusion-capacitance delay component increases in proportion to the product $\tau_F I_S$. This is illustrated in Fig. 8.4.

It should be noted that as long as C_{DE} is proportional to Q_{DE}, the total gate delay obtained from Eq. (8.12) is independent of the exact approximation used for C_{DE}. This is due to the fact that the weighing factors in Eq. (8.12) are obtained from a sensitivity analysis (Tang and Solomon, 1979; Chor *et al.*, 1988). Any "inaccuracy" in the coefficient of proportionality, which is $2/\Delta V$ in Eq. (8.14), is compensated by the corresponding weighing factor obtained from the sensitivity-analysis procedure.

8.2.2 Device Structure and Layout for Digital Circuits

Referring to Eq. (8.12), we see that the RC delay components can be grouped into two categories, namely delay components associated with the intrinsic-device parameters, and delay components associated with the extrinsic-device and circuit parameters. The intrinsic-device parameters are: τ_F, r_{bi}, $C_{dBCi,tot}$, $C_{dBE,tot}$, and C_{DE}. These parameters determine, or are closely related to, the intrinsic properties of the transistors. Designing the intrinsic parts of a transistor and how the intrinsic-device parameters relate to the device characteristics have already been discussed in Chapter 7.

The extrinsic-device and circuit parameters are just as important as the intrinsic-device parameters in determining the measured device and circuit characteristics. Furthermore, the parasitic emitter and base resistances affect the measured current–voltage characteristics

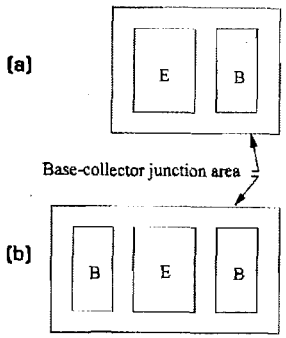


Figure 8.5. Schematics illustrating the layouts of the base–collector diode region for two bipolar transistors. The transistors are of the usual non-self-aligned type, and both transistors have the same emitter area. Layout (a) has base contact on only one side of the emitter, and layout (b) has base contact on both sides of the emitter.

directly, as noted in Section 6.3.1. Thus, the optimal design of the intrinsic-device parameters becomes a function of the extrinsic-device and circuit parameters. For instance, if $C_{dBCx,tot}$ is large compared to $C_{dBCi,tot}$ or $C_{dBE,tot}$, then reducing $C_{dBCi,tot}$ or $C_{dBE,tot}$ is not going to have much effect in improving the speed of a circuit. In general, the extrinsic-device resistance and capacitance of a transistor depend on its physical structure and fabrication process. Therefore, in optimizing the intrinsic-device parameters, we need to specify the device structure and process being used.

It should be noted that the physical structure of a transistor includes its physical layout as well. For the same design of the intrinsic-device parameters, the resultant device characteristics depend on how the transistor is laid out. As an illustration, the plan views of the base–collector diode portion of a non-self-aligned transistor are shown in Fig. 8.5 for two commonly used layouts. Both layouts have the same emitter area, and hence have the same intrinsic device characteristics when operated at the same current. The one in (b) with two base contacts has a much larger extrinsic-base–collector junction area, and hence much larger $C_{dBCx,tot}$, than the one in (a) with only one base contact. However, the base current for layout (b) can flow laterally in two directions, while that for layout (a) can flow in only one direction. As shown in Appendix 16, the intrinsic-base resistance for layout (b) is 1/4 that of layout (a).

In general, if a circuit is designed for low power dissipation, or for operation at low collector current densities, layouts such as that shown in Fig. 8.5(b) result in slower circuits because the reduction in base resistance is not enough to compensate for the increase in collector capacitance. For circuits designed to operate at large power dissipation, or for operation at large collector current densities, however, the reduction in base resistance can more than compensate for the increase in collector capacitance. In this case, layout (b) in Fig. 8.5 gives higher circuit speeds than layout (a) (Ranfft and Rein, 1982).

In order to reduce power dissipation and/or delay, most high-speed bipolar transistors now employ a structure such as the one described in Section 7.5, which has near-minimum

parasitic resistance and capacitance. To further improve circuit speed and/or reduce power dissipation, the intrinsic-device parameters need to be optimized, as discussed in the next section.

8.3 Bipolar Device Optimization for Digital Circuits

In the literature, the switching delay of an ECL gate is often plotted as a function of its power dissipation, or as a function of its switch or collector current. Both the delay-versus-current and the delay-versus-power-dissipation characteristics contain really the same information, since power dissipation is proportional to current. However, it is shown in Chapter 7 that ***the detailed design of a bipolar transistor is closely coupled to its collector current density.*** Therefore, in considering the design of a bipolar transistor, we need to translate the delay-versus-power-dissipation, or delay-versus-current, characteristics to delay-versus-collector-current-density characteristics. In this section, we discuss the optimization of a bipolar transistor for ECL circuits by examining the dependence of the dominant ECL delay components on collector current density.

8.3.1 Design Points for a Digital Circuit

For simplicity, we assume that all the transistors in the ECL circuit carry the same current density, and hence only one device design is needed for all the transistors in the circuit. In practice, this can be achieved easily by varying the emitter area of each transistor in proportion to its current, so that all the transistors in the circuit carry the same current density.

The logic swing of a bipolar digital circuit is usually fixed at some minimum value consistent with the noise-margin requirements. For an ECL gate, the typical logic swing is about 400 mV for a circuit driving other circuits on-chip, and about 800 to 1000 mV for a circuit driving a signal off-chip. The logic swing can be halved if the circuit is designed to operate in a differential mode. The larger off-chip logic swing is due to the noisier off-chip environment. Referring to the circuit configuration in Fig. 8.3, the switch current I_S can be varied readily by adjusting the voltage V_S and/or the resistor R_S. To maintain the same logic swing, the load resistors R_L are also varied according to Eq. (8.13).

It should be pointed out that the 400 mV logic swing of a typical ECL gate (200 mV if the circuit is designed to operate in a differential mode) is significantly smaller than the logic swing of a CMOS circuit, which is the same as its power-supply voltage V_{dd} (see Section 5.1). V_{dd} for CMOS has been decreasing with device scaling from 5 V towards about 1 V. To first order, it is their small logic swings that give bipolar circuits the speed advantage over CMOS circuits for driving heavy load capacitances, such as long wires.

For a given set of transistors in a circuit, as the operating current is changed by changing the resistors, the collector current density of the transistors is changed correspondingly. The expected gate delay as a function of collector current, or current density, is illustrated in Fig. 8.6. For simplicity, only the three dominant delay components in Fig. 8.4, namely the R_L component, the C_{DE} component, and the transit-time component,

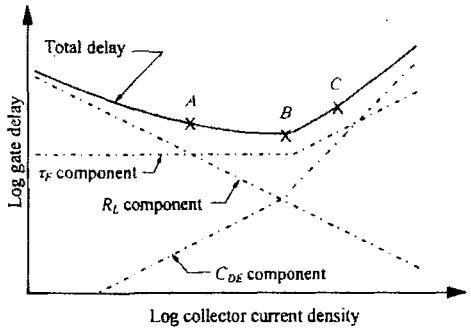


Figure 8.6. Schematic illustration of the dominant delay components, as well as their sum total, of an ECL gate as a function of collector current density. The design points *A*, *B*, and *C* are discussed in the text.

are shown in this illustration. Three possible design points, *A*, *B*, and *C*, are indicated in Fig. 8.6. These design points are discussed further next.

It is clear that design *C* is to be avoided. At *C*, the gate delay is completely dominated by base widening, which causes both the transit-time and the diffusion-capacitance components to increase rapidly with collector current density. The circuit actually runs slower as additional power is dissipated.

The minimum-delay point is *B*. Here the circuit is running at its maximum speed. For applications where speed is the most important consideration and power dissipation is not a factor at all, this is a reasonable design point.

However, if power dissipation is an important factor, then a low-power design point such as *A* is preferred. Design *A* can have a much smaller power–delay product than design *B*. Moving from design *B* to design *A*, a lot of power can be saved for a small increase in circuit delay. The device designs for points *A* and *B* are discussed further next.

8.3.2 Device Optimization When There Is Significant Base Widening

If the load-resistance delay component is large, the gate delay may not reach minimum until the collector current density is so large that there is significant base widening. This is illustrated in Fig. 8.7. To increase the circuit speed further, the transistors should be optimized to reduce base widening.

As discussed in Section 7.3.2, base-widening effects can be reduced by increasing the collector doping concentration and/or reducing the collector layer thickness, provided that the level of base–collector junction avalanche remains acceptable. Alternatively, or concurrently, the transistors can be designed with larger emitter areas to reduce their collector current density. Reducing the collector current density is effective in reducing base widening, as illustrated in Fig. 6.11. Of course, suppressing base widening by increasing the collector doping concentration and/or by increasing the emitter area increases the device capacitance and hence will increase the R_L component of the gate delay. However, often the net result is an increase in circuit speed (Tang and Solomon, 1979).

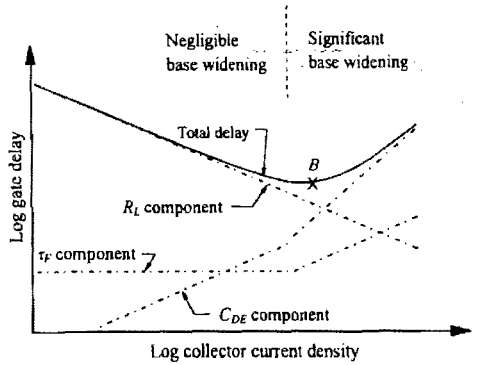


Figure 8.7. Schematic illustrating the characteristics of switching delay versus collector current density of an ECL gate, for a case where the load capacitance is large. The maximum speed occurs at point *B*, where there is significant base widening.

It should be noted that as long as there is significant base widening, the gate delay is not sensitive to the physical thickness of the intrinsic base, which determines the low-current value of the τ_F delay component in Fig. 8.7. Thus, when a transistor is operated in the base-widening mode, efforts to thin down the intrinsic base or to incorporate SiGe into the intrinsic base are not going to be effective in improving circuit speed.

In general, reducing device capacitance always improves circuit speed. However, when C_L is large compared to the total device capacitance, and when there is significant base widening, reducing device capacitance is not effective in improving circuit speed. The only effective method to improve the circuit speed in this case is to first minimize the base widening. Once the base widening is minimized, then efforts to reduce the device capacitance may be effective.

8.3.3 Device Optimization When There Is Negligible Base Widening

If the R_L delay component is relatively small, the gate delay can reach its minimum value at a collector current density where base widening is still negligible. This design point is illustrated in Fig. 8.8. To maximize the circuit speed, both the transit-time and the diffusion-capacitance components should be minimized.

Of course, reducing device capacitance also improves circuit speed. In particular, the collector doping concentration can be reduced to be just large enough to maintain negligible base widening, thereby reducing the collector–base junction capacitance. In so doing, the total delay, particularly in the region to the left side of the design point *B* in Fig. 8.8, can be reduced appreciably (Tang and Solomon, 1979). With the R_L delay component reduced by reducing the device capacitance, the design point *B* can be moved to a lower collector current density, thus improving the speed of the circuit and/or reducing its power dissipation.

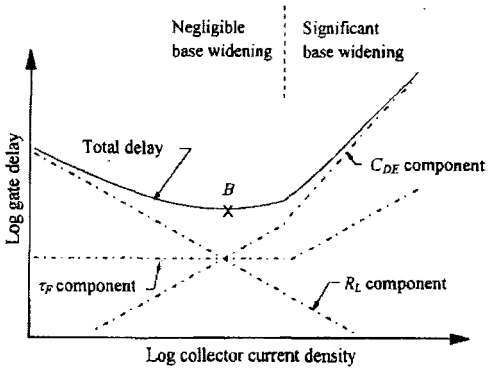


Figure 8.8. Schematic illustrating the characteristics of switching delay versus collector current density of an ECL gate, for a case where the maximum circuit speed is reached before significant base widening occurs.

The maximum speed, however, is still limited by the transit-time and diffusion-capacitance components. Since both of these components are proportional to τ_F, minimizing τ_F will increase the maximum circuit speed. In the rest of this subsection, we discuss how to minimize τ_F and to what extent it can be minimized.

8.3.3.1 Minimizing the Forward Transit Time

The forward transit time τ_F is given by Eq. (6.117), which is repeated here:

$$\tau_F = \tau_E + \tau_B + \tau_{BE} + \tau_{BC}, \tag{8.15}$$

where τ_E is the emitter delay time, τ_B is the base delay time, τ_{BE} is the base–emitter depletion-region delay time, and τ_{BC} is the base–collector depletion-region delay time. The relative contributions of these components to τ_F for a typical self-aligned polysilicon-emitter bipolar transistor have been evaluated (Ashburn, 1988). No one component dominates the total transit time when base widening is negligible, although τ_B and τ_{BC} are often the larger components. [In the literature, often the corresponding transit times, i.e., t_E, t_B, t_{BE} and t_{BC}, are used on the RHS of Eq. (8.15) instead. Depletion-layer delay time and depletion-layer transit time are two different labels for the same physical process, so there is no difference. However, as discussed in Section 6.4.4, there is a subtle difference between t_E and τ_E, and between t_B and τ_B. The transit times t_E and t_B are the average times for filling or emptying the emitter and the base, respectively, of all the excess minority carriers. The delay times τ_E and τ_B, on the other hand, represent only those excess minority carriers in the emitter and the base, respectively, that can contribute to the diffusion capacitance. At low currents, where base widening is negligible, $\tau_B = 2t_B/3$. If we assume the emitter to be wide emitter, then $\tau_E = t_E/2$.]

As shown in Section 7.4, a SiGe-base transistor can have a significantly smaller τ_E and τ_B than a Si-base transistor. As a result, a SiGe-base transistor has a significantly smaller

τ_F, and hence a higher f_T [see Eqs. (8.8) and (8.9)], than a Si-base transistor. The benefit of a higher f_T depends on the circuit family and its design point. For example, for an unloaded ECL gate the benefit of increasing f_T from 48 GHz to 70 GHz is only about 7% for a low-power design point, and about 20% for a high-power design point (Chuang *et al.*, 1992). For a loaded (FI = FO = 3, $C_L = 0.3$ pF) ECL gate, the benefit is less than 3% for the low-power design point, and only 10% for the high-power design point. The reader is referred to the literature (Chuang *et al.*, 1992) for more details and for a discussion on the benefit of higher f_T for other bipolar circuit families.

Of course, a higher f_T also means a higher f_{max} [see Eq. (8.10)]. In the literature, f_{max} is often used to indicate the speed of a transistor with the implication that it is also a good indicator of the speed of a digital circuit employing the transistor. For the example by Chuang *et al.* cited above, an increase of f_T from 48 GHz to 70 GHz should lead to a 20% increase in f_{max} ($f_{max} \propto f_T^{1/2}$), assuming there are no changes in base resistance and base–collector junction capacitance in achieving the increase in f_T. More recent experimental data also indicate a similar relationship (Jagannathan *et al.*, 2003). This suggests that f_{max} is a reasonable indicator of transistor speed for lightly loaded circuits designed to operate at high power or near their maximum-speed point, but not for circuits designed to operate at low power or for circuits having an appreciable load capacitance. In the rest of this subsection, we examine the characteristics of each of the components of τ_F. We also discuss how they can be reduced.

- *Emitter delay time.* As discussed in Section 6.4.4, the emitter delay time for a wide- or deep-emitter is

$$\tau_E(\text{deep-emitter}) = \frac{I_B \tau_{pE}}{2I_C} = \frac{\tau_{pE}}{2\beta_0}, \tag{8.16}$$

where $\beta_0 = I_C/I_B$ is the current gain. It can be seen readily from Fig. 2.24(b) that τ_{pE} can be reduced by increasing the emitter doping concentration. In practice, the emitter of a silicon bipolar transistor is always doped as heavily as possible already. For a typical emitter doping concentration of 2×10^{20} cm^{-3}, Fig. 2.24(b) gives $\tau_{pE} \sim 100$ ps. Therefore, τ_E(deep-emitter) $\sim$ 0.5 ps if $\beta_0 = 100$. The emitter delay time should be smaller for a shallow emitter than for a deep emitter, since the transit time for a shallow emitter is smaller than τ_{pE}. Also, for a polysilicon emitter, the emitter delay time should be smaller than τ_E (deep-emitter), since the current gain can be increased significantly with a polysilicon emitter. For a 1-μm self-aligned polysilicon-emitter n–p–n transistor, $\tau_E \approx 0.6$ ps (Ashburn, 1988). Equation (8.16) indicates that τ_E can be reduced by increasing the device current gain. For a given emitter process used to fabricate a transistor, the current gain can be increased by tailoring the base parameters to increase the collector current. As discussed in Chapter 7, this can be accomplished by increasing the base sheet resistivity and/or by using SiGe-base technology (see Exercise 8.5). However, in practice, a device designer almost never intentionally increases the current gain of a bipolar transistor in order to decrease its emitter delay time, since the emitter delay time is not a dominant component of τ_F. Even with SiGe-base technology, the larger current gain associated with a SiGe-base transistor is usually

traded off for a lower base resistance and/or for a larger Early voltage, instead of used to reduce the emitter delay time.

- *Base–collector depletion-layer delay time.* As mentioned earlier, the terms base–collector depletion-layer delay time τ_{BC} and base–collector depletion-layer transit time t_{BC} refer to the same physical process. Therefore, these two terms are used interchangeably. The base–collector depletion-layer transit time t_{BC} is given approximately by (Meyer and Muller, 1987; see also Exercise 8.6)

$$t_{BC} = \frac{W_{dBC}}{2v_{sat}} \tag{8.17}$$

where W_{dBC} is the width of the base–collector junction depletion layer, and v_{sat} is the saturated velocity of electrons. For a collector with $3 \times 10^{16}\,\text{cm}^{-3}$ doping concentration and reverse biased with respect to the base by 3 volts, W_{dBC} is about 0.4 μm, and hence $t_{BC} \approx 2$ ps if $v_{sat} \approx 10^7$ cm/s is assumed. For a 1-μm self-aligned polysilicon-emitter bipolar device, the value for t_{BC} calculated by numerical simulation is about 1.7 ps (Ashburn, 1988). As the collector doping concentration is increased to minimize base widening, W_{dBC}, and hence t_{BC}, will be reduced. However, in so doing, the base–collector junction depletion-layer capacitance will be increased, and the base–collector junction breakdown voltage will be reduced. In practice, the collector doping profile is usually designed to minimize base widening and to provide adequate breakdown voltage, instead of to reduce t_{BC}.
- *Base–emitter depletion-layer delay time.* The terms base–emitter depletion-layer delay time τ_{BE} and base–emitter depletion-layer transit time t_{BE} also refer to the same physical process, and hence are used interchangeably. t_{BE} is smaller than t_{BC}, since the width of the base–emitter junction depletion layer, W_{dBE}, is much smaller than W_{dBC}. This is due to the fact that, for a transistor biased in the forward-active mode, the potential drop across the base–emitter junction is much smaller than that across the base–collector junction. Furthermore, the base is typically about 10 to 100 times more heavily doped than the collector. The higher the doping concentration, the smaller the depletion-layer width (see Fig. 2.16). For a 1-μm self-aligned polysilicon-emitter bipolar device, the value of t_{BE} calculated by numerical simulation is about 0.7 ps (Ashburn, 1988).

  As will be shown in the next section, the base doping concentration is increased in bipolar device scaling. Therefore, W_{dBE} and t_{BE} are reduced in bipolar device scaling.
- *Base delay time.* At low current densities where base widening is negligible, the base delay time τ_B is equal to 2/3 times the base transit time t_B (see Section 6.4.4), which has been derived and discussed in detail in Section 7.2.4. For a boxlike base doping profile, t_B is given by Eq. (7.24), namely

$$t_B \approx \frac{W_B^2}{2D_{nB}}, \tag{8.18}$$

where W_B is the intrinsic-base width, and D_{nB} is the electron diffusion coefficient in the base. Base doping profiles obtained by ion implantation are seldom boxlike, but often

like a Gaussian distribution. The graded dopant distribution in a Gaussian profile results in a built-in electric field in the intrinsic base. However, as discussed in Section 7.2.3 and illustrated in Fig. 7.3, heavy doping in the intrinsic base of modern bipolar transistors also results in a built-in electric field which compensates substantially the electric field due to the graded dopant distribution. As a result, for modern bipolar transistors, the built-in electric fields in the base due to its nonuniform distribution of dopants have relatively little effect on the base transit time. However, this does not mean that the base transit time is totally independent of the base doping profile. In Eq. (8.18), W_B represents the width of the quasineutral region of the intrinsic base, which is always smaller than the physical base width, which is defined by the separation between where the emitter doping concentration equals the base doping concentration in the emitter–base junction and where the base doping concentration equals the collector doping concentration in the base–collector junction. For the same physical base width, W_B for a boxlike doping profile is larger than W_B for a Gaussian or graded base doping profile. This is due to the fact that for the same collector doping profile, W_{dBC} for a boxlike base doping profile is smaller than W_{dBC} for a Gaussian or graded base doping profile. In any event, Eq. (8.18) can be used to obtain a first-order estimate of t_B. For a transistor with $W_B = 90$ nm and average $N_B = 3 \times 10^{17}\,\text{cm}^{-3}$, Eq. (8.18) gives $t_B = 3.1$ ps, where an electron mobility of $500\,\text{cm}^2/\text{V-s}$, from Fig. 2.24(a), is assumed. Numerical simulation of a comparable Gaussian-like doping profile gives a value of 2.9 ps for t_B (Ashburn, 1988). It is clear from Eq. (8.18) that reducing W_B is an effective way of reducing the base transit time. However, when W_B is reduced, the base doping concentration must be increased appropriately to maintain adequately large emitter–collector punch-through voltage and adequately large Early voltage. As the base doping concentration is increased, the emitter–base junction capacitance will increase, and the emitter–base junction breakdown voltage will decrease. If the breakdown voltage becomes unacceptably low, then it may be necessary to design the base doping profile so that it peaks slightly away from the emitter–base junction. This can be accomplished by inserting an i-layer between the intrinsic-base region and the emitter region (Lu *et al.*, 1990), or by tailoring the base dopant implantation energy. A much more detailed discussion of the dependence of t_B on the base doping profile can be found in the literature (Suzuki, 1991). Also, using SiGe technology for the intrinsic-base layer can reduce the base transit time by a factor of 2 to 3, as shown in Section 7.4.

8.3.4 Device Optimization for Small Power–Delay Product

For circuits designed to operate at low current densities, the gate delay is dominated by the R_L-component. This is illustrated by the design point A in Fig. 8.9. This design point is not meant for maximum circuit speed, but for optimum power–delay tradeoff. To improve circuit speed, the capacitances in the R_L component should be minimized. Referring to Eq. (8.12), we see that these capacitances are the base–collector junction depletion-layer capacitances $C_{dBCi,tot}$ and $C_{dBCx,tot}$, the base–emitter junction depletion-layer capacitance $C_{dBE,tot}$ the collector–substrate junction depletion-layer capacitance $C_{dCS,tot}$, and the load capacitance C_L.

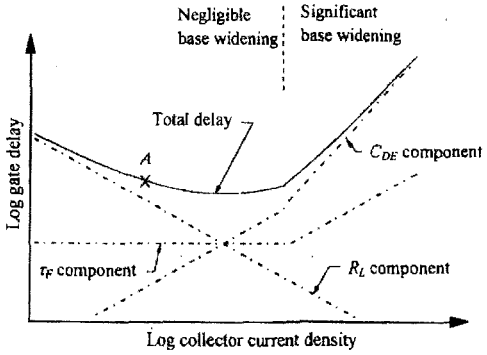


Figure 8.9. Schematic illustrating a design, point A, where the gate delay is dominated by the R_L component.

The base-collector junction capacitance can be reduced by reducing the collector doping concentration. However, the intrinsic-collector doping concentration must be kept high enough to maintain negligible base widening. This can be achieved with the pedestal-collector structure discussed in Section 7.5.

The base–emitter junction capacitance can be reduced by reducing the intrinsic base doping concentration. However, as the base doping concentration is reduced, the base width may have to be increased to avoid emitter–collector punch-through and to avoid the Early voltage becoming unacceptably low. For a SiGe-base transistor, the Early voltage is usually sufficiently large that it is not a design concern (see Section 7.4). As the base width increases, the transit-time delay component increases. Therefore, there will be an optimum base doping concentration for this design point. Alternatively, the base–emitter junction capacitance can be reduced by sandwiching an i-layer between the emitter and the base, as discussed in connection with the base transit time in the previous subsection. In practice, most bipolar transistors are not optimized specially for reducing the base–emitter junction capacitance, other than to use as small an emitter area as possible, consistent with the process technology and the intended collector current density.

The collector–substrate junction capacitance can be reduced by reducing the substrate doping concentration. Consequently, most modern bipolar transistors use lightly doped substrates, typically with a doping concentration of about $1 \times 10^{15}\,cm^{-3}$. In addition, as discussed in Section 7.5, the use of deep-trench isolation, instead of p-diffusion isolation, reduces the collector–substrate and the collector–isolation capacitances very significantly.

The load capacitance consists of two components, namely the interconnect wire capacitance and the input device capacitance. Only the input device capacitance is a function of the device design. It is reduced as the base–emitter junction and the base–collector junction capacitances are reduced. The modern device structure described in Section 7.5 has a base–collector junction area, and hence a base–collector junction capacitance, that is close to minimum. In addition, as discussed in Section 8.2.2, the base–collector diode capacitance can be reduced further if the base is contacted on only one side of the emitter.

8.3.5 Bipolar Device Optimization from Some Data Analyses

It is clear from the above discussion that ***a device optimized for one design point is likely to be nonoptimum for a different design point***. This is demonstrated by the experimental results discussed below.

- *Collector thickness effect.* Figure 8.10 shows the inverter gate delays for two bipolar transistors, one with a collector thickness (i.e., the distance between the intrinsic base and the subcollector) of 270 nm, and one with a collector thickness of 670 nm (Tang *et al.*, 1983). The thin-collector device has a larger base–collector junction capacitance than the thick-collector device. Therefore, at low collector current densities, where there is negligible base widening, the thin-collector device leads to a larger gate delay. However, as the collector current density increases, base widening, and hence emitter diffusion capacitance, increases. At sufficiently large collector current densities, it is the emitter diffusion capacitance that determines the gate delay. When that happens, the thin-collector device, with less collector volume than the thick-collector device for minority-carrier charge storage, leads to faster circuits. Figure 8.10 also demonstrates clearly one very important point about modern bipolar transistors for digital-circuit applications, i.e., the base-widening effect limits the maximum speed of modern bipolar devices.
- *SiGe-base versus Si-base.* Figure 8.11 shows the ECL delays versus current for a Si-base transistor and a SiGe-base transistor of the same design rule and approximately the same base dopant distribution (Harame *et al.*, 1995a). As expected, the SiGe-base transistor is faster than the Si-base transistor, with the speed advantage larger at high

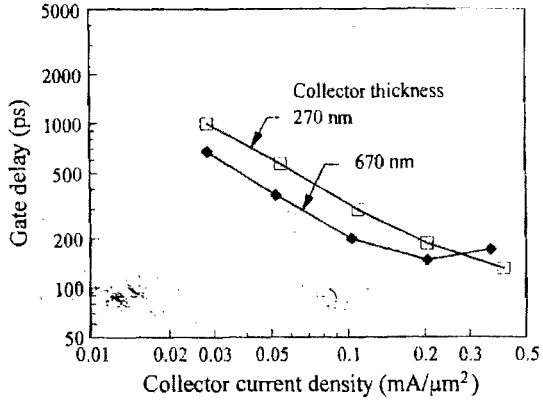


Figure 8.10. Typical switching delay of a bipolar circuit as a function of collector current density, with collector thickness as a parameter. (After Tang *et al.*, 1983.)

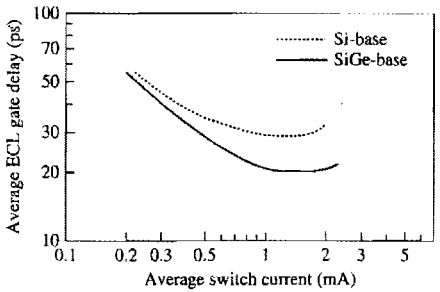


Figure 8.11. Typical comparison of an ECL ring oscillator made of a Si-base transistor to that made of a SiGe-base transistor as a function of collector current. Each ECL gate has FI = FO = 1, and an external wire load of 5 fF. All the transistors have the same emitter area ($0.5 \times 12.5\ \mu m^2$). The SiGe-base devices have a triangular Ge distribution in the quasineutral base. (After Harame *et al.*, 1995a.)

currents than at low currents. However, the maximum speed of both transistors is limited by Kirk effect, as evidenced by the slowdown at large currents. Optimizing the collector design to minimize Kirk effect should improve the maximum performance of both transistors.

- *Emitter length effect.* In general, the emitter stripe width is designed to be as narrow as possible, consistent with the lithography and other processes available. A small emitter stripe width leads to a small intrinsic-base resistance (see Appendix 16). For a given emitter stripe width, the emitter area can be varied by varying the emitter length. Figure 8.12 shows the ECL gate delay versus collector current for two SiGe-base transistors, one having twice the emitter length of the other (Washio *et al.*, 2002). The long-emitter device also has smaller emitter and collector series resistances. To first order, the device depletion-layer capacitance components are proportional to the emitter length, and hence are larger for the device with longer emitter. At low currents where the load-resistor delay component dominates the circuit delay, the smaller capacitance of the short-emitter device leads to a faster ECL gate. At intermediate and large currents, where the delay is no longer dominated by the load-resistor component, the intrinsic-base-resistance, parasitic-resistance, and diffusion-capacitance delay components combine to determine the circuit delay. For the long-emitter transistor, to first order, the decrease in base resistance and other parasitic resistances offsets the increase in the various depletion-layer capacitances. As a result, $r_{bi}C_{dBC,tot}$, $r_{bx}C_{dBC,tot}$, $r_eC_{dBE,tot}$, etc., are about the same for the two transistors. That is, the intrinsic-base-resistance and the parasitic-resistance delay components of the two transistors are about the same. The observed difference in delay for the two ECL gates at intermediate and high currents is due primarily to the difference in their diffusion-capacitance delay components. For a given collector current, the long-emitter device has a current density that is half that of the short-emitter device. That is, the diffusion capacitance of the short-emitter device is at least twice that of the long-emitter device (see Section 8.2.1.5). Once the diffusion-capacitance component dominates the circuit delay, the delay increases with further increase in current, as expected from earlier

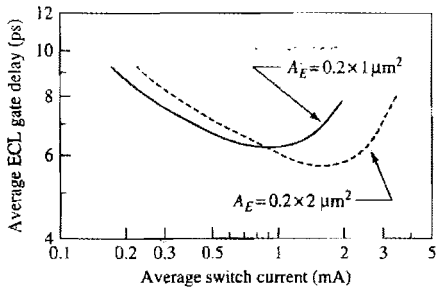


Figure 8.12. Typical delay-versus-current characteristics of an ECL ring oscillator showing the effect of emitter size. The transistors have a SiGe-base. The two transistors have the same device parameters other than emitter length. (After Washio *et al.*, 2002.)

discussions and as shown in Fig. 8.12 to the right of the minimum-delay points. Figure 8.12 shows that the long-emitter device has a smaller minimum gate delay than the short-emitter device.

Figure 8.12 suggests that, if we had plotted the gate delay as a function of collector current density instead of collector current for current densities to the right of the minimum-delay points, the curves for the two transistors would be about the same. This is consistent with the fact that emitter diffusion capacitance is a function of collector current density and collector doping profile, and the two transistors have the same collector doping profile. This result, together with the results discussed earlier in this subsection suggest that the maximum speed of a bipolar transistor is limited by its transit-time delay component, especially if the transistor is operated in a region where base widening is significant. ***Optimization of the collector doping profile and collector thickness to minimize base widening is key to realizing the maximum performance of modern bipolar transistors.***

8.4 Bipolar Device Scaling for ECL Circuits

Since the details of a device design depend on its circuit application, scaling of a bipolar device should be discussed in the context of its circuit application as well. A theory for scaling bipolar transistors for high-performance ECL circuits has been developed by Solomon and Tang (1979). The basic concept in this scaling theory is to reduce the dominant resistance and capacitance components in a coordinated manner so that all the dominant delay components are reduced proportionally as the horizontal dimensions of the transistor are scaled down. In this way, if a transistor is optimized for a given circuit design point before scaling, the transistor remains more or less optimized after scaling. This is accomplished by requiring the capacitance ratio $C_{DE}/C_{dBC,tot}$ and the resistance ratio r_b/R_L to be constant in scaling. Here $C_{dBC,tot} = C_{dBCi,tot} + C_{dBCx,tot}$ and $r_b = r_{bi} + r_{bx}$.

There are several additional constraints in bipolar scaling (Solomon and Tang, 1979). First of all, because of the exponential dependence of current on voltage, the turn-on voltage of a diode is insensitive to the diode area. That is, the diode turn-on voltage is roughly constant in scaling, increasing only about 60 mV for every tenfold increase in its current density. To first order, one can assume the diode turn-on voltage to be constant in scaling. As a result, unlike the scaling of MOSFETs (see Section 4.1), the voltages in a bipolar circuit, including the logic swing ΔV, cannot be reduced in scaling. If the voltages are already optimally small to begin with, then they should remain constant in scaling. Secondly, as explained in Section 6.3.3, the collector doping concentration N_C should be varied in proportion to the collector current density J_C in order to maintain the same degree of base widening in scaling. Thirdly, to avoid emitter–collector punch-through as the base width is reduced in scaling, the base doping concentration must be increased. The base is depleted on the emitter side as well as on the collector side, but the depletion on the emitter side is usually more severe than on the collector side because the emitter is more heavily doped than the base, while the collector is more lightly doped than the base. To avoid excessive base-region depletion near the emitter–base junction, the emitter–base junction depletion-layer width W_{dBE} should remain the same fraction of the base width W_B as W_B is reduced. From the dependence of W_{dBE} on N_B [see Eq. (2.85)], we see that this requirement is met if the base doping concentration N_B is increased so that $N_B \propto W_B^{-2}$.

As shown in Eq. (8.13), for a constant logic swing, R_L is inversely proportional to the switch current I_S. The requirement of r_b/R_L being constant means that r_b should be varied inversely proportional to I_S as well, which would greatly complicate the device layout and design, and would also greatly narrow the device design window. It is much more practical to drop this resistance-ratio requirement and only keep the capacitance-ratio requirement in scaling. This approximation is quite reasonable, since, as discussed in Section 8.2.1, the r_b-component of the gate delay is relatively small to start with. Furthermore, as shown in Appendix 16, the base resistance can be reduced readily, if desired, by modifying the physical layout of the transistor.

Actually, as will be shown in the next subsection, I_S is often kept constant in scaling in order to achieve circuit delay reduction in proportion to the emitter-stripe width. (This scaling objective is analogous to the constant-field scaling of MOSFETs, the results of which are shown in Table 4.1.) In this case, to maintain a constant logic swing in scaling, R_L is also kept constant. Therefore, the ratio r_b/R_L is constant if r_b is kept constant. As shown in Appendix 16, for a given emitter geometry, r_b is constant if the intrinsic-base sheet resistivity is constant. In the case of a Si-base transistor, the intrinsic-base sheet resistivity is indeed often maintained around 10 K Ω/□, partly to maintain a current gain of about 100 and partly to maintain a sufficiently large Early voltage. In the case of a SiGe-base transistor, the intrinsic-base sheet resistivity is usually smaller, typically in the 3–5 K Ω/□ range. That is, without requiring special effort, r_b / R_L is more-or-less constant in the scaling of bipolar devices for high-speed digital applications.

8.4.1 Device Scaling Rules

The scaling constraints, together with the requirement on the capacitance ratio in scaling, are summarized in Table 8.1. The resulting scaling rules for the device design and circuit

Table 8.1 Constraints and Requirements in ECL Scaling

Parameter	Constraint or Requirement
Voltage	V, ΔV = constant
Capacitance	$C_{DE}/C_{dBC,tot}$ = constant
Base doping concentration	$N_B \propto W_B^{-2}$
Collector doping concentration	$N_C \propto J_C$

Table 8.2 Scaling Rules for ECL Circuits

Parameter	Scaling Rule*
Feature size or emitter-stripe width	$1/\kappa$
Base width W_B	$1/\kappa^{0.8}$
Collector current density J_C	κ^2
Circuit delay	$1/\kappa$

* Scaling factor $\kappa > 1$.

delay are summarized in Table 8.2. These rules are for the case where the ECL gate delay is reduced in proportion to the emitter-stripe width of the transistors, or in proportion to the minimum lithographic feature size (Solomon and Tang, 1979).

8.4.1.1 A Qualitative Derivation of the ECL Scaling Rules

The scaling rules shown in Tables 8.1 and 8.2 can also be "derived" qualitatively as follows. The emitter–base and base–collector junctions can be approximated by one-sided diodes. With the power-supply voltage V and the logic swing ΔV held constant, the voltages across these junctions do not change in scaling. As a result, the capacitance per unit area for the emitter–base junction, as given by Eqs. (2.83) and (2.85), is simply proportional to $N_B^{1/2}$. The device areas decrease as $1/\kappa^2$. Therefore, if N_B is increased in proportion to κ^2 to avoid emitter–collector punch-through, then the emitter–base junction capacitance decreases as $1/\kappa$. (The more accurate scaling rules of Solomon and Tang, shown in Tables 8.1 and 8.2, suggest that N_B increases as $\kappa^{1.6}$ for the case where W_B decreases as $1/\kappa^{0.8}$.) Similarly, as N_C is increased in proportion to κ^2 in scaling to maintain the same degree of base widening, the base–collector junction capacitance also decreases as $1/\kappa$ in scaling.

It is shown in Section 5.2.4 that the wiring capacitance per unit length is approximately constant, independent of the wire physical dimensions. Therefore, the capacitance due to wire loading decreases as $1/\kappa$ in scaling. This fact, together with the scaling properties of the device capacitance components discussed above, suggests that the total capacitance C in a bipolar circuit decreases as $1/\kappa$ in scaling.

The gate delay scales as $C\Delta V/I$, where I is the device current charging and discharging the capacitance C. Since C scales as $1/\kappa$ and ΔV is constant, the gate delay scales as $1/\kappa$ if I is held constant. In other words, in order to obtain a gate delay decreasing as $1/\kappa$ in scaling, I should be kept constant, and hence the current density should increase as κ^2, as indicated in Table 8.2.

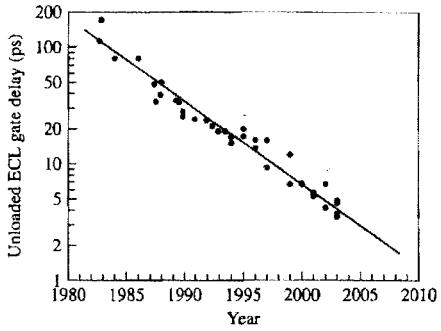


Figure 8.13. Reported ECL gate delays over time.

8.4.1.2 ECL Circuit Scaling in Practice

In practice, device designers do not follow any scaling rules exactly in designing transistors for product applications. Nonetheless, the Solomon–Tang ECL scaling rules have provided a guide to identifying the important delay components in bipolar device scaling. Once these important delay components have been identified, device designers can then focus on minimizing them. As lithographic dimensions were reduced over time, and advances in process technology, such as SiGe-base to reducing base transit time and emitter delay time and incorporation of carbon in the base to minimize boron diffusion, are developed, the reported best-of-breed ECL delays were also reduced steadily over time, as expected from scaling. This is illustrated in Fig. 8.13, which was compiled from published data in the literature.

8.4.2 Limits in Bipolar Device Scaling for ECL Circuits

In this subsection, we examine the limits in scaling bipolar devices for ECL circuits. We do this by examining the scaling constraints in Table 8.1 and the scaling rules in Table 8.2, and by examining the implications of these constraints and rules on the design of small-dimension bipolar transistors.

8.4.2.1 Collector-Current-Density Limit

The scaling rules in Table 8.2 suggest that in order to reduce the delay of an ECL circuit in scaling in proportion to the minimum lithographic feature size, the collector current density should be increased in proportion to κ^2. This in turn suggests that the collector doping concentration should be increased as κ^2 in scaling in order to maintain the same degree of base widening.

The constant power supply voltage in bipolar scaling implies that the maximum reverse-bias voltage across the collector–base junction does not change in scaling. The increasing collector doping concentration therefore could lead to a rapid increase in base–collector junction avalanche. Thus, ***base–collector junction avalanche limits how far a bipolar transistor can be scaled down in physical dimensions and can still yield proportionally large speed improvements.***

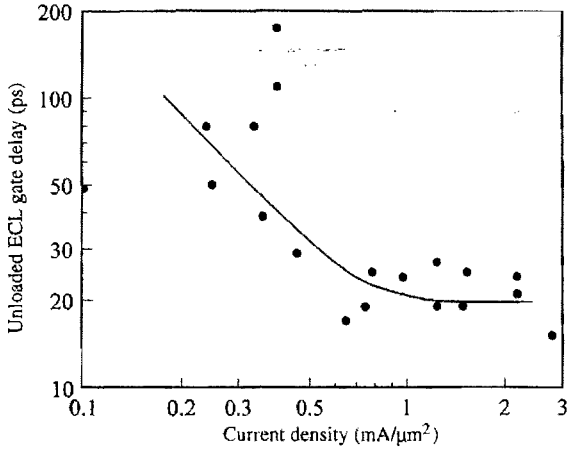


Figure 8.14. Reported ECL circuit delays plotted as a function of the quoted current densities. (After Warnock, 1995.)

Several design approaches for reducing base–collector junction avalanche have been discussed in Section 7.3. In order to maintain acceptable base–collector junction avalanche characteristics, bipolar transistors of small dimensions are often designed with their collector doping concentration lower than suggested by scaling. As a result, these devices have more severe base-widening problems, and the circuit delays tend to saturate with collector current density. This is illustrated in Fig. 8.14, where the reported delays of many Si-base ECL circuits are plotted as a function of the quoted collector current densities (Warnock, 1995). The same speed saturation phenomenon occurs in SiGe-base circuits as well, as suggested by the results in Fig. 8.12.

8.4.2.2 Limitation Due to Device Breakdown

With the power supply voltage held constant, the minimum breakdown voltage required for the proper operation of a bipolar transistor does not change in scaling. As discussed in Section 6.5, the most important breakdown voltage to consider is BV_{CEO}, which is related to but is much smaller than BV_{CBO}. In the literature, there are many reports of high-speed bipolar devices with BV_{CEO} of less than 2.0 V (see Fig. 6.24). If these devices were optimally designed already, then they cannot be scaled much further, if at all.

It should be pointed out that there are many reports of ECL circuits running at "record" speeds in the literature. Unfortunately, many of these reports do not state clearly if the bipolar transistors also have acceptable breakdown voltages. Without such information, it is impossible to judge the significance of the record speeds claimed. However, one point for sure is that transistors for circuits that operate with a smaller power supply voltage can be scaled to larger collector current densities than those for circuits that operate with a larger power supply voltage.

8.4.2.3 Limitation Due to Emitter Series Resistance

It was shown in Section 6.3.1 that the emitter series resistance r_e introduces an emitter–base diode voltage drop of $-I_E r_e = (I_C + I_B) r_e$ and causes the I_C-versus-V_{BE} characteristics to saturate rapidly once this voltage drop is larger than about 60 mV. As an emitter is scaled down in size, the emitter series resistance increases in inverse proportion to its area. Therefore, the emitter–base diode voltage drop due to emitter series resistance increases rapidly in scaling, and could severely limit the current-carrying capability of small-emitter-area transistors.

Fortunately, for an ECL gate, the switching delay attributable to emitter series resistance is relatively small. For a properly optimized 0.5-μm self-aligned bipolar transistor, with an emitter series resistance of 70 Ω, the ECL gate delay attributable to r_e is only about 10% of the total gate delay (Chor *et al.*, 1988).

Also helping to alleviate the emitter-series-resistance problem is the fact that the collector-current-density limit discussed earlier forces designers to use emitters that are significantly larger than the minimum allowed by the process technology, in order to minimize the much more detrimental effects of base widening. A relatively narrow emitter stripe is often employed, with the total emitter area determined by the desired operating current and the current-density limit. The stripe-emitter geometry is chosen to reduce base resistance (see Appendix 16).

8.4.2.4 Limitation Due to Power Density

With the current density increasing as κ^2, the device currents are in effect kept constant in scaling. This implies that the power dissipation of the circuit, which is simply the total current times the power supply voltage, is approximately constant. That is, the power density increases as κ^2 in scaling. Thus, if a 1000-circuit bipolar chip at 1-μm design rules dissipates 3 W, the same-size chip at 0.1-μm design rules could hold 100 000 circuits, with each circuit running 10 times as fast, but the chip would dissipate 300 W.

Unlike a CMOS gate, which can be designed to dissipate negligible power during standby (see Section 5.1.1), a bipolar circuit typically dissipates about the same power whether it is in standby or switching. It is this large standby power dissipation of a bipolar circuit that makes the averaged power dissipation of a bipolar chip much larger than that of a CMOS chip for the same logic function. The very high averaged power dissipation of bipolar circuits has limited bipolar chips to circuit integration levels that are small compared to CMOS chips. As a result, bipolar technology has lost out to CMOS for digital VLSI applications where system-level speed is determined not just by transistor speed, but by chip-level integration and package-level integration as well. Since the mid 1990s, CMOS has replaced bipolar for building all computers, including high-end mainframe computers. Today, designers no longer develop bipolar technology for digital VLSI application. Rather, bipolar technology is developed primarily for RF and analog applications where designers often prefer bipolar devices, particularly SiGe-base bipolar devices, over CMOS devices because of the superior characteristics of bipolar devices for these applications. In that case, designers prefer BiCMOS technology (bipolar and CMOS integrated on the same chip) where the CMOS is used primarily for the digital logic functions and the bipolar is used primarily for the analog and RF functions.

8.5 Bipolar Device Optimization and Scaling for RF and Analog Circuits

In general, the techniques used to optimize a bipolar transistor for digital circuits, such as minimizing the base resistance and collector capacitance, are also applicable to optimizing a bipolar transistor for RF and analog circuits. Also, the current-density limit due to base widening applies to transistors for digital as well as RF and analog circuits. However, there are some device design differences between digital circuits and RF or analog circuits.

For a digital circuit, the overall circuit switching speed and power dissipation are the important factors governing the device design, provided that the design meets the breakdown voltage requirements of the circuit. For RF and analog circuits, perhaps the most important device parameters or figures of merit are the cutoff frequency f_T, the maximum oscillation frequency f_{max}, the base resistance, and the Early voltage. The cutoff frequency and the maximum oscillation frequency should be high, the base resistance should be small, and the Early voltage should be large. The importance of high f_T, high f_{max} and small base resistance is obvious, especially for high-frequency RF circuits. Perhaps less obvious is the desire for large Early voltage.

In this section, we first consider a bipolar transistor as an amplifier for examining the merit of large Early voltage. We then discuss how various parameters or figures of merit can be optimized in general, and how they can be traded off among one another. We also discuss how the technology of growing the intrinsic base epitaxially can lead to superior device characteristics.

8.5.1 The Single-Transistor Amplifier

Important insights into the figures of merit of an analog transistor can be obtained by examining the single-transistor amplifier. The circuit configuration of a bipolar transistor biased to operate as an amplifier is shown in Fig. 8.15(a), where R_L is the load resistor. Notice that an increase in the input voltage causes a decrease in the output voltage. For

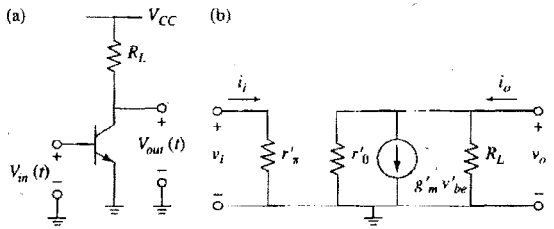


Figure 8.15. (a) Circuit configuration of a single-transistor bipolar amplifier. (b) The small-signal low-frequency equivalent circuit.

simplicity, let us consider the low-frequency case. In this case, the small-signal equivalent circuit in Fig. 6.19 can be adapted to give the small-signal equivalent circuit shown in Fig. 8.15(b) for the amplifier. If no load is attached to the output, the output voltage is

$$v_o = -i_o(r_0' \parallel R_L) = -g_m'(r_0' \parallel R_L)v_i, \tag{8.19}$$

where $(r_0' \parallel R_L)$ denotes the resistance of the two resistors r_0' and R_L in parallel. The minus sign is for the fact that a positive v_i leads to a negative v_o. Thus, the open-circuit or unloaded voltage gain is

$$a_v = \frac{v_o}{v_i} = -g_m'(r_0' \parallel R_L). \tag{8.20}$$

The maximum low-frequency open-circuit voltage gain is given by letting R_L be very large, i.e.,

$$\lim_{R_L \to \infty} a_v = -g_m' r_0'. \tag{8.21}$$

For a bipolar transistor, substituting Eqs. (6.99) and (6.101) into Eq. (8.21) gives

$$\lim_{R_L \to \infty} a_v = -\frac{qV_A}{kT}, \tag{8.22}$$

where V_A is the Early voltage. That is, the intrinsic voltage amplification capability of a bipolar transistor is proportional to its Early voltage, provided that the transistor has sufficiently large breakdown voltages to handle the amplified voltage. For a Si-base bipolar transistor, V_A is typically about 40 V (see Section 6.3.2.1), implying an intrinsic voltage gain of about 1600. The intrinsic voltage gain of a bipolar transistor is significantly larger than that of a MOSFET which is about 17 for a 0.1 μm nMOSFET (see Section 5.4.1.2).

For an analog circuit where large voltage amplification is required, a transistor with a large Early voltage is desired. For high-frequency RF circuits, transistors having high f_T and high f_{max} are needed. As can be seen from Eq. (8.10), small base resistance and small base–collector junction capacitance lead to high f_{max}. As discussed in Section 7.4, a SiGe-base bipolar transistor is far superior to a Si-base bipolar transistor in all these figures of merits.

8.5.2 Optimizing the Individual Parameters

The frequency response, the base resistance, and the Early voltage are all closely coupled. In a practical device design, they are and should be considered together. Here we describe how each of them can be optimized and discuss how the optimization of one may adversely affect the others.

8.5.2.1 Maximizing the Cutoff Frequency

The cutoff frequency is given by Eq. (8.8) or (8.9). To maximize f_T, the capacitances $C_{dBE,tot}$ and $C_{dBC,tot}$ and the forward transit time τ_F should all be minimized. The simplest

way to minimize these capacitances is to use advanced device structures that have small parasitic capacitance, the same as for digital circuits (see Section 7.5). The techniques for minimizing τ_F have already been discussed in Section 8.3.3, in connection with the device optimization for digital circuits. The most important component of τ_F is the base transit time, which can be reduced effectively by reducing the intrinsic-base width W_B. However, reducing W_B alone will lead to a larger intrinsic-base resistance and a smaller Early voltage.

It should be noted that τ_F is a function of both the intrinsic-device doping profiles and the collector current density. It is independent of the horizontal device dimensions and geometry. That is, for a given vertical device doping profile, a large-emitter device has the same τ_F as a small-emitter device, if both devices are operated at the same collector current density. Equation (8.8) shows that the maximum intrinsic cutoff frequency is determined by τ_F. Therefore, in theory, the maximum cutoff frequency of a transistor should be independent of its emitter area, but dependent on its vertical doping profile. However, to reach the maximum f_T value determined by τ_F, the transistor must be operated at sufficiently high current such that the term $kT(C_{dBE,tot}+C_{dBC,tot})/qI_C$ is small compared to τ_F. Unfortunately, base widening becomes important at large collector current densities. Once base widening becomes appreciable, τ_F increases rapidly with further increase in current density. When base widening happens, instead of decreasing with increasing current, the measured f_T actually decreases rapidly with further increase in current. This is illustrated in Fig. 8.16, where the measured f_T as a function of collector current is shown for two transistors of different collector doping concentrations (Crabbé *et al.*, 1993a). It clearly shows that the f_T rolloff characteristics shift towards higher current in proportion to the increased collector doping concentration, as expected from base-widening effects. It also shows that a simple way to increase the peak f_T of a transistor is to increase its

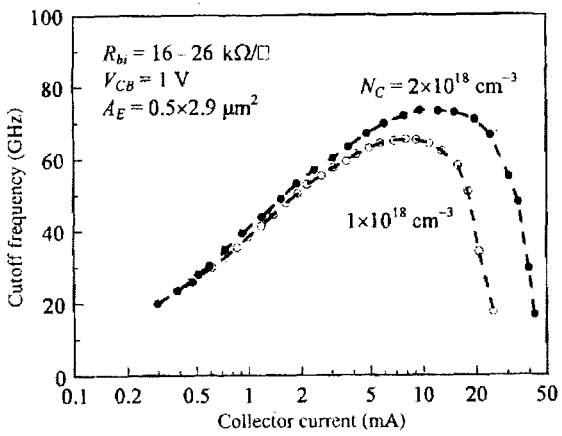


Figure 8.16. Typical measured cutoff frequency as a function of collector current, with collector doping concentration as a parameter. (After Crabbé *et al.*, 1993a.)

collector doping concentration, provided that the device breakdown voltages remain acceptable.

The cutoff frequency is degraded by parasitic emitter and collector resistances r_e and r_c, respectively [see Eq. (8.9)]. Thus, to maximize f_T, it is important that the emitter and collector parasitic resistances are kept small. For a modern bipolar transistor, $C_{dBC,tot}$ is typically about 5 fF. Therefore, the amount of degradation caused by r_e and r_c should be quite small.

8.5.2.2 Minimizing the Intrinsic-Base Resistance

As shown in Appendix 16, the intrinsic-base resistance r_{bi} for an emitter stripe of width W and length L, is proportional to $(W/L)R_{Sbi}$, where R_{Sbi} is the sheet resistivity of the intrinsic-base layer. Thus the intrinsic-base resistance can be reduced by making the emitter stripe as narrow as possible and using long emitter stripes. Emitter stripes less than 0.2 μm wide can now be made readily. Furthermore, contacting the intrinsic base on both sides of an emitter stripe, instead of simply on one side, reduces r_{bi} by a factor of four. In practice, all high-performance bipolar transistors have base contacts on both sides of their emitter stripes.

With polysilicon emitter, current gain is usually not an issue. As a result, the intrinsic-base layer can be doped rather heavily to reduce R_{Sbi}, which in turn reduces r_{bi}. However, if the base doping concentration is too high, emitter–base band-to-band tunneling current could degrade the current gain at low and moderate currents (see Section 6.3.4). If necessary, the peak of the base dopant distribution can be located somewhat deeper than the emitter junction to minimize excessive emitter–base tunneling current. This approach will reduce emitter–base junction depletion-layer capacitance as well.

The design window in the tradeoff between f_T and r_{bi} can be enlarged by using a boxlike instead of a graded intrinsic-base doping profile. It was shown in Section 7.2 that, as a result of heavy-doping effect, the built-in electric field in the intrinsic base of a modern bipolar transistor is not significant whether the intrinsic-base doping profile is graded or boxlike. Thus, for the same W_B and peak base doping concentration, a boxlike intrinsic-base doping profile and a graded intrinsic-base doping profile should give about the same maximum f_T. The boxlike base profile, however, has a larger majority-carrier hole charge, and hence a smaller R_{Sbi} and r_{bi}, than a graded base profile.

8.5.2.3 Maximizing the Maximum Oscillation Frequency

The maximum oscillation frequency is given by Eq. (8.10). It is a function of f_T, r_b, and $C_{dBC,tot}$. Therefore, designs that increase f_T at the expense of increasing r_b could result in a decrease in f_{max}. In fact, if a slightly reduced f_T allows a significantly reduced r_b, the net result could be a larger f_{max}.

Reducing $C_{dBC,tot}$ increases f_{max}. However, if $C_{dBC,tot}$ is reduced by reducing the collector doping concentration, base widening will set in at a lower collector current density, which in turn can reduce the maximum f_T of the transistor (see Fig. 8.16). Thus, maximizing f_{max} is a complex tradeoff process. This point will be discussed further in the next subsection.

8.5.2.4 Maximizing the Early Voltage

The Early voltage for a uniformly doped intrinsic base is give by Eq. (6.72). It is the ratio of the majority-carrier hole charge per unit area, Q_{pB}, to the intrinsic-base–collector depletion-layer capacitance per unit area, C_{dBCi}. An obvious way to increase the Early voltage is to increase the majority-carrier hole charge. If this is done at a fixed W_B by increasing the base doping concentration, then it has relatively little effect on τ_F, and it will reduce r_{bi}. It will have relatively little effect on f_T, but it will lead to a higher f_{max}. However, it will also reduce the current gain, since, as shown in Eq. (7.7), the collector current, hence the current gain, is proportional to the intrinsic-base sheet resistivity. This should not be much of an issue because, as discussed earlier, current gain is usually not an issue for polysilicon-emitter transistors. On the other hand, if the majority-carrier hole charge is increased by increasing W_B, it will decrease f_T, and could adversely affect f_{max} as well. It will definitely reduce the current gain.

The Early voltage can also be increased by reducing C_{dBCi}. As long as base widening is not an issue, this capacitance can be reduced by reducing the collector doping concentration and/or increasing the collector layer thickness.

For a given collector doping profile, a boxlike intrinsic-base doping profile has a larger C_{dBCi} as well as a large Q_{pB} than a graded base doping profile of the same width and same peak doping concentration. However, the increase in C_{dBCi} is not enough to offset the increase in Q_{pB}, and the net effect is that the Early voltage, $V_A = Q_{pB}/C_{dBCi}$, is larger for a boxlike base doping profile than for a graded base doping profile (see Exercise 8.8).

8.5.3 Technology for RF and Analog Bipolar Devices

It is apparent from the above discussion that ***a boxlike intrinsic-base doping profile is preferred for RF and analog circuits***. It provides a larger device design window, and hence allows more optimized tradeoffs to be made among the various device parameters. In practice, a boxlike intrinsic-base doping profile can be obtained readily by using an epitaxially grown intrinsic-base layer. The intrinsic base is doped *in situ* with boron during growth, instead of doped subsequently by ion implantation or diffusion. By incorporating carbon into the base layer, diffusion of boron in the base layer can be suppressed during subsequent process steps (Stolk *et al.*, 1995; Lanzerotti *et al.*, 1996). Using a boxlike base doping profile, a design tradeoff has been made (Yoshino *et al.*, 1995). The results are shown in Figs 8.17 and 8.18. These results clearly show that f_T can be traded off for a larger f_{max} and/or a larger V_A.

It is shown in Section 7.4 that, compared to the Si-base transistor, a SiGe-base transistor has significantly larger collector current for the same base–emitter forward bias, significantly smaller base transit time, and significantly larger Early voltage. The larger collector current can be traded off for a smaller base resistance. Also, as shown in Section 8.3.3, a SiGe-base transistor has a smaller emitter delay time than a Si-base transistor because of its larger current gain. Thus, SiGe-base technology is superior to Si-base technology for RF and analog circuit applications. Since the development of incorporating carbon into the base layer to suppress boron diffusion (Stolk *et al.*, 1995;

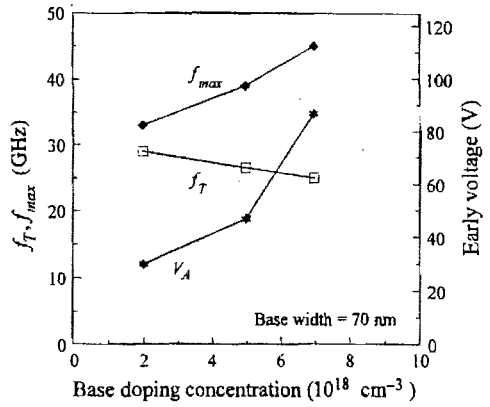


Figure 8.17. Experimental results showing the dependence of f_T, f_{max}, and V_A on base doping concentration N_B. The base has a boxlike doping profile, formed by epitaxial growth of the intrinsic-base layer. (After Yoshino *et al.*, 1995.)

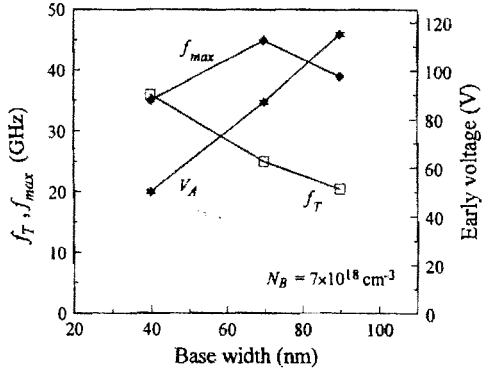


Figure 8.18. Experimental results showing the dependence of f_T, f_{max}, and V_A on intrinsic-base layer thickness. The intrinsic base is formed by epitaxial growth of silicon. (After Yoshino *et al.*, 1995.)

Lanzerotti *et al.*, 1996), the most advanced SiGe-base bipolar transistors are made with carbon in the base layer as well. The reader is referred to Section 7.4.6 for an in-depth discussion on the optimization of the Ge profile in a SiGe-base transistor.

8.5.4 Limits in Scaling Bipolar Transistors for RF and Analog Applications

Although quantitatively the impacts are different, the physical mechanisms that limit the scaling of bipolar transistors for digital circuits also limit the scaling of bipolar transistors for RF and analog circuits. Thus base widening limits the maximum collector current density, and collector–emitter breakdown limits how heavily the collector can

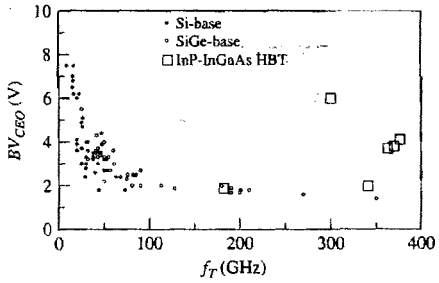


Figure 8.19. BV_{CEO} versus f_T for some n–p–n transistors reported in the literature.

be doped to minimize base widening and hence limits the maximum obtainable cutoff frequency.

Figure 8.19 is a plot of BV_{CEO} versus f_T for some n–p–n transistors reported in the literature. Data for Si-base and SiGe-base transistors, as well as some InP-based HBTs, are included. It shows that there is definitely a tradeoff between BV_{CEO} and f_T. As expected, for the same BV_{CEO} value, the SiGe-base transistors can reach a higher f_T. However, it is also apparent that if a minimum BV_{CEO} value of 3.0 V is required, then the maximum f_T obtainable is only about 60 GHz even with SiGe-base technology. On the other hand, if a BV_{CEO} value of only 2.0 V is acceptable, then SiGe-base transistors with peak f_T values about 150 GHz should be realizable.

8.6 Comparing a SiGe-Base Bipolar Transistor with a GaAs HBT

Driven primarily by wireless applications, there has been very rapid progress in the development of SiGe-base bipolar technology in recent years. In terms of frequency response, SiGe-base bipolar transistors can be comparable to compound-semiconductor HBTs. The competition between SiGe-base bipolar technology and compound-semiconductor HBT technology has led to many reports comparing their performance characteristics. In most cases, a report compares some brand new laboratory results of one transistor with some published results of some other transistors, with little attempt to normalize the data in the comparison. As a result, instead of clarifying the merits and limitations of the transistors being compared, the comparison often leaves the reader more confused.

In this section, we examine the fundamental differences between a SiGe-base transistor and a compound-semiconductor HBT. We do this by first normalizing the two transistors to some comparably advanced structure and the same design rule. This is important because at a given design rule, a device of more advanced structure has better performance than a device of less advanced structure. And, for a given device structure, a design using smaller rules results in better performance than a design using larger rules. Furthermore, we make the comparison at the same collector current densities. This is

Table 8.3 Comparison of relevant material properties

Properties	For SiGe-base bipolar	For GaAs HBTs
Dielectric constant	11.9	13.1
Energy gap (eV)	1.12	1.424
Typical average emitter doping conc. (cm^{-3})	2×10^{20}	5×10^{17}
Typical average base doping conc. (cm^{-3})	2×10^{18}	4×10^{19}
Typical base layer resistivity (Ω-cm)	3×10^{-2}	3×10^{-3}
Electron drift mobility (cm^2/V-s)	1500	8500
(at typical base doping conc.)	(~200 at $2 \times 10^{18} cm^{-3}$)	(~700 at $4 \times 10^{19} cm^{-3}$)
Electron drift veft velocity (cm/s)		
at $\mathscr{E} = 10^4$ V/cm	0.72×10^7	1.35×10^7
at $\mathscr{E} = 4 \times 10^4$ V/cm	0.97×10^7	0.89×10^7

important because, as explained in Section 8.3, the design and optimization of a bipolar transistor are closely coupled to its intended collector current density. Once the two transistors have been thus normalized, a fair comparison of their fundamental differences can be made by examining their device parameters (Ning, 1995). Recently, GaAs HBTs comparable in structure to the SiGe-base bipolar transistor shown in Fig. 7.29 have been reported (Oka *et al.*, 1997). Here we discuss the comparison results for such advanced SiGe-base bipolar transistors and GaAs HBTs (Ning, 2001).

- *Comparison of basic material properties.* At low electric fields, the electron mobility in GaAs is about 10 times that in Si. As a result, the low-field velocity of electrons in GaAs is also about 10 times that in Si. However, at electric fields greater than about 10^4 V/cm, which are typical in the base–collector junction depletion region of a bipolar transistor, the electron velocities in GaAs and in Si are comparable (see e.g., Sze, 1981). Some properties of Si and GaAs relevant to typical SiGe-base transistors and GaAs HBTs are compared in Table 8.3. To first order, the Si and GaAs have about the same dielectric constant, which means comparison of the emitter–base junction capacitance per unit area, C_{dBE}, and base–collector junction capacitance per unit area, C_{dBC}, of the two transistors can be made simply by comparing the doping concentrations of the corresponding diodes.
- *Comparison of the emitter–base diodes.* The heterojunction nature of a SiGe-base bipolar transistor has been discussed in Section 7.4.9, where it is argued that a SiGe-base transistor does not behave like a wide-gap-emitter transistor even when the Ge concentration is more-or-less spatially constant. For the emitter–base diode, a SiGe-base transistor has the same C_{dBE} as a Si-base transistor. C_{dBE} of a SiGe-base transistor is determined primarily by the base doping concentration, which is lower than the emitter doping concentration. On the other hand, a GaAs HBT has a wide-gap emitter which enables the emitter to be doped more lightly than the base. As a result, C_{dBE} of a GaAs HBT is determined primarily by the emitter doping concentration. Since the doping concentration of the emitter of a GaAs HBT is lower than the doping concentration of the base of a SiGe-base transistor, C_{dBE} is smaller for a GaAs HBT than for a SiGe-base transistor.

- *Comparison of the intrinsic-base regions.* For a GaAs HBT, the wide-gap emitter allows the intrinsic base to be doped heavily and yet maintain sufficient current gain. For a SiGe-base transistor, the narrow base bandgap allows the intrinsic base to be doped heavily and yet maintain sufficient current gain. In general, the wide-gap emitter in a GaAs HBT is more efficient in enhancing current gain than the narrow-gap base in a SiGe-base transistor. As a result, the average base doping concentration is usually higher in a GaAs HBT than in a SiGe-base bipolar transistor. For the same base width and emitter geometry, a higher base doping concentration leads to a lower intrinsic-base resistance, r_{bi}. For the parameters shown in Table 8.3, r_{bi} of a GaAs HBT is about 10 times smaller than that of a SiGe-base transistor. As for the base transit time t_B, a SiGe-base bipolar transistor has a smaller t_B than a Si-base transistor because of its graded base bandgap. However, it should be noted that graded base bandgap is also possible in GaAs HBTs (Hayes *et al.*, 1983). Also, a higher electron mobility means a smaller base transit time because $t_B \propto 1/\mu_{nB}$ [see Eq. (7.24)]. For the parameters in Table 8.3 and Fig. 7.6, it can be shown that a SiGe-base transistor with $\Delta E_{gmax} = 150$ meV has about the same t_B as a GaAs HBT with the same base width but without base-bandgap grading. If base-bandgap grading is employed, t_B of a GaAs HBT can definitely be significantly smaller than that of a SiGe-base transistor. If base-bandgap grading is employed, the current gain of a GaAs HBT can be improved, and hence its r_{bi} can be further reduced.
- *Comparison of the collector regions.* As discussed earlier, the electron drift velocities in the high-field collector depletion regions of the two transistors are about the same. That means the two transistors have about the same collector doping concentration when designed to operate at the same current density. This in turn implies that the two transistors have about the same C_{dBC} and about the same base–collector junction depletion-layer transit time, t_{BC}.
- *Comparison of f_T and f_{max}.* With a smaller C_{dBE}, smaller r_{bi}, and comparable or smaller t_B, a GaAs HBT should have a higher f_T and a higher f_{max} than a SiGe-base bipolar transistor if the two transistors are of comparably advanced structure, designed in the same rules, and operated at the same collector current densities.
- *Comparison of power dissipation.* The larger energy bandgap of GaAs means that the turn on voltage V_{on} of a GaAs diode is also larger than that of a Si diode. The relationship between the bias voltage and current density of a forward-biased diode is given by Eq. (2.123), which implies that for the same forward current density,

$$V_{on}(\text{GaAs}) \approx V_{on}(\text{Si}) + 0.28\text{V}. \tag{8.23}$$

  For a transistor operated at 1 mA/μm^2, V_{on}(Si) is about 0.9 V. That is, a bipolar circuit designed to run at a certain current using GaAs HBTs dissipates about 30% more power than the same circuit designed to run at the same current using Si-base or SiGe-base bipolar transistors.
- *Comparison of scaling limits.* The scaling of a Si-base or SiGe-base bipolar transistor is limited by its base–collector junction avalanche effect, which causes BV_{CEO} to be less than 2 V when f_T is larger than 150 GHz (see Fig. 8.19). Compared to a SiGe-base transistor, the larger energy gap of GaAs means that the collector of a GaAs HBT can

be doped more heavily and still maintain sufficiently large BV_{CEO}. That is, a GaAs HBT can be designed to operate at much higher current densities without suffering from excessive base-widening effect than a Si-base or SiGe-base transistor. ***A GaAs HBT is more scaleable than a SiGe-base bipolar transistor because the energy gap of GaAs is larger than that of Si.*** Compound-semiconductor HBTs having f_T or f_{max} greater than 300 GHz and BV_{CEO} larger than 3.5 V have been demonstrated (see e.g., Hafez *et al.*, 2003; Yu *et al.*, 2003; and the f_T data in Fig. 8.19).

Exercises

8.1 The small-signal equivalent circuit for a bipolar transistor, including its internal series resistances, is shown in Fig. 6.20. Ignore the collector-substrate capacitance $C_{dCS,tot}$, which is quite small in modern bipolar transistors. Show that the cutoff frequency is given by Eq. (8.9), i.e.,

$$\frac{1}{2\pi f_T} = \tau_F + \frac{kT}{qI_C}\left(C_{dBE,tot} + C_{dBC,tot}\right) + C_{dBC,tot}(r_e + r_c).$$

State the assumptions used in the derivation.

8.2 Consider an n–p–n bipolar transistor, with an emitter area of $A_E = 1 \times 2\ \mu m^2$ and a base–collector junction area of $A_{BC} = 10\ \mu m^2$, a deep emitter with $N_E = 10^{20}\ cm^{-3}$, a box-like intrinsic base with $N_B = 1 \times 10^{18}\ cm^{-3}$ and $W_B = 100$ nm, and a uniformly doped collector of $N_C = 5 \times 10^{16}\ cm^{-3}$. Assume one-sided junction approximation for all the junctions, and that the transistor is biased with $V_{BE} = 0.8$ V (for purposes of t_{BE} and $C_{dBE,tot}$ calculations) and $V_{CB} = 2$ V (for purposes of t_{BC} and $C_{dBC,tot}$ calculations).

(a) Estimate the low-current values of the delay or transit times τ_E, τ_B, t_{BE}, t_{BC}, and τ_F, assuming $\beta_0 = 100$.

(b) Estimate $C_{dBE,tot}$ and $C_{dBC,tot}$.

(c) The maximum cutoff frequency can be estimated if we assume the transistor is operated at its maximum current density without significant base widening, i.e., at $J_C = (0.3qv_{sat}N_C)$. (Note that the maximum J_C increases with N_C.) Estimate the maximum obtainable cutoff frequency from

$$\frac{1}{2\pi f_T} = \tau_F + \frac{kT}{qI_C}\left(C_{dBE,tot} + C_{dBC,tot}\right).$$

(d) The effect of the emitter and collector series resistances on the cutoff frequency can be estimated from

$$\frac{1}{2\pi f_T} = \tau_F + \frac{kT}{qI_C}\left(C_{dBE,tot} + C_{dBC,tot}\right) + C_{dBC,tot}(r_e + r_c).$$

Assume $(r_e + r_c) = 50\ \Omega$. Estimate the maximum cutoff frequency.

8.3 This exercise is designed to illustrate the advantage of the pedestal-collector design. Consider the n–p–n transistor of Exercise 8.2. Let N_C (int) $= 5 \times 10^{16}\ cm^{-3}$ be the collector doping concentration directly underneath the emitter and intrinsic base, and N_C (ext) $= 5 \times 10^{15}\ cm^{-3}$ be the doping concentration of the extrinsic part of the collector. $C_{dBC,tot}$ of the pedestal-collector design is smaller than that of the non-pedestal-collector design. Repeat Exercise 8.2 for this pedestal-collector transistor.

8.4 This exercise is designed to illustrate the sensitivity of the maximum cutoff frequency to the pedestal-collector doping concentration. Consider the pedestal-collector transistor of Exercise 8.3. The maximum collector current density without significant base widening can be increased by increasing N_C (int). Repeat Exercise 8.3 for the case of N_C (int) $= 1 \times 10^{17}\ cm^{-3}$.

8.5 The incorporation of Ge into the intrinsic base will reduce the base transit time and the emitter delay time. For a linearly graded Ge profile, the transit time is reduced by a factor [see Eq. (7.41)]

$$\frac{t_B(\text{SiGe})}{t_B(\text{Si})} = \frac{2kT}{\Delta E_{g\max}}\left\{1 - \frac{kT}{\Delta E_{g\max}}\left[1 - \exp\left(-\Delta E_{g\max}/kT\right)\right]\right\},$$

and the emitter delay time is reduced by a factor [see Eq. (8.16)]

$$\frac{\tau_E(\text{SiGe})}{\tau_E(\text{Si})} = \frac{\beta_0(\text{Si})}{\beta_0(\text{SiGe})},$$

where $\Delta E_{g,\max}$ is the total bandgap narrowing due to Ge. The current gain ratio is given by Eq. (7.33), namely

$$\frac{\beta_0(\text{SiGe})}{\beta_0(\text{Si})} = \frac{\Delta E_{g,\max}/kT}{1 - \exp\left(-\Delta E_{g,\max}/kT\right)}.$$

Repeat Exercise 8.4 for a SiGe-base transistor with $\Delta E_{g,\max} = 100$ meV. (Assume the same V_{BE} of 0.8 V.) (Comparison of the results from Exercises 8.4 and 8.5 illustrates the effect of SiGe-base technology on f_T.)

8.6 The depletion-layer transit time is given by Eq. (8.17). Here we want to show that it can also be derived by considering the average transit time of the stored charge after the charging current is switched off. Consider a depletion layer of width W, located between $x = 0$ and $x = W$. Assume there is a constant current density J flowing through this depletion layer, and all the charges contributing to this current flow are traveling at the same velocity v, so that $J = qnv$, where qn is the mobile charge density. We further assume that qn is small enough so that the presence of the current flow does not affect the depletion-layer thickness. Imagine at time $t = 0$ the current is suddenly switched off, and the mobile charge carriers stored in the depletion layer continue to drain out at the same velocity v. The time needed to drain out the last mobile charge carrier from the depletion layer is W/v. Show that the average transit time t_{avg}, i.e., the transit time averaged over all the stored charge carriers, is only $W/2v$. [*Hint*: The transit time for a charge carrier located at x is $(W - x)/v$.]

8.7 This exercise is designed to evaluate the tradeoff between performance and base-collector junction avalanche in an n–p–n transistor. In Exercise 8.4, as N_C (int) is increased from $5 \times 10^{16}\ cm^{-3}$ to $1 \times 10^{17}\ cm^{-3}$, the maximum electric field $\mathscr{E}_m$ in the

base–collector junction depletion layer is increased, which could lead to significantly increased avalanche multiplication in the junction. One way to minimize this problem is to sandwich an i-layer between the base and the collector doped regions.

(a) Assume the base–collector junction is reverse-biased at 2 V, calculate the i-layer thickness d needed so that the maximum base-collector junction fields for the following two designs are the same (i) N_C (int) = 1×10^{17} cm^{-3} with an i-layer, and (ii) N_C (int) = 5×10^{16} cm^{-3} without an i-layer.

(b) Assume the emitter and base designs are the same as the transistor in Exercise 8.2. Calculate the delay or transit time τ_E, τ_B, t_{BE}, t_{BC}, and τ_F, assuming $\beta_0 = 100$, for the two designs in (a).

(c) Calculate $C_{dBE,tot}$ and $C_{dBC,tot}$ for the two pedestal-collector designs in (a).

(d) Estimate the maximum cutoff frequency for the two designs from

$$\frac{1}{2\pi f_T} = \tau_F + \frac{kT}{qI_C}\left(C_{dBE,tot} + C_{dBC,tot}\right),$$

assuming that in each case the transistor is operated at its maximum collector current density of $J_C = [0.3qv_{sat}N_C(\text{int})]$ to avoid significant base widening. [*Hint*: The depletion-layer width, maximum electric field, and capacitance for a p–i–n diode are derived in Section 2.2.2.4. Use one-sided junction approximation for both the emitter–base and the base–collector diodes.]

8.8 This exercise is design to show that for the same peak base doping concentration and base width, a boxlike doping profile gives a larger Early voltage than a graded base doping profile. The Early voltage V_A is given approximately by Eq. (6.72), namely

$$V_A \approx \frac{Q_{pB}}{C_{dBC}},$$

where Q_{pB} is the majority-carrier hole charge per unit area in the base [see Eq. (6.66)] and C_{dBC} is the base–collector junction depletion-layer capacitance per unit area. Consider the following two intrinsic-base profiles (see Fig. 8.20):

(i) a boxlike base of doping concentration N_B and width W_B, and

(ii) a steplike base doping profile (as an approximation for a graded doping profile) such that the doping concentration between $x = 0$ (emitter–base junction) and $x = W_B - d$ is N_B, and the doping concentration between $x = W_B - d$ and $x = W_B$ (base–collector junction) is zero.

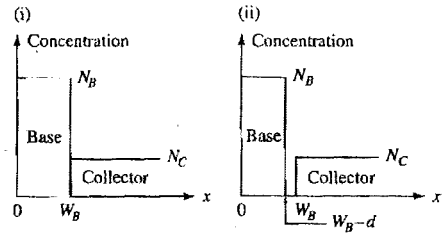


Figure 8.20. Base and collector doping profiles for Exercise 8.8.

[Thus, profile (i) is like profile (ii) with $d = 0$.] If we assume the collector doping concentration is N_C for both transistors, then it is clear that Q_{pB} is larger for profile (i) than for profile (ii), i.e., $Q_{pB}(d=0) > Q_{pB}(d)$, and C_{dBC} is also larger for profile (i) than for profile (ii), i.e., $C_{dBC}(d=0) > C_{dBC}(d)$. We want to show that $V_A(d) < V_A(d=0)$.

(a) For simplicity, we assume W_B to be large enough so that the quasineutral base width is also W_B. Show that

$$\frac{V_A(d)}{V_A(d=0)} = \left(1 - \frac{d}{W_B}\right)\sqrt{1 + \frac{d^2}{W_{dBC}^2(d=0)}},$$

where $W_{dBC}(d=0)$ is the base–collector junction depletion-layer width for profile (i). [*Hint*: Use Eq. (2.97) for the capacitance ratio $C_{dBC}(d) / C_{dBC}(d=0)$.]

(b) The results in (a) show that the ratio $V_A(d) / V_A(d=0)$ is less than 1 unless the ratio $d / W_{dBC}(d=0)$ is quite large. Estimate the largest value of $d / W_{dBC}(d=0)$ for which the ratio $V_A(d) / V_A(d=0)$ is still less than 1, for $d / W_B = 0.1$, and for $d / W_B = 0.2$.

(c) Let us put in some real numbers. Estimate the ratio $V_A(d) / V_A(d=0)$ for $d = 10$ nm and for $d = 20$ nm, assuming $N_B = 1 \times 10^{18}$ cm^{-3}, $N_C = 5 \times 10^{16}$ cm^{-3}, $W_B = 100$ nm. Assume the one-sided junction approximation and $V_{BC} = 0$ in calculating W_{dBC}.

8.9 Referring to Appendix 18 (on f_T and f_{max} of bipolar transistors), show that $r_e \text{Re}(Y'_{11}Y'_{22}/Y'_{21}) << 1$ for a typical bipolar transistor.

9 Memory Devices

The previous chapters have considered the operation of CMOS and bipolar devices mainly in the context of logic circuits. This chapter addresses another basic functional block in modern VLSI chips – memory. A predominant majority of the VLSI devices produced today are in various forms of *random-access memory* (RAM).

Viewed from the operation standpoint, a RAM functional unit is usually organized into an *array* of memory cells (or *bits*) together with its supporting circuits for selecting, writing, and reading the memory cells. In an array, the bits on the same row are selected by a *word* signal. A schematic block diagram of a RAM unit is shown in Fig. 9.1. The array consists of W words with B bits each, for a total memory capacity of $W \times B$ bits. A random bit in the array can be accessed through signals applied to its *wordline* and *bitline*.

Depending on the retention of information in the cells of a memory array, random-access memories can be classified into three categories: *static random-access memory* (SRAM), *dynamic random-access memory* (DRAM), and *nonvolatile random-access memory* (NVRAM). NVRAM is often referred to as nonvolatile memory for short. SRAMs have fast access times. They retain data as long as they are connected to the power supply. Practically every VLSI chip contains a certain amount of SRAM which is usually built using basically the same devices as in the logic circuits. DRAMs have relatively slow access times. They require periodic refresh in order to prevent loss of data. On a per-bit basis, DRAMs have a much lower cost than SRAMs because a DRAM cell is typically only about one tenth the size of an SRAM cell. For systems that require much more SRAM than can be contained on the logic chip, stand-alone SRAM chips are often used to meet the need. However, in order to reduce system cost and size, designers often use stand-alone DRAM chips instead of stand-alone SRAM chips. In that case, some form of memory-hierarchy architecture is usually employed to minimize the impact of the relatively slow DRAM on the system performance. Both SRAMs and DRAMs are volatile in that data are lost once the power supply to the chip is disconnected.

For systems that must retain all or some of their data at all times, nonvolatile memories are often used in place of or in addition to DRAMs. Nonvolatile memories can be classified into three categories: read-only or non-programmable, programmable once, and erasable and programmable. By far the most versatile nonvolatile memory technology is the erasable and programmable type, particularly the *electrically erasable and programmable read-only memory* (EEPROM).

Semiconductor memory development reached a critical high point when bipolar SRAMs were used as the main memory in the IBM System 370 Model 145 mainframe

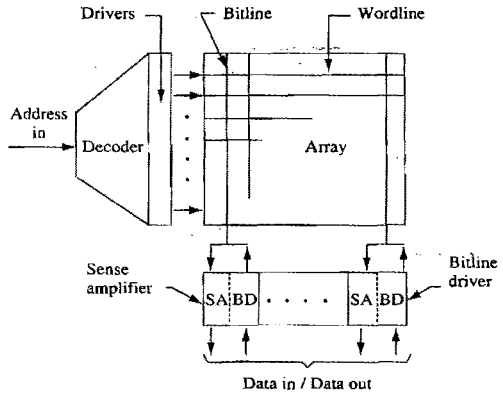


Figure 9.1. Block diagram of a random-access memory functional unit. In this schematic, a bitline represents either a single line or a pair of lines. (After Terman, 1971.)

computer first shipped in 1971 (Pugh *et al.*, 1981). However, as explained in Section 8.4.2, the standby power of bipolar circuits is high, making bipolar devices unsuitable for memory applications that require a very large number of bits on a single chip. CMOS devices, with their low standby power characteristics, are uniquely suitable for building large SRAM arrays. As a result, CMOS is now used to build SRAM functions whenever CMOS devices are available on the chip, e.g., in CMOS and BiCMOS technologies. Relatively small arrays of bipolar SRAMs are still used in applications built with bipolar-only technology.

The papers by Terman and Sah (Terman, 1971; Sah, 1988) give a detailed account of the early efforts on exploring various kinds of dynamic semiconductor memory cells. The one-transistor, one-capacitor memory cell (Dennard, 1968) is by far the densest dynamic memory cell. It has been subsequently adopted universally as the standard DRAM cell. In the case of NVRAM and EEPROM, the early development of many of the device concepts has been summarized in the literature (Sah, 1988; Hu, 1991). They remain areas of very active research. In this chapter, we discuss the basic operational principles and device design and scaling issues of the CMOS SRAM cell, the one-transistor, one-capacitor DRAM cell, and several commonly used EEPROM devices. Only the most commonly used bipolar SRAM cell will be discussed.

9.1 Static Random-Access Memory

In principle, any device or arrangement of devices that can be programmed into two distinct, electrically stable states can be used as a storage element of an SRAM cell. Depending on the storage element, one or more *access transistors* are connected to the storage element to make an SRAM cell. In this section, we first discuss the basic operation of CMOS SRAM cells and the design and scaling issues of the

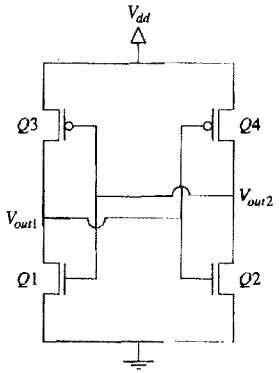


Figure 9.2. Circuit schematic of two cross-coupled CMOS inverters. The output of inverter 1 ($Q1$ and $Q3$) is the input to inverter 2 ($Q2$ and $Q4$), i.e., $V_{in2} = V_{out1}$, and vice versa.

constituent devices. After that, the basic operation of the commonly used bipolar SRAM cell is given.

9.1.1 CMOS SRAM Cell

In CMOS VLSI designs, the most commonly used SRAM storage element is the *bistable latch* consisting of two cross-coupled CMOS inverters shown in Fig. 9.2. It can be built using a standard CMOS logic fabrication process. Inverter 1 consists of nMOSFET $Q1$ and pMOSFET $Q3$ while inverter 2 consists of nMOSFET $Q2$ and pMOSFET $Q4$. The two stable states can be readily recognized by plotting the transfer curves (Section 5.1.1.1) of the two inverters back to back, as illustrated in Fig. 9.3, often referred to as the "butterfly" plot of a pair of cross-coupled inverters.

In Fig. 9.2, one of the inverters has its input at high and output at low, while the other inverter has its input at low and output at high. The first inverter, with its output at low, keeps the second inverter in the state described above, and vice versa. Thus, a CMOS SRAM storage element has two stable states: one at the intersection A of the two inverter transfer curves in Fig. 9.3 with $V_{out1} = V_{in2} = V_{dd}$, and the other at the intersection B with $V_{out2} = V_{in1} = V_{dd}$. The two stable states can be interpreted as logical "0" and "1". Here we designate logical "1" as $V_{out1} = 0$ and $V_{out2} = V_{dd}$, i.e., point B, and logical "0" as $V_{out1} = V_{dd}$ and $V_{out2} = 0$, i.e., point A. A bistable latch will remain in one of its two stable states until it is forced by an external signal to flip to the other stable state.

The most commonly used SRAM cell is a six-transistor cell consisting of two cross-coupled CMOS inverters and two access transistors. The circuit schematic for a CMOS SRAM cell is shown in Fig. 9.4. The cross-coupled inverters are connected to two bitlines, *BLT* (bitline true) and *BLC* (bitline complement), through n-channel access transistors $Q5$ and $Q6$. The access transistors are controlled by the wordline (*WL*) voltage. In the standby mode, *WL* is kept low ($V_{WL} = 0$ V), thus turning off the access transistors and isolating the bitlines from the cross-coupled inverter pair.

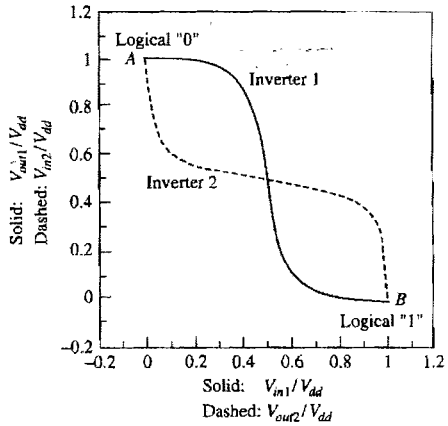


Figure 9.3. Butterfly plot for two cross-coupled CMOS inverters. The transfer curve of inverter 1 (solid) is plotted as V_{out1} vs. V_{in1}, and inverter 2 (dashed) as V_{in2} vs. V_{out2}. Identical and symmetric inverters are assumed. In a CMOS SRAM cell, nMOSFETs $Q1$, $Q2$ are usually made stronger than pMOSFETs $Q3$, $Q4$ and the high-to-low transition of the transfer curves is not symmetric. See Fig. 9.7(c).

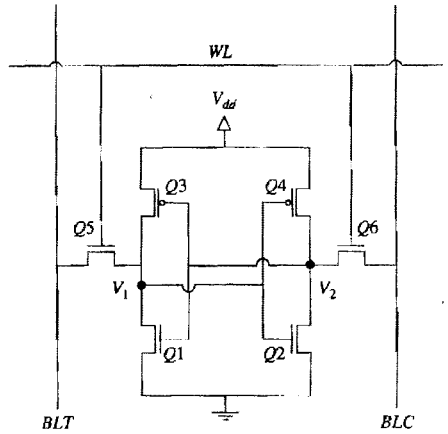


Figure 9.4. Circuit configuration of a CMOS SRAM cell. In the text, we assume the cell is storing a "1" when V_2 = high (V_1 = low) and a "0" when V_1 = high (V_2 = low).

9.1.1.1 Basic Operation of a CMOS SRAM Cell

Here we give a description of the basic read and write operations of a CMOS SRAM cell. The reader is referred to the literature for details on the circuits involved in operating an SRAM array or chip (see e.g., Itoh, 2001).

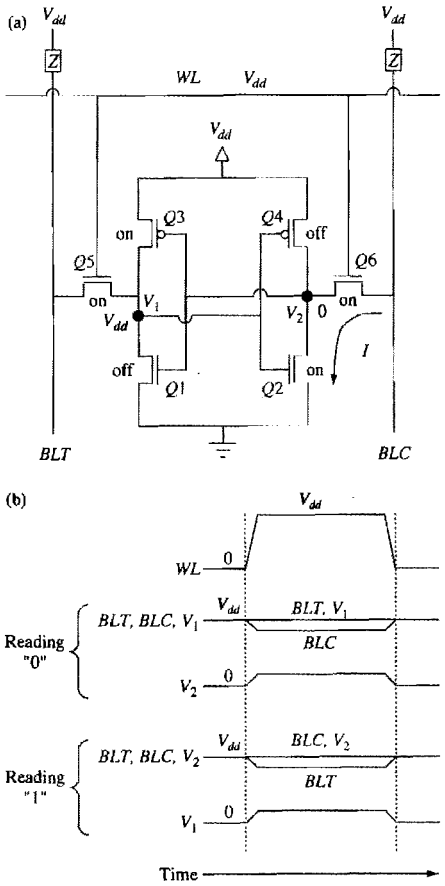


Figure 9.5. Read operation of a CMOS SRAM cell. (a) Voltage and current in reading a "0". (b) Node waveforms in reading a "0" and reading a "1".

- *Read operation*. In a read operation, the wordline is kept high ($V_{WL} = V_{dd}$), thus turning on the access transistors and allowing the logical state of the cell, represented by the values of V_1 and V_2, to be sensed via the bitlines. The signal voltages involved in the read operation and their timing are shown in Fig. 9.5. Prior to the wordline being selected (V_{WL} raised from 0 to V_{dd}), both bitlines are precharged to V_{dd} via low-impedance loads Z. Let us first consider reading a logical "0". In this case, prior to the wordline being selected, we have V_{BLT}, V_{BLC} and node V_1 all at V_{dd}, and node V_2 at 0. After the wordline is selected, $Q5$ and $Q6$ are on. Charge flows from BLC to V_2 through the conducting $Q6$, causing V_2 to rise and V_{BLC} to drop. The bitline voltage difference, $V_{BLT} - V_{BLC} > 0$, is read by the sense amplifier connected to the bitlines. In reading a logical "1", the bitline voltage difference is such that $V_{BLT} - V_{BLC} < 0$. At the end of the read operation, the wordline is turned off, thus isolating the cell from the bitlines and allowing V_1 and V_2 to return to their standby values before the read cycle. It should be noted that voltages V_1 and V_2 deviate from their standby values during the read operation. This makes the SRAM cell less stable while being read. Proper design needs be exercised to ensure that a memory bit is not flipped by the read operation itself. The read instability issue will be discussed later.

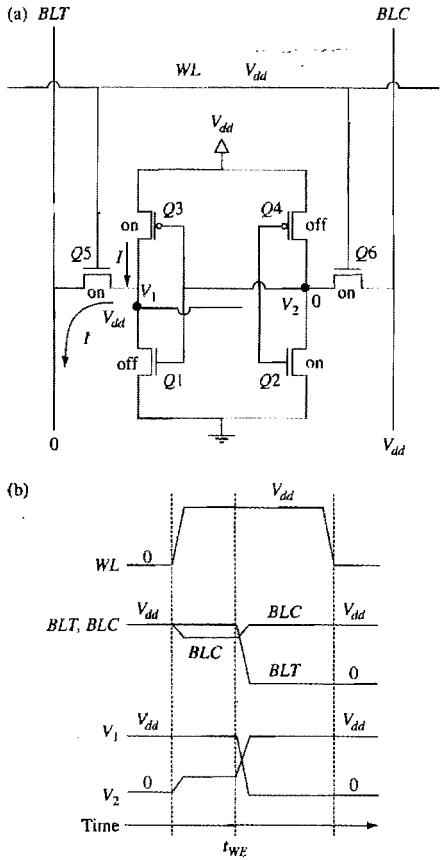


Figure 9.6. Write operation of a CMOS SRAM cell. (a) Voltage and current in writing a "1" to a cell originally storing a "0", i.e., flipping (V_1, V_2) from (V_{dd}, 0) to (0, V_{dd}). The "on", "off" labels refer to the transistor states before flipping. (b) Node waveforms in writing a "1" to a cell storing a "0".

- *Write operation*. In a write operation, appropriate write voltages are applied to the bitlines to force the cell into the intended logical state. As an example, let us consider writing a cell from logical "0" to logical "1". The signal voltages involved and their timing are illustrated in Fig. 9.6. The bitlines are precharged to V_{dd} prior to the

wordline being selected. A "write enable" signal is given at time t_{WE} when a voltage $V = V_{dd}$ is applied to *BLC* and a voltage $V = 0$ is applied to *BLT*. The voltage on *BLT* forces V_1 to 0, while the voltage on *BTC* forces V_2 to V_{dd}, thus writing a logical "1" to the cell. At the end of the write operation, the wordline is turned off, thus leaving the isolated cell in a logical "1" state.

9.1.1.2 Device Sizing for a CMOS SRAM Cell

From density consideration, it is desirable to have all the transistors in a CMOS SRAM cell as small as possible. However, for stability in read and instability in write operations, the devices must have the correct on conductance or "strength" relative to one another. As a result, not all transistors can be of minimum size.

Let us consider the current path during a read "0" operation illustrated in Fig. 9.5(a). We want the current through transistors $Q6$ and $Q2$ to cause V_{BLC} to drop but without causing V_2 to rise too much to affect the stability of the cell. That is, we want the on resistance of $Q2$ to be small compared to that of $Q6$, i.e., we want $R(Q2) < R(Q6)$, or $Q2$ to be stronger (wider) than $Q6$. By symmetry, $Q1$ needs to be stronger than $Q5$, or $R(Q1) < R(Q5)$.

The relative device strength for the write operation is considered in Fig. 9.6(a), where a "1" is written to a cell originally storing a "0". Here we want to pull V_1 readily from high to low and flip the cell. Hence we want $Q5$ to be strong compared with $Q3$, or $R(Q5) < R(Q3)$. Similarly, we want $Q6$ to be stronger than $Q4$, or $R(Q6) < R(Q4)$.

It is straightforward to make $Q3$ and $Q4$ the weakest transistors in a CMOS SRAM cell because they are pMOSFETs. For the same device channel length and threshold voltage magnitude, a pMOSFET has about half of the current per width of an nMOSFET. Thus, by using minimum-size pMOSFETs for $Q3$ and $Q4$ and minimum-size nMOSFETs for $Q5$ and $Q6$, we meet the device sizing requirement for write operation. To meet the device sizing requirement for read operation, $Q1$ and $Q2$ must be larger than minimum size. Designers typically make $Q1$ and $Q2$ about twice the width of $Q5$ and $Q6$ (see e.g., Seevinck *et al.*, 1987).

9.1.1.3 Static Noise Margin of a CMOS SRAM Cell

A memory cell should maintain its logical state when the memory array or chip is in use in a system environment. The stability of a memory cell is often characterized by its *static noise margin* (SNM), i.e., by the magnitude of noise voltage needed to cause the memory bit to flip to the other logical state. As discussed earlier, when a memory cell is in standby, it is isolated from the bitlines. However, when the cell is accessed during a read or write operation, it is coupled to the bitlines through the access transistors. We shall discuss the noise margins of a CMOS SRAM cell when it is in the standby mode, when during a read operation, and when during a write operation.

- *SNM in standby mode.* In the standby mode, the access transistors are turned off. Consider the cell in a logical "0" state, i.e., at the stable point A with $V_2 = 0$ and $V_1 = V_{dd}$. Let us assume there is a noise voltage of magnitude V_n that tends to flip the cell, i.e., the noise adds a bias of $+V_n$ to the input of inverter 1 and a bias of $-V_n$ to the input

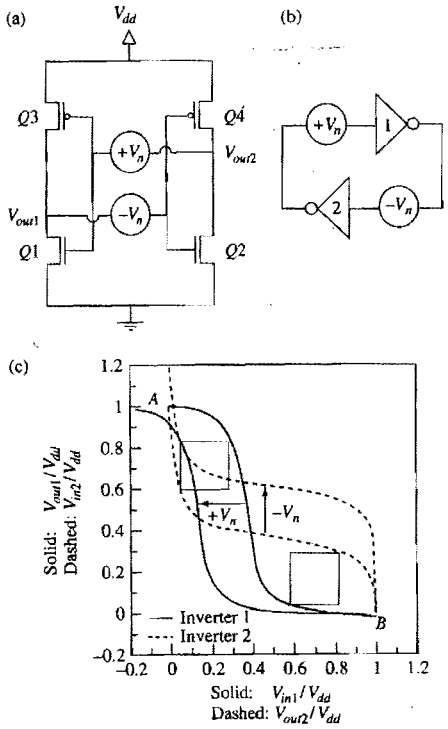


Figure 9.7. A CMOS SRAM cell having a noise voltage that tends to flip the cell from "0" to "1". (a) Schematic showing the transistor connections. (b) Schematic showing the circuit configuration. (c) Effect of the noise on the transfer curves. Note that the high-to-low transitions of the transfer curves before noise occur at $V_{in}/V_{dd} < 0.5$, as discussed in the text.

of inverter 2, as illustrated schematically in Figs 9.7(a) and 9.7(b). The corresponding transfer curves are illustrated in Fig. 9.7(c). Because read and write operations require that the bottom nMOSFETs be stronger than the top pMOSFETs, the transfer curve is not symmetric, with its high-to-low transition at $V_{in} < V_{dd}/2$. The effect of the noise voltages is to shift the transfer curve of inverter 1 horizontally to the left and the transfer curve of inverter 2 vertically upward. The SNM is the minimum noise voltage that shifts the two transfer curves until they no longer intersect at point A, i.e., the only intersection is at point B. The cell has flipped from "0" (stable point A) to "1" (stable point B). ***Graphically, the static noise margin is measured by the side of the maximum square that can be nested between the two transfer curves*** as indicated in Fig. 9.7(c) (Lohstroh *et al.*, 1983), just like that of a cascade chain of inverters discussed in Section 5.1.1.2. The static noise margin for flipping a cell from "1" to "0" can be derived in a similar manner by switching the noise voltage polarities. In that

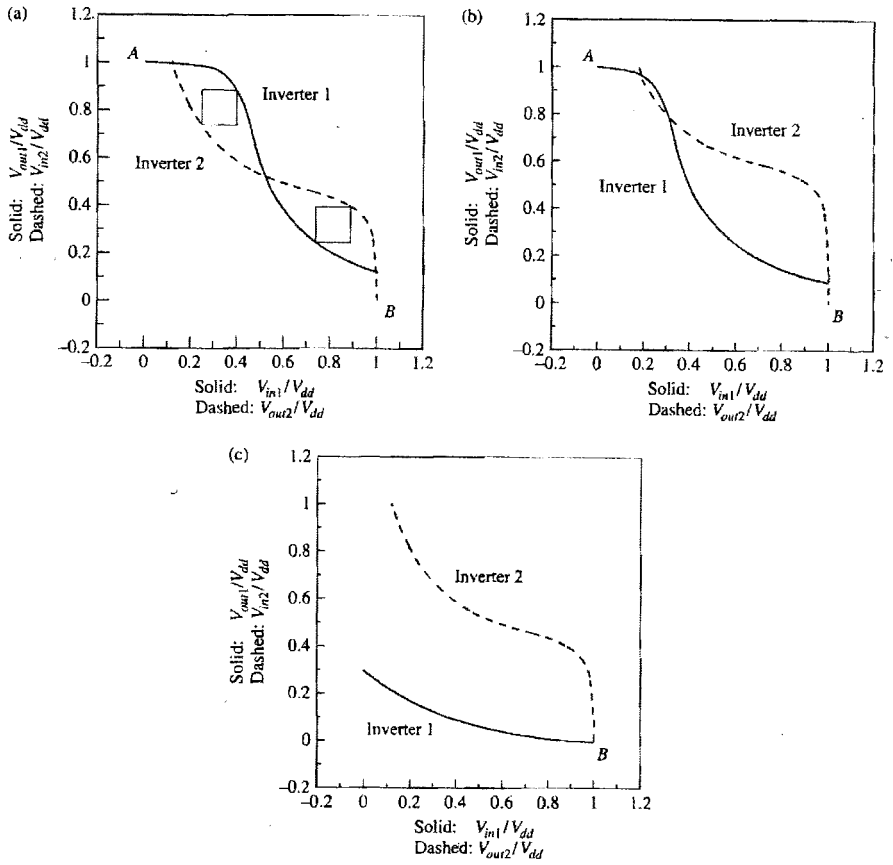


Figure 9.8. The transfer curves of a CMOS SRAM cell (a) during a read operation with identical inverters, (b) during a read operation with mismatched inverters, and (c) during a write operation.

case, the transfer curve of inverter 1 is shifted horizontally to the right while that of inverter 2 is shifted vertically downward until they only intersect at point A.

- *SNM during read access.* As shown in Fig. 9.5, when a CMOS SRAM cell is being read, the node that is low is pulled up by the bitline while the node that is high remains high. When the cell is reading a "0" (V_1 = high and V_2 = low), $V_{out1} = V_1$ remains high ($= V_{dd}$), but $V_{out2} = V_2$ is pulled up by the bitline to a value > 0. Similarly, when the cell is reading a "1" (V_1 = low and V_2 = high), $V_{out2} = V_2$ remains high ($= V_{dd}$), but $V_{out1} = V_1$ is pulled up by the bitline to a value > 0. The resulting butterfly curves during a read operation are illustrated in Fig. 9.8(a). Because of the added connection of the output node through the turned-on access transistor to the bitline at V_{dd}, the high-to-low transition of each inverter becomes less steep and the output voltage at input high

(V_{dd}) does not go all the way to zero. ***In practice, a CMOS SRAM cell is more vulnerable to noise disturbance during the read operation because of the smaller noise margin.***

For nominal bits with nearly symmetric inverters and transfer curves, the normalized SNM, V_{NM}/V_{dd}, does not change significantly with V_{dd} until V_{dd} is only a few kT/q, as discussed in Section 5.1.1.2. However, for the worst-case bits with severely mismatched devices such that one of the noise margins is barely above zero (Fig. 9.8(b)), reduction of V_{dd} even slightly could push those bits over the edge, i.e., one of their SNM goes to zero and they fail to function properly. This is further discussed in the next section.

- *SNM during write access.* In the write operation shown in Fig. 9.6, voltages (0, V_{dd}) are applied to (*BLT*, *BLC*) with the access transistors turned on to flip the SRAM cell from a "0" state to a "1" state, i.e., flipping (V_1, V_2) from (V_{dd}, 0) to (0, V_{dd}). This can be illustrated in terms of the static transfer curves plotted in Fig. 9.8(c). Since $Q5$ is stronger than $Q3$, i.e., $R(Q5) < R(Q3)$, the connection to $BLT = 0$ causes $V_{out1}(V_{in1} = 0)$ to fall far enough below V_{dd} that the noise margin for state "0" (intersection A) completely vanishes and the only allowed state for the cross-coupled latch is state "1" (intersection B).

9.1.1.4 Scaling Issues of CMOS SRAM Cells

The scaling properties of CMOS devices and technology have been discussed in Sections 4.1 and 4.2, with the device parameter trends in scaling shown in Fig. 4.7. In general, the overdrive ratio V_{dd}/V_t has decreased as both V_{dd} and V_t are scaled down with the gate length. This means that the I–V characteristics and hence the transfer curves become more sensitive to V_t variations that do not scale with V_{dd}. Since the SNM of a CMOS SRAM cell depends critically on device matching and since a typical SRAM array has a very large number of cells with a wide statistical distribution of device parameters, these factors pose specific scaling issues for CMOS SRAMs as discussed below.

- *Threshold voltage variation due to short-channel effect.* Because the built-in potential ψ_{bi} is nearly a constant for silicon, the short-channel threshold-voltage rolloff, Eq. (3.67), does not scale with V_{dd}. Also, any process variation of t_{ox} and doping concentration can add to the V_t variation. They can cause device mismatch and reduction of the SNM of SRAM cells. Gate length mismatch due to lithography can be minimized by laying out the transistors within a cell in a symmetrical manner. Other process variations usually track well within the close proximity of a cell.
- *Threshold voltage variation due to statistical dopant fluctuation.* Threshold voltage variation caused by statistical fluctuation of the number of dopant atoms in a small MOSFET has been covered in Section 4.2.5. For CMOS SRAM devices with a minimum width in the 100-nm range, the standard deviation $\sigma_{V_{on}}$ of V_t fluctuation is of the order of 30 mV (Wong and Taur, 1993; Frank *et al.*, 1999; Bhavnagarwala *et al.*, 2001). Threshold voltage variation due to dopant number fluctuation is completely random and hence cannot be minimized by placing transistors close to one another or by layout. In fact, V_t fluctuation is usually determined experimentally by using matched pairs of adjacent transistors and measuring the V_t difference between the two

transistors in a pair (Mizuno *et al.*, 1994; Tuinhout *et al.*, 1996, 1997). Threshold voltage fluctuation can cause one or more of the transistor pairs in a CMOS SRAM cell to become mismatched in V_t, which in the worst case can cause one of the SNM of the cell to vanish.

- *Threshold voltage variation due to high-field effects.* As discussed in Section 3.2.5, the characteristics of a CMOS device can change due to high-field effects. In general, high-field degradation results in an increase of V_t magnitude. Thus, even if a device pair is well matched as fabricated, their characteristics may become mismatched appreciably after burn-in stress or during operation. For advanced CMOS generations, threshold voltage instability in pMOSFETs due to negative-bias-temperature instability (NBTI) is of the most concern. NBTI can cause relatively large V_t shifts as well as V_t mismatch in pMOSFETs (Rauch III, 2002).

For a given CMOS SRAM array, there is a minimum supply voltage V_{min} below which at least one bit in the array ceases to function. The failed bit usually has severely mismatched devices and inverters. For example, one inverter (#1 in Fig. 9.8(b)) may have its high-to-low transition shifted to $V_{in} < V_{dd}/2$ because the nMOSFET, with its V_t at the low end of the distribution, is much stronger than the pMOSFET whose V_t (magnitude) happens to be on the high side. The other inverter (#2 in Fig. 9.8(b)) in the cross-coupled cell could be on the opposite side of the spectrum, i.e., its low-V_t (magnitude) pMOSFET is much stronger than its high-V_t nMOSFET such that the high-to-low transition is at $V_{in} > V_{dd}/2$. Reduction of V_{dd} has a more pronounced effect percentage-wise on the current drive of the high-V_t device whose overdrive ratio V_{dd}/V_t is already low, than on the current drive of the low-V_t device. In other words, reduction of V_{dd} makes the weak device even weaker thus further worsens the mismatch until it becomes so severe that one of the SNM goes to zero.

For a large array of the order of 10^9 devices with a Gaussian distribution, on average two of the devices will have V_t deviated more than $6\sigma_{V_{on}}$ from the nominal value. This is of the order of 0.2 V due to dopant number fluctuations alone. With a scaled down $V_{dd} \sim 1$ V and $V_t \sim 0.3$ V in the sub-100-nm CMOS generations, it is difficult to design the transistors in an SRAM cell to guard against such a large percentage of V_t variation (Bhavnagarwala *et al.*, 2001). Tradeoffs in device size (larger width) and V_t (may impact either standby power or read speed) are often necessary to ensure functionality of the SRAM array in the intended voltage range.

9.1.2 Other Bistable MOSFET SRAM Cells

As explained earlier, the storage element in a CMOS SRAM cell is the bistable latch consisting of two cross-coupled CMOS inverters. Other bistable latches can be employed as storage elements to make other types of SRAM cells. Figure 9.9 shows schematically three of the common types that are in use in applications built using older generations of technologies. The access transistors, not shown, are connected in the same manner as in a full CMOS SRAM cell. In Fig. 9.9(a), the depletion-mode nMOSFETs are normally on devices with a negative V_t, in contrast to the normally off enhancement-mode nMOSFETs with a positive V_t. In Fig. 9.9(c), the thin-film transistors (TFT) are

Table 9.1 Comparison of the Characteristics of SRAM Cells

	Full CMOS	Depletion Load	Resistor Load	TFT Load
Standby current	Low	High	Medium	Low
Cell stability	High	High	Low	Medium
Cell density	Low	Medium	High	High

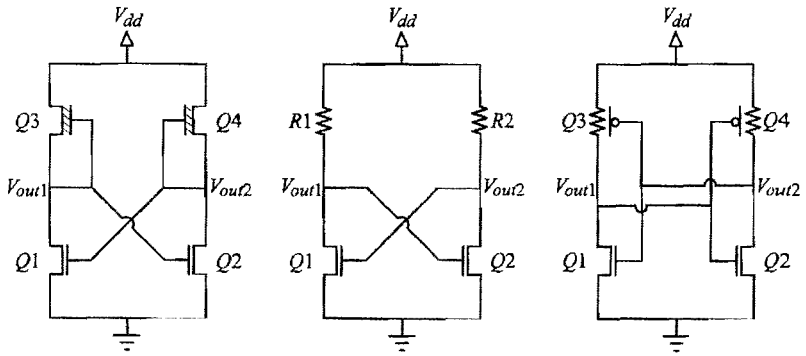


Figure 9.9. Schematics of three other bistable storage elements that can be used to form SRAM cells. (a) A depletion-load cell, where $Q3$ and $Q4$ are depletion-mode nMOSFETs. (b) A resistor-load cell, where $R1$ and $R2$ are high-resistance resistors. (c) A TFT-load cell, where $Q3$ and $Q4$ are p-channel thin-film transistors.

low-mobility MOSFETs made in polysilicon film. A TFT-load cell offers density advantage since the p-channel TFTs can be stacked on top of the regular nMOSFETs in the cell. The noise margin for these storage elements can be analyzed in a similar manner, and the standby current is determined by the off current of the load devices ($Q3$ and $Q4$ in the depletion-load cell and the TFT-load cell, and $R1$ and $R2$ in the resistor-load cell). Table 9.1 gives a comparison of these SRAM cells with the full CMOS SRAM cell (Itoh, 2001). The only cell that has noise margin comparable to the full CMOS SRAM cell is the depletion-load cell. However, its standby current is much too large for modern VLSI applications where the number of SRAM cells on a chip can be larger than 20 MB. The resistor-load cell inherently has inferior noise margin compared to the full CMOS cell (Scevinck *et al.*, 1987). From a process complexity point of view, the full CMOS SRAM cell is free in that it is fabricated using a CMOS logic process without modification. With best noise margin and standby power characteristics, the full CMOS cell is the SRAM cell of choice in high performance scaled technologies.

9.1.3 Bipolar SRAM Cell

The storage element in a bipolar SRAM is a bistable latch shown in Fig. 9.10(a). Each bipolar inverter consists of one bipolar transistor and a load resistor. In normal operation,

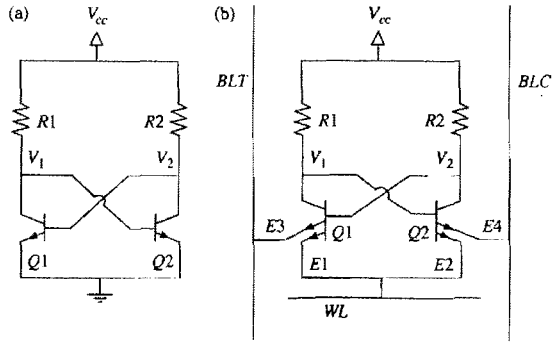


Figure 9.10. (a) A bipolar latch. (b) An emitter-coupled bipolar SRAM cell.

one of the transistors in the latch is on while the other transistor is off. Let us assume transistor $Q1$ to be on. There is a relatively large current flow through resistor $R1$. The IR drop in $R1$ means that the base voltage (which is V_1) of transistor $Q2$ is low, keeping $Q2$ in the off state. Similarly, if $Q2$ is on, the IR drop in $R2$ keeps the base voltage (which is V_2) of $Q1$ low and hence keeps $Q1$ in the off state.

To form an SRAM cell, each bipolar transistor in the latch is coupled to a bitline through an additional emitter. This emitter-coupled bipolar SRAM cell is illustrated in Fig. 9.10(b). Instead of having just one emitter, each of the bipolar transistors in the latch now has two emitters, one for forming the basic bistable latch, and one for coupling to a bitline. The emitters of $Q1$ are labeled $E1$ and $E3$, and the emitters of $Q2$ are labeled $E2$ and E4. $E1$ and $E2$ are used to form the bistable latch, while $E3$ and $E4$ couple the true bitline BLT and the complement bitline BLC, respectively, to the latch.

The operation of an emitter-coupled bipolar SRAM cell is based on the fact that ***the current flow in a multi-emitter transistor is carried primarily by the emitter that has the largest base–emitter forward bias voltage V_{BE}***. Consider the transistor $Q1$ in Fig. 9.10(b). Let us assume both emitters $E1$ and $E3$ to have the same area. Let V_{BE1} and V_{BE3} be the base–emitter forward bias voltage of $E1$ and $E3$, respectively. The collector current is carried by the two emitters. The portion of collector current in $E1$ is proportional to $\exp(qV_{BE1}/kT)$, and the portion in $E3$ is proportional to $\exp(qV_{BE3}/kT)$. At room temperature, the collector current changes by 10× for 60 mV change in V_{BE}. Therefore, if $V_{BE1} - V_{BE3} > 60$ mV, the collector current can be assumed to be carried entirely by emitter $E1$. Similarly, if $V_{BE3} - V_{BE1} > 60$ mV, the collector current can be assumed to be carried entirely by emitter $E3$.

There are other means of coupling a bipolar latch to bitlines to form an SRAM cell. However, the emitter-coupled cell is the simplest because it can be built using just bipolar transistors and resistors, the same as used for building fast bipolar logic circuits. Furthermore, bipolar SRAM is now used only in niche applications where process simplicity is more important than power dissipation and/or density. As a result, even though the emitter-coupled cell is not as low power as some other cells (Lynes and Hodges, 1970), it is the most commonly used bipolar SRAM cell. Here we discuss the behavior of a bipolar transistor in the operation of an emitter-coupled cell. The reader is referred to the literature for discussion on other bipolar memory cells (Wiedmann and Berger, 1971; Hodges, 1972; Farber and Schlig, 1972; Nokubo *et al.*, 1983).

9.1.3.1 Bipolar Transistor as an Ideal On–Off Switch

In considering bipolar circuits, designers find it convenient to think of a bipolar transistor as an ideal switch which is off when V_{BE} is less than V_{on} and on when V_{BE} is larger than V_{on}. This approach works well because the collector current of a bipolar transistor increases exponentially with V_{BE}, with the collector current changing 10 × for every 60 mV change in V_{BE} at room temperature. Once the desired on current has been established, the transistor current increases or decreases by large amounts for just a small change in V_{BE}, a property expected of an ideal switch. The exact value of V_{on} is determined by the target on current for the circuit application. For a modern silicon-base bipolar transistor, V_{on} is typically about 0.9 V, corresponding to a collector current density of about 1 mA/μm^2 (see Fig. 6.5). [In the literature, V_{on} is often taken as 0.8 V or smaller (Meyer *et al.*, 1968). The smaller value for V_{on} in these older publications is primarily caused by the larger emitter areas, and hence smaller collector current densities, of the bipolar devices used at the time. For instance, if the emitter is 10 μm^2, and the target on current is 1 mA, then the collector current density is only 0.1 mA/μm^2, and hence V_{on} can be taken as 0.84 V instead of 0.9 V.] For SiGe-base bipolar transistor, the target collector current density can be reached at a lower V_{BE}, which means V_{on} is lower by the same amount. Compared to a Si-base bipolar transistor, V_{on} is typically 50–100 mV smaller for a SiGe-base transistor, depending on the details of the Ge distribution in the base (see Section 7.4).

9.1.3.2 Operation of a Bipolar Inverter

Let us consider one of the inverters used to build the bistable latch in Fig. 9.10(a). This inverter by itself is shown in Fig. 9.11(a). It consists of an n–p–n bipolar transistor Q and a collector resistor R_C, connected between power supply V_{cc} and ground. This inverter is the basic building block for forming a *direct-coupled transistor logic* (DCTL) circuit (Meyer *et al.*, 1968). For simplicity, we assume the parasitic resistances of the transistor to be negligible. We will return to discuss the effect of parasitic resistances later. It is apparent from Fig. 9.11(a) that the output voltage V_{out} is always lower than V_{cc} by $I_C R_C$. When the input voltage V_{in} is much smaller than V_{on}, I_C is negligibly small and V_{out} approaches V_{cc}. As V_{in} is increased, I_C increases exponentially with V_{in}, and V_{out} decreases in proportion to I_C. Above a certain V_{in} value, we have V_{out} lower than V_{in}. For proper operation in a logic circuit, the bias condition must be such that the logic swing, which is equal to $V_{in} - V_{out}$ when the transistor is on, has adequate noise margin for the intended application.

It is tempting to conclude that the output voltage of an inverter switches between 0 V (ground) and V_{cc}. [The output voltage of a CMOS inverter switches between the power supply voltage and ground (see Section 5.1).] But that is incorrect because an isolated inverter by itself does not contain information about how its input node will vary when

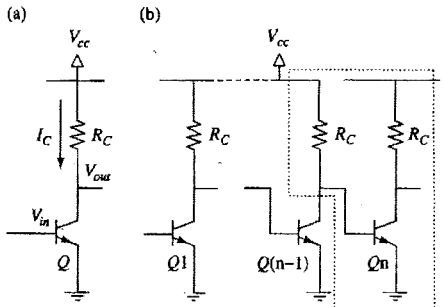


Figure 9.11. (a) An isolated bipolar inverter. (b) A chain of bipolar inverters.

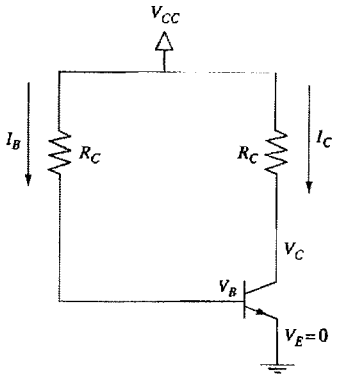


Figure 9.12. Biasing of a basic bipolar inverter in an inverter chain. There is no external resistor at the emitter terminal.

the inverter is used in a circuit. To obtain the correct output voltage swing for a bipolar inverter in a circuit, we need to include the circuit driving the inverter input. For instance, consider an inverter chain shown schematically in Fig. 9.11(b). If we assume transistor Qn to be on, then transistor Q(n − 1) is off. In this case, the current flowing through the load resistor of inverter n − 1 is not the collector current of transistor Q(n − 1), but the base current of transistor Qn. In other words, the load resistor of inverter n − 1 determines the bias current for the base node of inverter n. The quiescent voltage of an inverter in the chain is determined by the circuit configuration enclosed within the dotted line in Fig. 9.11(b). We consider the voltages of this circuit configuration next.

From Fig. 9.11(b), the biasing scheme for a bipolar inverter in an inverter chain is shown in Fig. 9.12. The emitter is held at ground potential, i.e., $V_E = 0$. The power supply voltage V_{cc} is usually larger than 2 V. I_B is the base current and I_C is the collector current. We have

$$V_{BE} \equiv V_B - V_E = V_{cc} - I_B R_C, \tag{9.1}$$

and

$$V_{CE} \equiv V_C - V_E = V_{cc} - I_C R_C. \tag{9.2}$$

For a given combination of V_{cc} and R_C, Eqs. (9.1) and (9.2) relate the transistor currents I_B and I_C to the transistor terminal voltages V_{BE} and V_{CE}. These relationships are rather complex because the currents themselves are functions of the terminal voltages. The difference between V_B and V_C gives the logic swing ΔV of the inverter circuits in the chain, i.e.,

$$\Delta V \equiv V_B - V_C = V_{BE} - V_{CE} = (I_C - I_B)R_C. \tag{9.3}$$

For proper circuit operation, $\Delta V > 0$ or $V_{BE} > V_{CE}$. That is, a bipolar transistor in an inverter circuit is operated in the saturation region. It can be shown (see Exercise 9.1) that I_C remains much larger than I_B even in deep saturation. Therefore, the logic swing is, to first order, given by the $I_C R_C$ drop. In the rest of this subsection, we examine the characteristics of a transistor operated in saturation and the dependence of the inverter terminal voltages on load resistance and power supply voltage.

- *Current–voltage characteristics in the saturation region.* The current equations derived in Chapters 6 to 8 are for a transistor operated in the nonsaturation region. Here we modify them for the saturation region so that they are applicable to a bipolar inverter circuit. With V_B larger than V_C, the collector–base diode is forward biased, causing an electron current to flow from the collector to the emitter and a hole current to flow from the base into the collector. The complete collector current in the Ebers–Moll model, i.e., Eq. (6.89), can be rearranged to give

$$\begin{aligned} I_C &= \alpha_F I_{F0}(e^{qV_{BE}/kT} - 1) - I_{R0}(e^{qV_{BC}/kT} - 1) \\ &= I_{C0F}(e^{qV_{BE}/kT} - 1) - (I_{C0R} + I_{B0R})(e^{qV_{BC}/kT} - 1) \\ &= I_{C0F}(e^{qV_{BE}/kT} - 1) - (I_{C0F} + I_{B0R})(e^{qV_{BC}/kT} - 1). \end{aligned} \tag{9.4}$$

In Eq. (9.4), $I_{C0F} = \alpha_F I_{F0}$ is the saturated collector current in the forward active mode, and I_{C0R} is the saturated collector current in the reverse active mode and I_{B0R} is the saturated base current in the reverse active mode. Notice that the base current in the forward-biased emitter–base diode does not contribute to the collector current. Also, in writing the third line of Eq. (9.4), we have used the reciprocity characteristics of the emitter and collector for electron injection into the base, which give $I_{C0R} = I_{C0F}$ (see Section 6.4.1). For a properly designed inverter, we have V_{BE} and V_{BC} much larger than kT/q, so that Eq. (9.4) can be simplified to

$$\begin{aligned} I_C &\approx I_{C0F}e^{qV_{BE}/kT} - I_{C0F}e^{qV_{BC}/kT} - I_{B0R}e^{qV_{BC}/kT} \\ &= I_{B0F}e^{qV_{BE}/kT}\left[\beta_0\left(1 - e^{-qV_{CE}/kT}\right) - \frac{I_{B0R}}{I_{B0F}}e^{-qV_{CE}/kT}\right], \end{aligned} \tag{9.5}$$

where I_{B0F} is the saturated base current in the forward mode, and $\beta_0 = I_{C0F}/I_{B0F}$ is the common-emitter current gain in the nonsaturation region. In a similar manner, we can write the base current approximately as

$$I_B \approx I_{B0F}e^{qV_{BE}/kT} + I_{B0R}e^{qV_{BC}/kT}$$
$$= I_{B0F}e^{qV_{BE}/kT}\left(1 + \frac{I_{B0R}}{I_{B0F}}e^{-qV_{CE}/kT}\right), \quad (9.6)$$

where the first term is for hole injection into the emitter and the second term is for hole injection into the collector. Notice that I_C is the difference of the forward and reverse electron currents, while I_B is the sum of the forward and reverse hole currents. Also, in general, we have $I_{B0R} > I_{B0F}$ for a modern bipolar transistor because the forward base current sees a polysilicon emitter while the reverse base current sees a regular n^+ silicon region. In addition, the junction area for base current injection is significantly larger for the reverse mode than for the forward mode, as evident from the device cross section in Fig. 7.29. For some transistors, I_{B0R} can be ten times as large as I_{B0F} (Rieh *et al.*, 2005). β_0 is typically around 100 (see Section 6.2.3). Equations (9.5) and (9.6) imply that the ratio I_C/I_B decreases as V_{BC} increases, or as V_{CE} decreases. However, it is readily shown that as long as V_{CE} is positive and larger than some fraction of kT/q, the collector current remains larger than the base current for typical bipolar transistors (see Exercise 9.1).

- *V_{CE} for a bipolar inverter.* Equations (9.1) and (9.6) can be combined to give

$$V_{cc} - V_{BE} = R_C I_{B0F}e^{qV_{BE}/kT}\left(1 + \frac{I_{B0R}}{I_{B0F}}e^{-qV_{CE}/kT}\right), \quad (9.7)$$

which can be rearranged to give

$$V_{CE} = -\frac{kT}{q}\ln\left\{\frac{[\beta_0(V_{cc} - V_{BE})/R_C I_{C0F}e^{qV_{BE}/kT}] - 1}{I_{B0R}/I_{B0F}}\right\}. \quad (9.8)$$

Similarly, Eqs. (9.2) and (9.5) can be combined to give

$$V_{cc} - V_{CE} = R_C I_{B0F}e^{qV_{BE}/kT}\left[\beta_0(1 - e^{-qV_{CE}/kT}) - \frac{I_{B0R}}{I_{B0F}}e^{-qV_{CE}/kT}\right], \quad (9.9)$$

which can be rearranged to give

$$V_{BE} = \frac{kT}{q}\ln\left\{\frac{V_{cc} - V_{CE}}{R_C[I_{C0F}(1 - e^{-qV_{CE}/kT}) - (I_{B0R}/I_{B0F})(I_{C0F}/\beta_0)e^{-qV_{CE}/kT}]}\right\}. \quad (9.10)$$

Equation (9.8) gives V_{CE} as a function of V_{BE}, while Eq. (9.10) gives V_{BE} as a function of V_{CE}. They can be used to determine V_{BE} and V_{CE} in terms of V_{cc}, R_C, and the transistor parameters I_{C0F}, I_{B0R}/I_{B0F} and β_0. The relationship between V_{BE} and V_{CE} with load resistance and V_{cc} as parameters is illustrated in Fig. 9.13.

Let us examine Fig. 9.13(b). It shows that increasing the load resistance from 2 KΩ to 20 KΩ decreases V_{CE} from about 0.022 V to about 0.015 V, while decreasing V_{BE} from about 0.90 V to about 0.84 V. Equation (9.2) suggests that the collector current is reduced by about 10 ×, consistent with the fact that V_{BE} is reduced by about 60 mV. The logic swing, given by Eq. (9.3), is also reduced by about 60 mV, but remains more than adequate. The power dissipation of the inverter circuit is reduced by about 10 ×.

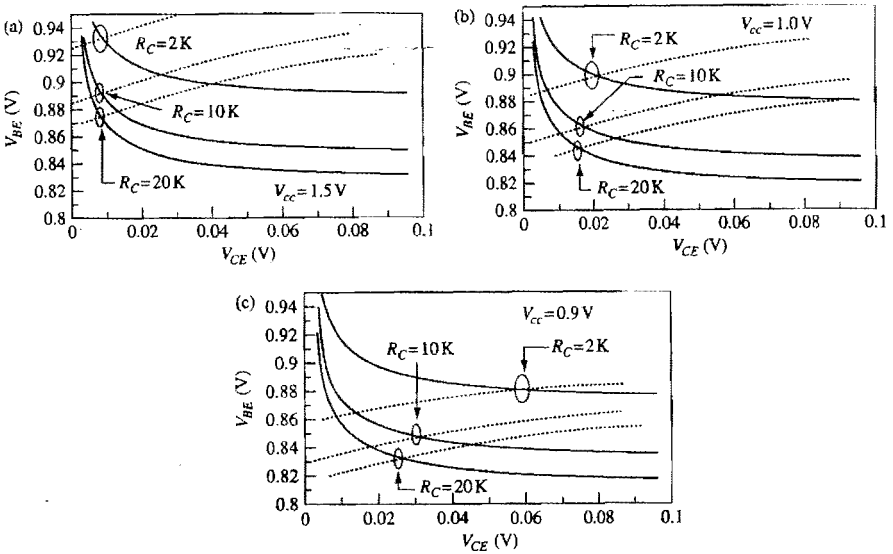


Figure 9.13. Simulated relationship between V_{CE} and V_{BE} for a bipolar in an inverter chain at room temperature. The transistor parameters used in the simulation are: $\beta_0 = 100$, $I_{B0R}/I_{B0F} = 10$, and I_{C0F} chosen such that the forward collector current I_{CF} is 1 mA at $V_{BE} = 0.9$ V. The V_{CE} curves (dashed) are given by Eq. (9.8) and the V_{BE} curves (solid) are given by Eq. (9.10). R_C is in units of Ω. (a) For $V_{cc} = 1.5$ V, (b) for $V_{cc} = 1.0$ V, and (c) for $V_{cc} = 0.9$ V. For each value of R_C, the V_{BE} and V_{CE} values for the inverter are determined by the intersection (circled) of the V_{CE} and V_{BE} curves.

In general, the choice of the load resistance in a bipolar inverter circuit is based primarily on speed or drive current requirement. There is an approximately linear relationship between power dissipation and drive current.

Comparison of the results in the three figures shows that V_{CE} increases as V_{cc} is reduced. Let us compare Figs 9.13(a) and 9.13(c) and focus on the curves for $R_C = 2$ KΩ. At $V_{cc} = 1.5$ V, we have $V_{BE} \sim 0.93$ V and $V_{CE} \sim 0.01$ V, which implies $V_{BC} \sim 0.92$ V. The transistor is in deep saturation. At $V_{cc} = 0.9$ V, we have $V_{BE} \sim 0.88$ V and $V_{CE} \sim 0.06$ V, which implies $V_{BC} \sim 0.82$ V. That is, the collector–base diode is forward biased less by 100 mV at $V_{cc} = 0.9$ V than at $V_{cc} = 1.5$ V. The difference in V_{BC} may not seem much, but it is enough to result in the amount of minority holes stored in the collector region being about 50 × less at $V_{cc} = 0.9$ V than at $V_{cc} = 1.5$ V. [At room temperature, the hole injection current from the base into the collector changes by about 10 × for every 60 mV change in V_{BC}.] The important point is that ***reducing V_{cc} across a bipolar inverter reduces the deleterious effects associated with operation in deep saturation.***

- *Effect of parasitic collector and base resistances.* The effect of parasitic resistances on the device terminal voltages is discussed in Section 6.4.3. Here we extend the discussion to a bipolar inverter. If the device has an internal collector resistance r_c that is appreciable, then

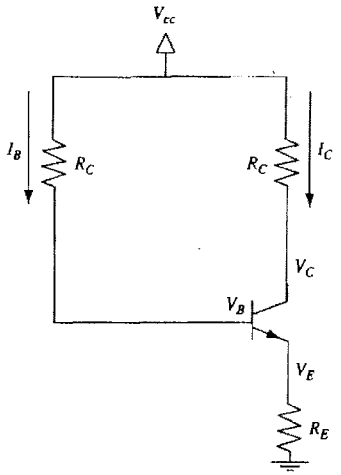


Figure 9.14. A bipolar inverter circuit having an emitter resistor.

R_C in Eq. (9.2) should be replaced by $R_C + r_c$. The presence of r_c causes the collector current to be reduced from $\sim V_{cc}/R_C$ to $\sim V_{cc}/(R_C + r_c)$. The intrinsic V'_{CE} value remains very small, but the terminal V_{CE} value is now given by $V_{CE} = V'_{CE} + r_c I_C$. There should be little change in V_{BE} because even if r_c is sufficiently large to cause a 2 × decrease in I_C, V_{BE} is reduced by only about 18 mV. We can treat parasitic base resistance in a similar manner. If the device has an appreciable internal base resistance r_b, then R_C in Eq. (9.1) should be replaced by $R_C + r_b$. The base current is reduced somewhat, according to Eq. (9.1), but there should be little change to the intrinsic V'_{BE} value or to V_{CE}.

- *Effect of parasitic emitter resistance or an external emitter resistor.* For a given power supply voltage V_{cc}, the effective voltage across a bipolar inverter can be reduced by adding an external resistor to the emitter node. This is illustrated in Fig. 9.14. An emitter resistor R_E has been added to the inverter. For a given value of V_{cc}, the IR drop in the emitter resistor has the effect of raising the emitter voltage V_E from ground to R_E $(I_C + I_B)$ above ground. As far as the inverter operation is concerned, the effective power supply is now $V_{cc} - V_E = V_{cc} - R_E (I_C + I_B)$, resulting in the transistor being operated in a less saturated region. It is standard practice in bipolar circuit design to add an external resistor to the emitter node to reduce saturation effect and hence improve circuit speed. A bipolar SRAM cell having an emitter resistor in the bistable latch is faster than one without an emitter resistor (Mayumi *et al.*, 1974).

9.1.3.3 Basic Opertion of a Bipolar SRAM Cell

For simplicity, we choose a cell without an external emitter resistor, i.e., the cell in Fig. 9.10(b). In standby mode, one of the transistors in the latch is on and one is off. Let us assume the on transistor has $V_{BE} = V_{on}$. To read a cell, we want to transfer the current in the on transistor from one emitter to another without upsetting the off transistor, and without upsetting the memory state of any of the other cells in the array. To write a

cell, we want to force its off transistor to turn on by forcing its V_{BE} to become larger than V_{on}, without upsetting the memory state of any of the other cells in the array. There are many schemes for biasing the wordline and bitlines in standby and during read and write operations. In general, V_{cc} for the latch is typically more than 2 V above the standby wordline voltage, but the wordline standby voltage varies with design. The wordline in standby can be ground, positive or even negative. Here, for simplicity, we assume the wordline standby voltage is ground, and examine how the bitlines and wordline should be varied relative to ground in order to operate the SRAM cell.

- *Standby mode and cell selection.* Referring to Fig. 9.10(b), if $Q1$ is on, then $V_{BE1} = V_{on}$. In standby mode, the wordline is at ground while the bitlines are kept at a voltage higher than V_{on} to ensure that no significant amount of current flows to either bitline. The current in the latch flows through $E1$ to the wordline. There is no current flow in $E3$ because bitline BLT is at a voltage higher than V_{on}. There is no current flow in $Q2$ which is off. A cell is selected by raising its wordline voltage, reducing the voltage across the latch. As discussed in the subsection above, both V_{BE} and V_{CE} of a transistor in an inverter is relatively insensitive to the voltage across the inverter. That is, to first order, ***V_B of the on transistor and V_B of the off transistor are shifted upward by about the same amount as the wordline.*** Now, consider the cells connected to a bitline. With the V_B of the transistors of the selected cell higher than V_B of the transistors of the nonselected cells, a voltage can be applied to the bitline to read or write the selected cell without upsetting the nonselected cells.
- *Read operation.* A cell is read by raising its wordline and lowering its bitlines to the point where the wordline becomes higher than the bitlines. Referring to Fig. 9.10(b), if $Q1$ is on, raising the wordline voltage to above the bitlines causes $V_{BE3} > V_{BE1}$, forcing the current to transfer from $E1$ to $E3$ and a current to flow in bitline BLT. All the while, $Q2$ remains off and no current flows in bitline BLC. Thus the state of the memory cell is read. At the end of the read operation, the wordline and bitlines are returned to their respective standby voltages.
- *Write operation.* Note that transistor $Q2$ can be turned on by forcing a current through one or both of its emitters $E2$ and $E4$. Assuming $Q1$ is on and we want to write the cell such that $Q2$ becomes on (and $Q1$ becomes off). The cell is selected by raising its wordline, but keeping it several kT/q below the bitline standby voltage. At this raised voltage, with bitlines at standby, the cells are not disturbed and no current flows in the bitlines. $Q2$ is then forced to turn on by lowering the voltage on bitline BLC so that V_{BE4} becomes larger than V_{BE1}, causing the flip-flop current to switch from $E1$ of $Q1$ to $E4$ of $Q2$. In this way, $Q2$ is forced to turn on while $Q1$ is forced to turn off. The write operation is complete after BLC is returned to standby.

9.2 Dynamic Random-Access Memory

A DRAM cell consists of one MOSFET and one capacitor. It is considerably smaller in silicon area than a CMOS SRAM cell which consists of six MOSFETs. The cell is in state

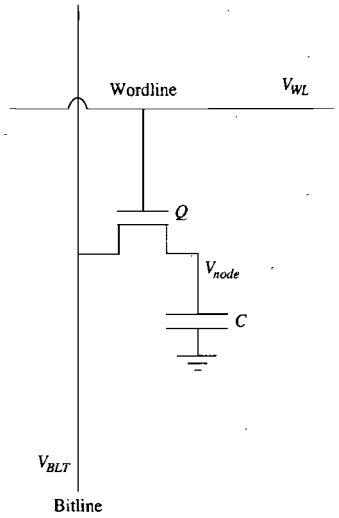


Figure 9.15. Schematic of a DRAM cell.

"0" when there is no charge and in state "1" when there is charge in the capacitor. The charge stored in the capacitor leaks away over time if left alone. Thus periodic read and refresh cycles are necessary to restore the charge state ("0" or "1") in the cell. In this section, we describe the basic operation of a DRAM cell and discuss its design and scaling issues.

9.2.1 Basic DRAM Cell and Its Operation

A DRAM cell is shown schematically in Fig. 9.15. The MOSFET Q is for accessing and transferring charge into and out of the capacitor C. The MOSFET is often referred to as the *access device* or the *transfer device*. It is usually an n-channel device. In a memory array arrangement, the gate electrode of Q is connected to the wordline while the source and drain regions are connected to the capacitor and the bitline. In Fig. 9.15, V_{node} denotes the voltage on the storage capacitor.

- *Cell structures.* The storage capacitance is determined by the minimum memory cell charge required for data retention and read operation, which will be discussed later. A typical value for the storage capacitance per cell is 30 fF. If the capacitor is a simple planar structure constructed using silicon dioxide of 10 nm thick, then the capacitor will take up an area of $A_C = Ct_{ox}/\varepsilon_{ox} \sim 9\ \mu m^2$. Such an area is excessively large for design rules of less than 1 μm. Since the generation of 4Mb DRAM, built using 1.2 μm minimum lithographic features, either a trench capacitor buried deep beneath the silicon surface (Lu *et al.*, 1985, 1986) or a stacked capacitor constructed above the transfer device (Koyanagi *et al.*, 1978) has been used to significantly reduce the area

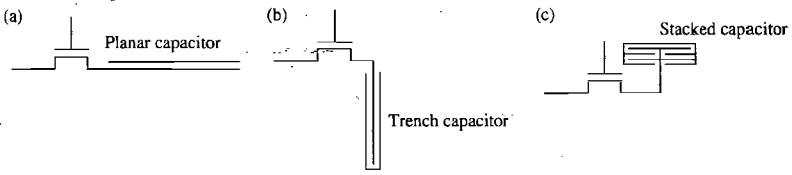


Figure 9.16. Schematics showing three DRAM cell structures: (a) planar-capacitor cell, (b) trench-capacitor cell, and (c) stacked-capacitor cell.

taken up by the storage capacitor. DRAM cells using these capacitor structures are illustrated schematically in Fig. 9.16. Also, insulators with a higher dielectric constant than SiO_2 are often used to increase the capacitance per area. Most of the effort in the development of DRAM technology is devoted to finding manufacturable means for making the cell area, particularly the storage capacitor area, small.

- *Write operation.* It is straightforward to write a "0" to a DRAM cell. Referring to Fig. 9.15, all one has to do is to turn on the access device by applying V_{dd} to the wordline and then set $V_{BLT} = 0$ to discharge any charge stored on C until $V_{node} = 0$. In writing a "1", one would like to charge the storage capacitor to $V_{node} = V_{dd}$. However, it is not sufficient to just set both V_{WL} and V_{BLT} to V_{dd}. This is because the transfer device Q turns off, i.e., goes into subthreshold, when V_{node} reaches $V_{dd} - V_t$. For V_{node} above $V_{dd} - V_t$, the gate-to-source voltage of Q becomes $V_{gs} = V_{WL} - V_{node} < V_t$, and the charging rate drops precipitously. To ensure that V_{node} reaches V_{dd} during a write operation, the wordline must be *boosted* to $V_{WL} > V_{dd} + V_t$ so that Q stays on through $V_{node} = V_{BLT} = V_{dd}$. After the storage node is fully charged or discharged, the wordline is brought back to its standby voltage, thus turning Q off and isolating the storage capacitor from the bitline. (In most designs, the standby voltage for the wordline is zero. However, in some designs, the standby voltage for the wordline can be negative. See discussion in Section 9.2.2 below.)
- *Read operation.* The read operation of a DRAM cell is illustrated in Fig. 9.17. Figure 9.17(a) shows the bitline (*BLT*) and bitline complementary (*BLC*) connections to a cross-coupled CMOS sense amplifier that not only senses the stored bit in the cell but also writes the same bit back to the cell at the end of the read cycle. The *BLC* acts to provide a reference voltage to make up the differential signal. It can be the bitline of another array of cells not on the same wordline as the cell to be read (see folded bitline in Fig. 9.18). The waveforms of the voltages involved in a read operation are shown in Fig. 9.17(b). First, both *BLT* and *BLC* are precharged to $V_{dd}/2$. Then the access transistor Q is turned on. The same boost of the wordline to $V_{WL} > V_{dd} + V_t$ is applied to ensure a full V_{dd} write-back later. If a "0" bit is stored in the cell, $V_{node} = 0$, the bitline voltage V_{BLT} is discharged to below $V_{dd}/2$ and a sense signal $V_s = V_{BLT} - V_{BLC} < 0$ is developed. If a "1" bit is stored in the cell, $V_{node} = V_{dd}$, the bitline voltage V_{BLT} is charged to above $V_{dd}/2$ and a sense signal $V_s = V_{BLT} - V_{BLC} > 0$ is developed. These are shown in Fig. 9.17(b). Before turn-on of the wordline, the sense amplifier is biased in a neutral state with both V_{SP} and V_{SN} at $V_{dd}/2$. After a differential signal is developed between *BLT* and *BLC*, the sense amplifier is activated by setting V_{SN} to 0 and V_{SP} to V_{dd}.

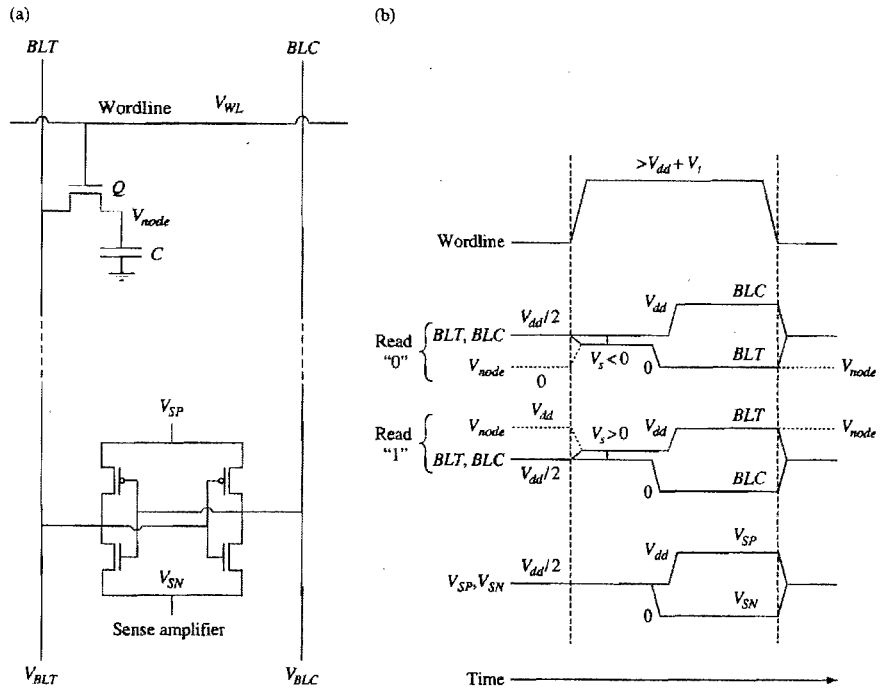


Figure 9.17. (a) Schematic connection of a DRAM cell to sense amplifier for read and write back. (b) Voltage waveforms for read and write back operation. After V_s is developed, V_{SN} is set to 0 first, which causes the lower of *BLT, BLC* to fall to 0. The higher of *BLT, BLC* is then pulled up from $V_{dd}/2$ to V_{dd} by setting V_{SP} to V_{dd}.

This turns the sense amplifier into a cross-coupled latch that must settle into one of the two stable states, depending on the polarity of the signal V_s. If $V_s < 0$, i.e., if $V_{BLT} < V_{BLC}$, then the latch settles into the state of $V_{BLT} = 0$ and $V_{BLC} = V_{dd}$. Conversely, if $V_s > 0$, the latch ends up in the other stable state, namely, $V_{BLT} = V_{dd}$ and $V_{BLC} = 0$. This also restores the cell voltage V_{node} back to either 0 or V_{dd}, the same as its starting value. After the read cycle, the wordline is turned off and all *BLT, BLC*, V_{SP}, V_{SN} are returned to their neutral voltage of $V_{dd}/2$. The cell, now isolated, is back to the originally stored "0" or "1" state.

- *The read signal.* Let us assume that C_{cell} is the capacitance of the storage capacitor and $C_{bitline}$ is the capacitance associated with the bitline. Then the differential signal developed between *BLT* and *BLC* is given by

$$V_s = \left(V_{node} - \frac{V_{dd}}{2}\right) \frac{C_{cell}}{C_{cell} + C_{bitline}}, \tag{9.11}$$

where $C_{cell}/(C_{cell} + C_{bitline})$ is referred to as the *transfer ratio*. To obtain a large read signal, V_{node} should be close to V_{dd} for a "1" state and close to zero for a "0" state, and

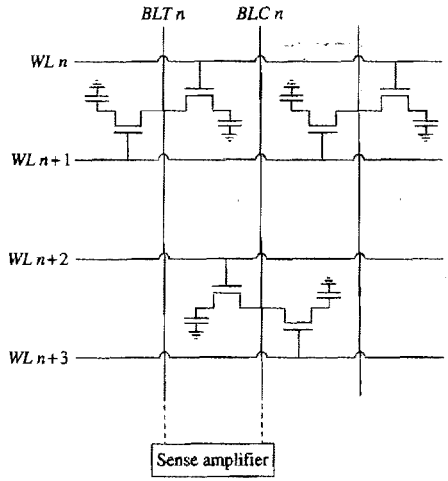


Figure 9.18. Schematic showing the cell arrangement in a folded-bitline architecture of DRAM array. Cells on adjacent bitlines are not on the same wordline.

$C_{cell}/C_{bitline}$ should not be too small. To minimize the chip area taken up by sense amplifiers and associated circuits, designers usually hang 256 to 1024 cells on one bitline. The capacitance contribution of each unselected cell to the bitline can be estimated to be ~1 fF/μm (per device width) from the drain-to-gate fringe capacitance and the drain junction capacitance of an off-state MOSFET (Table 5.2). The total $C_{bitline}$ would then be of the order of 100–300 fF assuming an effective transfer device width of ~0.3 μm. For a typical DRAM design with a choice of $C_{cell} = 30$ fF and a maximum $V_{node} = 1$ V, the transfer ratio is about 0.2 and the maximum V_s is about 100 mV.

- *Folded-bitline architecture.* Even though the maximum read signal may not be small, the DRAM circuits are usually designed to detect just a fraction of the signal before it is fully developed to achieve a short read time. Noise on a bitline can make it difficult to read a memory cell fast and reliably. To minimize the effect of noise, designers employ the so-called *folded-bitline architecture*, illustrated schematically in Fig. 9.18. In this case, the *BLC* is the bitline of another array of cells not connected to the same wordline as the cell to be read. Both *BLT* and *BLC* cross over the same wordline so that any voltage transient on the wordline is coupled equally to both bitlines. Thus, the signal being sensed is insensitive to noise originating from the wordlines. In a folded-bitline architecture, the minimum area taken up by one memory bit is $8F^2$ where $2F$ is the pitch of the wordlines and bitlines.

9.2.2 Device Design and Scaling Considerations for a DRAM Cell

Even though the storage capacitor may be fully charged in a write operation, leakage current can significantly degrade the signal by the time the cell is read. The design of a

DRAM cell is primarily driven by the desire to achieve simultaneously the smallest cell possible consistent with the lithography groundrules and the specified read signal and data *retention time* requirements. Data retention time is the time interval between data refreshes. More frequent read and refresh cycles mean higher chip power. A typical worst-case data retention time specified for a DRAM chip is about 100 ms. The retention time requirement sets the upper limit of the total leakage current allowed in a DRAM cell. For instance, for a capacitor storing 30 fC of charge (30 fF capacitor charged to 1 V), if we want to limit the charge loss to less than 10% between data refreshes, the maximum allowed leakage current from all sources is 30 fA. The insulator forming the storage capacitor is typically sufficiently thick so that the tunneling current through it is negligible. Also negligible is the diffusion-controlled reverse-biased junction leakage, of the order of 10^{-16} A/μm^2 at 100 °C (Section 2.2.4.11), as the junction area is only a fraction of 1 μm^2. The leakage current requirement mainly affects the design and scaling of the transfer device, which are discussed below.

- *Threshold voltage of the transfer device.* It was discussed in Section 4.2.1.2 that the MOSFET current at threshold is approximately 10^{-7} A for $W = L = 0.1$ μm, insensitive to temperature, and that the inverse slope of subthreshold current is 100 mV/decade at 100 °C. To satisfy the 30 fA off-current requirement, the threshold voltage of the transfer device would have to be at least 0.6 V at 100 °C or 0.7 V at 25 °C. This V_t value for the transfer device cannot be reduced in scaling unless the retention time requirement is reduced. In some designs, a transfer device with a natural V_t of less than 0.7 V can be used as long as the wordline at standby is held at a voltage more than 0.7 V below the natural V_t. This can be accomplished by holding the wordline at a negative voltage in standby.
- *Gate insulator thickness of the transfer device.* The gate insulator of the transfer device must be sufficiently thick so that the gate leakage current is less than 30 fA. Under the standby condition, the gate is low and the MOSFET is off. Electrons could tunnel from the gate to the positively biased drain if charge is stored (a "1" state). If we assume a gate-to-drain tunneling area of 0.1 × 0.01 μm^2, then the gate leakage current should be less than 10^{-3} A/cm^2. Figure 2.62 suggests that the gate oxide should be thicker than ~2 nm for $V_{dd} = 1$ V. This means that the DRAM transfer device cannot be scaled to as short a gate length as the high-performance logic device which can use a 1 nm thick gate oxide (see Fig. 4.7).

9.3 Nonvolatile Memory

In theory, any bistable device that retains its state when the power supply is disconnected can make a nonvolatile memory cell. If the memory cell cannot be reprogrammed, e.g., a *fuse* (which changes from a conducting state to a nonconducting state in programming) or an *antifuse* (which changes from a nonconducting state to a conducting state in programming), it is called a *programmable read-only memory* (PROM). A nonvolatile memory that can be erased and reprogrammed is called an *erasable programmable read-only memory* (EPROM). In the literature, the term EPROM includes nonvolatile memories that are reprogrammable but the memory erasure is done by nonelectrical means, e.g., by exposure to ultra-violet light. A nonvolatile memory that can be programmed and erased electrically is called an *electrically erasable and programmable read-only memory* (EEPROM). Notice that, as suggested by their full names, these nonvolatile memories are read-only memories. This means that, when used in a system, these nonvolatile memories really function as storage for data and program codes, and not as memories (like SRAM and DRAM) for running computer program codes. The reason for this read-only restriction is, as will be evident in the discussions to follow, that these nonvolatile memories do not have the read, write, and/or endurance properties required for running computer program codes.

The field of nonvolatile memory technology is extremely broad and rapidly evolving since many bistable elements or devices can retain their state when disconnected from their power supply. The technical considerations of a nonvolatile memory technology are: (i) memory speed, which includes access time, program time, and erase time; (ii) memory retention time, which measures how long a memory bit retains its state after being programmed; (iii) memory endurance, which measures how many cycles a memory bit can be programmed and erased while still functioning properly; (iv) power, which includes the power dissipation in programming, accessing, and erasing a memory bit; (v) power supply voltages, which include the voltages needed for program and erase operations; (vi) memory cell size; (vii) scaling properties of the memory technology. The choice of a nonvolatile memory technology depends on the application requirements and the cost involved. In this section, we discuss MOSFET-based nonvolatile memory devices and the basic principles of their operation. The reader is referred to the vast literature for further reading on the circuit and chip design aspects of the technology (e.g., Itoh, 2001; Hu, 1991, and the references therein).

9.3.1 MOSFET Nonvolatile Memory Devices

Figure 9.19 illustrates the basic principles of a MOSFET nonvolatile memory device. It is shown in Section 2.3.7.1 that the flat-band voltage depends on the amount and distribution of charge in the gate insulator. Let us assume that by some programming means we are able to inject a charge distribution $\rho_{net}(x)$ in the gate insulator (Fig. 2.50). This charge distribution causes a shift in the gate voltage needed to maintain the flatband condition, which means a shift in the device threshold voltage. From Eq. (2.220), the shift in threshold voltage is given by

$$\Delta V_t = -\frac{1}{\varepsilon_{ox}} \int_0^{t_{ox}} x \rho_{net}(x) dx = -\frac{1}{C_{ox}} \int_0^{t_{ox}} \frac{x}{t_{ox}} \rho_{net}(x) dx, \tag{9.12}$$

where $x = 0$ is the gate–oxide interface and $x = t_{ox}$ is the oxide–silicon interface. The injected charge is trapped in the gate insulator with a retention time of years, without the need of a power supply.

Figure 9.19(b) shows schematically the I_{ds}–V_{gs} characteristics before and after charge injection. For electron injections in an nMOSFET, ΔV_t is positive, i.e., the threshold

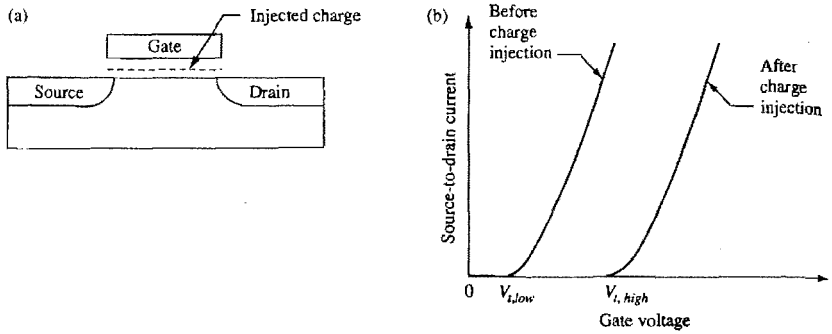


Figure 9.19. (a) Schematic diagram of a MOSFET nonvolatile memory device. (b) The MOSFET threshold voltage shifts from $V_{t,low}$ to $V_{t,high}$ after electron injection.

voltage increases from $V_{t,low}$ to $V_{t,high}$ after charge injection. Typically, ΔV_t can be a few volts. One can readily identify one of the states as logical "0" and the other logical "1". It is straightforward to read out the bit stored in the MOSFET by setting the gate voltage between $V_{t,low}$ and $V_{t,high}$. In one logical state, the MOSFET is on or conducting, and in the other, the MOSFET is off or nonconducting.

It is indicated in Fig. 2.28 that the energy barrier for electron injection into SiO_2 is 3.1 eV, significantly lower than that for hole injection, 4.6 eV. As a result, MOSFET-based nonvolatile devices usually employ electrons instead of holes for memory programming and erasure. In programming, the threshold voltage is shifted in a positive direction by injecting electrons from the channel region into the gate insulator and storing part or all of the injected charge within the gate insulator. In memory erasure, the stored charge is neutralized, usually by tunneling electrons out of the gate insulator region. In the subsections below, we consider the device physics related to charge injection, charge storage, and charge erasure in a MOSFET-based nonvolatile memory.

9.3.1.1 Charge Injection

- *By hot electrons*. Electron injection from silicon into SiO_2 can be by tunneling or by hot electron injection. If an n-channel MOSFET is used, usually channel hot electron injection is employed. Typical gate current and channel current in an n-channel MOSFET are shown in Fig. 9.20(a) and those in a p-channel MOSFET are shown in Fig. 9.20(b) (Hsu *et al.*, 1992). Note that in both cases, the gate current is an electron current.

  As discussed in Section 3.2.5.1, the substrate current is a direct measure of the amount of secondary carriers generated by impact ionization, which in turn is an indirect measure of the amount of primary hot carriers in the device channel region. From Fig. 3.34, we see that the substrate current, hence the amount of secondary carriers, is largest at gate voltages slightly above threshold voltage. In the case of an n-channel MOSFET, this is the region where the voltage difference between the gate and the drain, $V_{gs} - V_{ds}$, is negative and therefore does not favor injection of hot

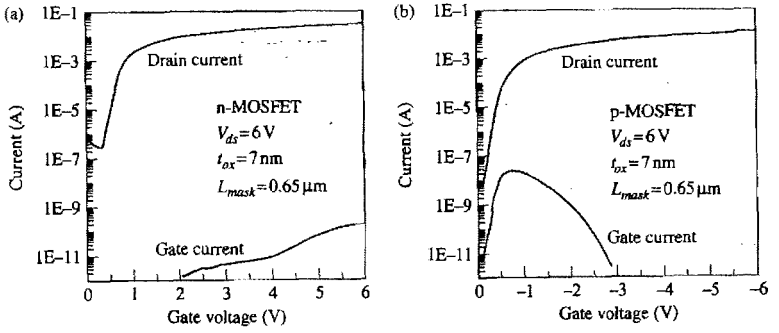


Figure 9.20. Typical drain and gate current characteristics in MOSFETs. (a) n-channel MOSFET. The gate current source is primarily the hot electrons flowing from source to drain. (b) p-channel MOSFET. The gate current is due to the injection of electrons generated via avalanche multiplication. (After Hsu *et al.*, 1992.)

electrons into the gate insulator. Hot electron injection becomes more favorable as V_{gs} approaches or exceeds V_{ds}, but, as explained in Section 3.2.5.1, the maximum electric field in the silicon decreases in that case. This mismatch between the gate voltage for maximum electric field in the channel and the gate voltage for maximum electric field in the gate insulator for hot electron injection makes the channel hot electron induced gate current in an nMOSFET quite low and the injection process highly inefficient. For the example in Fig. 9.20(a), the maximum electron current into the gate is about 10^{-10} A and the maximum I_g/I_{ds} ratio is only about 10^{-8}.

The situation is quite different when a p-channel MOSFET is used. In this case, electron injection is accomplished by injecting secondary hot electrons created by avalanche multiplication. Just like the nMOSFET in Fig. 3.34, the pMOSFET substrate current, which is a measure of the amount of avalanche generated electrons, also increases with gate voltage and peaks at V_{gs} slightly above V_t (magnitude-wise). At low gate voltages where the substrate current is rising or at its peak, the voltage difference between the gate and the drain, $V_{gs} - V_{ds}$, is positive (i.e., V_{ds} more negative than V_{gs}) and therefore in favor of hot electron injection from the silicon into the gate insulator. That is, the gate voltage favorable for generating avalanche hot electrons is also favorable for injection of hot electrons into the gate insulator. The hot electron gate current in Fig. 9.20(b) indeed has very similar dependence on gate voltage as the substrate current of nMOSFET in Fig. 3.34. Note that at the gate voltages where the avalanche hot electron gate current rises and peaks, the drain current is low. As a result, avalanche hot electron injection from silicon into gate insulator in a p-channel MOSFET is much more efficient. For the example in Fig. 9.20(b), the maximum gate current is about 10^{-8} A and the maximum I_g/I_{ds} ratio is about 10^{-4}.

Avalanche hot carrier injection can also be induced with the source of the MOSFET floating, i.e., without any channel current flowing. If a p-channel MOSFET source is left floating, the drain behaves like a p^+–n gated diode. As explained in Section 2.5.5, in a

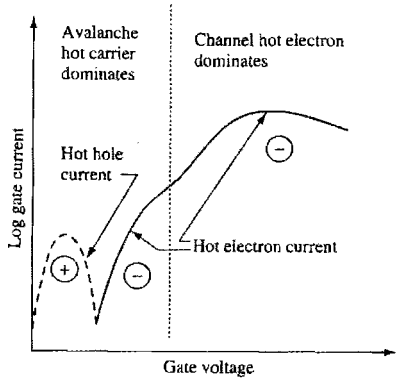


Figure 9.21. Schematic illustrating the injection of hot holes and hot electrons into the gate insulator region in an n-channel MOSFET as a function of gate voltage at large V_{ds}.

gated-diode mode, substrate carrier (electrons in a p-channel MOSFET) can be injected into the gate insulator if the gate is biased towards surface accumulation.

The injection currents shown in Fig. 9.20 do not tell the whole story. The fact that there is avalanche hot electron injection in a p-channel MOSFET at low gate voltages suggests that there should also be avalanche hot hole injection at low gate voltages in an n-channel MOSFET. Indeed, avalanche hot hole injection in an n-channel MOSFET at low gate voltages can be measured using ultra-sensitive current monitors (Takeda *et al.*, 1983; Nissan-Cohen, 1986). The expected gate current dependence on gate voltage is illustrated in Fig. 9.21. At low gate voltages, the electric field in the gate insulator favors the injection of avalanche hot holes. At somewhat higher gate voltages, the field in the gate insulator becomes less favorable for hot hole injection and more favorable for avalanche hot electron injection. As the gate voltage is increased further, the amount of hot carriers produced by avalanche multiplication actually decreases, as evidenced by the decrease in the substrate current in Fig. 3.34, and the field in the insulator becomes even more favorable for hot electron injection. Thus, the gate current becomes more and more dominated by channel hot electron injection as the gate voltage increases. Similarly, for a p-channel memory device, we should see avalanche hot electron injection at low gate voltages and channel hot hole injection at high gate voltages.

In an n-channel nonvolatile memory device using channel hot electron injection for programming, the injection of secondary hot holes at low gate voltages can have unintended consequences on non-selected cells on the same bitline. The injected positive charge shifts the V_t of non-selected devices in the negative direction, which, after repeated write cycles, may either unintentionally erase a previously programmed bit or cause device leakage when the V_t becomes too low. Care should be exercised to keep the wordline voltage of non-selected devices far enough below the threshold to avoid such "write disturbs" (Yamada *et al.*, 1996).

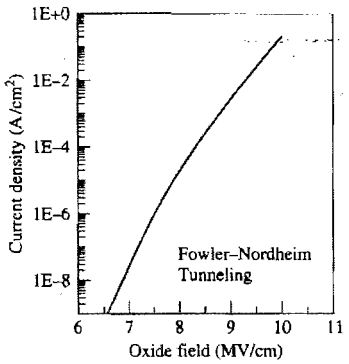


Figure 9.22. Fowler–Nordheim tunneling current density as a function of electric field in oxide. (Pavan *et al.*, 1997.)

- *By Fowler–Nordheim tunneling.* Electron injection in nonvolatile memory devices can also be accomplished by Fowler–Nordheim tunneling, discussed in Section 2.5.3.1. As shown in Fig. 2.61(b), electrons in the silicon conduction band tunnel through a triangular energy barrier into the oxide conduction band. The tunneling current density is a strong function of the electric field in the oxide, as plotted in Fig. 9.22. Typical field for programming is in the range of 8–10 MV/cm, employing an oxide 10 nm thick. Too thick an oxide would require higher voltages for programming. Too thin an oxide would lead to direct tunneling and charge leakage. Typical programming time for a nonvolatile memory cell is on the order of 1 μs to 1 ms.

9.3.1.2 Charge Storage

- *In silicon nitride layer.* In theory, electron traps in silicon dioxide can be used for charge storage for nonvolatile memory applications. However, the capture efficiency, which is the product of the trap density and the capture cross section, is just too small for silicon dioxide to be an effective electron storage medium for such an application (Ning and Yu, 1974; also see Section 2.3.6.4 on the properties of electron traps in silicon dioxide). An oxide–nitride–oxide (ONO) composite layer that can store plenty of electrons is commonly used instead. Electrons are stored mostly in the nitride layer. If charge injection is uniform and if ΔQ is the stored charge per unit area in the thin nitride layer at an average distance t_q from the gate electrode, then the threshold voltage shift is $\Delta V_t = -t_q \Delta Q/\varepsilon_{ox}$ from Eq. (9.12). The effect on V_t in case of nonuniform charge injection is discussed in Subsections 9.3.4.2 and 9.3.4.3.
- *In floating gate.* Another commonly used technique for enhancing charge storage within the gate insulator region of a MOSFET is to embed a conductive *floating gate*, typically just a thin layer of polysilicon, in the gate insulator between the silicon

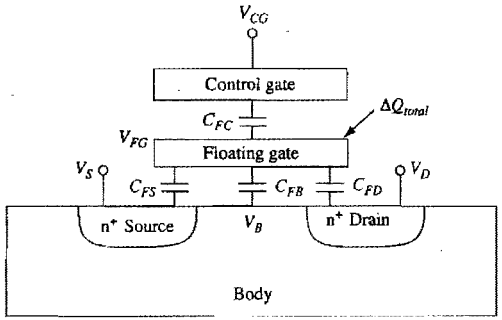


Figure 9.23. Capacitive coupling of the floating gate to other electrodes in an n-channel MOSFET nonvolatile memory device. Any depletion capacitance in the silicon body is absorbed in C_{FB}.

and the gate electrode. Electrons are stored in the floating gate, illustrated in Fig. 9.23. The stored charge spreads uniformly over the entire floating gate. When a floating gate is present, the usual gate electrode is referred to as the *control gate*, to differentiate it from the floating gate. The oxide between the floating gate and the silicon is called the tunnel oxide, typically 10 nm thick. The oxide between the control gate and the floating gate is called the inter-poly oxide, typically 20 nm thick. The overlap between the floating gate and the control gate can extend beyond the device region.

The potential of the floating gate, V_{FG}, is determined by the stored charge and the capacitive coupling of the floating gate to other electrodes surrounding it and their voltages. From the capacitive equivalent circuit in Fig. 9.23 (Wang, 1979),

$$\Delta Q_{total} = C_{FC}(V_{FG} - V_{CG}) + C_{FS}(V_{FG} - V_S) + C_{FB}(V_{FG} - V_B) + C_{FD}(V_{FG} - V_D). \tag{9.13}$$

Therefore,

$$V_{FG} = \frac{\Delta Q_{total} + C_{FC}V_{CG} + C_{FS}V_S + C_{FB}V_B + C_{FD}V_D}{C_{FC} + C_{FS} + C_{FB} + C_{FD}}. \tag{9.14}$$

One can define coupling factors of the various electrodes to the floating gate by the ratio of their individual capacitances to the total capacitance in the denominator. The presence of ΔQ_{total} shifts the floating gate potential by $\Delta V_{FG} = \Delta Q_{total}/(C_{FC} + C_{FS} + C_{FB} + C_{FD})$. From the control gate point of view, an additional $\Delta V_{CG} = -\Delta Q_{total}/C_{FC}$ needs to be applied to offset the effect of the stored charge and restore V_{FG} to the zero charge value. In other words, the threshold voltage shift due to the stored charge ΔQ_{total} on the floating gate is

$$\Delta V_t = -\frac{\Delta Q_{total}}{C_{FC}}. \tag{9.15}$$

Note that a larger control-gate coupling factor, $C_{FC}/(C_{FC} + C_{FS} + C_{FB} + C_{FD})$, gives more control of the floating gate potential to the control gate and yields a higher V_{FG} and therefore a higher field in the tunnel oxide for a given applied V_{CG}. High C_{FC}, however, means a smaller threshold voltage shift for a given charge injection.

9.3.1.3 Charge Erasure

In an EPROM, erasure is typically accomplished by exposing the device to high-energy photons, such as ultra-violet light or X-rays. The high-energy photons excite the stored electrons in the gate insulator or in the floating gate to an energy level above the oxide conduction band, thus allowing the excited electrons to flow back into the silicon substrate. It is cumbersome to carry out such a operation for a packaged chip.

In EEPROM devices, erasure is done by Fowler–Nordheim tunneling of electrons from the floating gate back to a positively biased source or drain region with the control gate grounded. For an applied erasure voltage V_S to the source, the potential difference across the tunnel oxide, $V_S - V_{FG}$, is obtained from Eq. (9.14) with $V_{CG} = 0$ (ignoring the V_B, V_D terms),

$$V_S - V_{FG} = \left(1 - \frac{C_{FS}}{C_{FC} + C_{FS} + C_{FB} + C_{FD}}\right)V_S - \frac{\Delta Q_{total}}{C_{FC} + C_{FS} + C_{FB} + C_{FD}}. \tag{9.16}$$

A small coupling factor from the source to the floating gate, $C_{FS}/(C_{FC} + C_{FS} + C_{FB} + C_{FD})$, is desired for achieving a high field in the tunnel oxide for erasure. Note that the stored negative charge adds to the field and helps the initial erasure speed. (By the same token, negative charge on the floating gate retards the field during programming and slows down the programming speed towards the end.) For an erasure voltage V_S in the range of 10–12 V, the erasure time is typically 0.1–1 s. The common practice is to carry out the erasure operation for a large block of cells simultaneously. This is discussed in the Flash memory section in 9.3.2.

For EEPROM devices using a nitride layer for charge storage, erasure can also be carried out by hole injection to neutralize the stored negative charge.

9.3.2 Flash Memory Arrays

So far we have discussed the basic program and erase operations of an individual bit in a nonvolatile memory array. In designing a nonvolatile memory array, it is desirable to be able to erase at once a block of memory bits or the entire array. In the case of UV-erased EPROM, the entire array is erased when it is exposed to radiation. In the case of EEPROM, this can be accomplished by connecting the bits to be erased together in an erase process. Masuoka first proposed to provide a special erase gate to an EEPROM array such that "the contents of all memory cells are simultaneously erased by using field emission of electrons from a floating gate to an erase gate in a *flash*" (Masuoka *et al.*, 1984). Since then, flash erasure of various forms has become common in most if not all EEPROM designs. Furthermore, the terms Flash, Flash memory, and EEPROM are now used interchangeably.

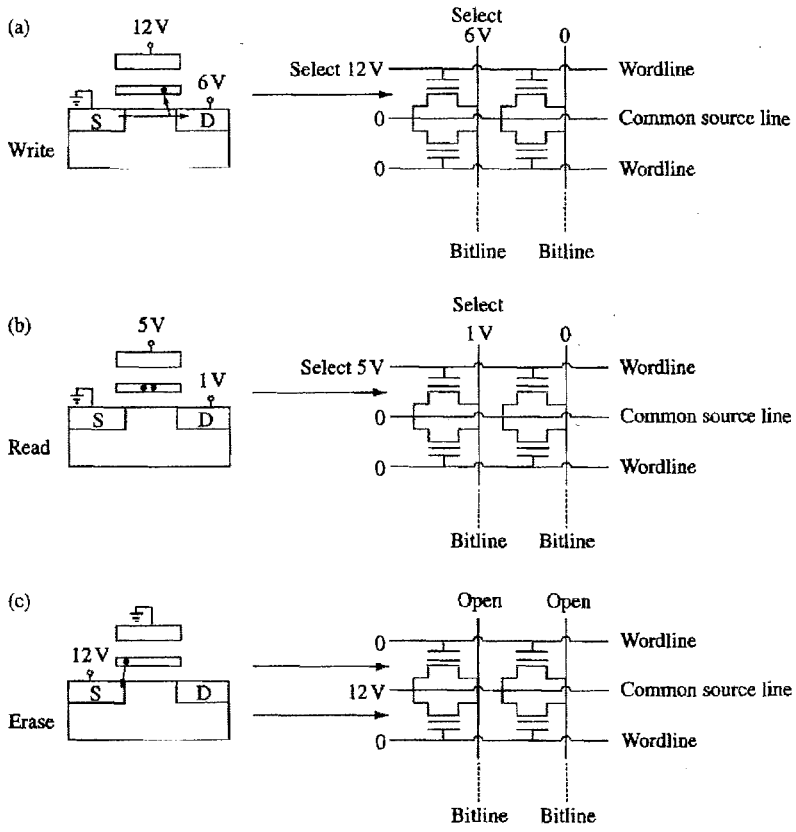


Figure 9.24. Schematics showing the connection of EEPROM devices to wordlines and bitlines in a memory array and their bias voltages for write, read, and erase operations. This is a *NOR* array. (After Itoh, 2001.)

9.3.2.1 Write, Read and Erase Operations

The write, read, and erase operations of a typical stacked-gate Flash memory array are illustrated in Fig. 9.24. In a write or program operation, a large positive voltage is applied to turn on the selected wordline and a high drain voltage is applied to the selected bitline to generate hot electrons near the drain where they are injected onto the floating gate. In a read operation, the selected wordline is biased at a voltage (5 V in this example) between $V_{t,low}$ and $V_{t,high}$ depicted in Fig. 9.19(b) and a positive voltage is applied to the selected bitline. The current in the bitline reflects the threshold voltage and therefore the state of charge storage in the cell. Note that all the non-selected cells on the same bitline are biased below $V_{t,low}$ to ensure that they are off regardless of their charge state. In an erase operation, all the sources are connected to a large positive voltage with all the control gates grounded. This causes the electrons stored on the floating gates to tunnel back to the source and erases a large block of cells simultaneously. Because the Fowler–Nordheim tunneling current is very sensitive to the field (Fig. 9.22), a slight variation in the thickness of the tunnel oxide could lead to a large spread of the erasure speed. It is usually necessary to verify that all the bits are back to their un-programmed state after each erasure. Repetitive erasure operations are carried out if there are bits not completely erased the first time.

A common problem encountered in a Flash memory array is *over erasure*. It happens when the number of electrons tunneled out of the floating gate is larger than the number of electrons injected into the floating gate during programming. It can also happen when holes tunnel or are injected into the floating gate (see Fig. 9.21). Over erasure results in threshold voltages lower than $V_{t,low}$ in Fig. 9.19(b) of un-programmed devices, causing a higher device off current than intended. In a memory array, if many of the non-selected cells (e.g., those with zero wordline voltage in Fig. 9.24(b)) on the same bitline as the selected cell are over-erased, their combined leakage current may be so large as to hinder sensing the on–off state of the selected cell. That is, over erasure can render parts of a memory array nonfunctional. Problems due to over erasure can be avoided by using a split-gate device (see Figs 9.29(b) and 9.30). With a sufficiently high threshold voltage on the single-gate part of the split-gate memory device, the cell off current can be controllably low independent of the charge level in the floating gate. The disadvantage of a split-gate device is its larger area.

9.3.2.2 NOR and NAND Architecture

The cell array architecture shown in Fig. 9.24 is called a *NOR* configuration, in which cells on the same bitline are connected in parallel and the unselected devices are turned off, i.e., with their wordlines biased below $V_{t,low}$ (Fig. 9.19(b)). Another commonly employed nonvolatile memory array architecture is called the *NAND* configuration, shown in Fig. 9.25. In a *NAND* configuration, memory cells on the same bitline are connected in series and the unselected devices are turned on, i.e., with their wordlines biased above $V_{t,high}$ (Masuoka *et al.*, 1987; Itoh *et al.*, 1989). In fact, $V_{t,low}$ is often made negative so un-programmed devices are normally on nMOSFETs. For read operation, the selected wordline is biased between $V_{t,low}$ and $V_{t,high}$ which can be zero if $V_{t,low} < 0 < V_{t,high}$.

Write and erase of a *NAND* array are done by Fowler–Nordheim tunneling. For write, the selected wordline is biased at a high positive voltage, say, 20 V, while the unselected wordlines are biased moderately above $V_{t,high}$, say, 10 V. To program a cell for charge injection, its bitline is grounded so that a high field for tunneling is established in the tunnel oxide. The field in the unselected cells on the same bitline is not high enough for charge injection. If no charge injection to the selected cell is desired, the bitline is biased positively at, say, 10 V, so there is no high field in any of the cells on that bitline. For block erase, a high positive voltage of 20 V is applied to the bitline and the substrate (to avoid junction breakdown) with all the wordlines tied to zero voltage to tunnel the stored electrons back to silicon.

Since no contacts to the source and drain of the memory devices are needed in a *NAND* architecture, it has significant density advantage over the *NOR* configuration. However, with the memory bits connected in series, the read current in a *NAND* array is low, resulting in relatively long access times for the memory array. *NAND* Flash is used

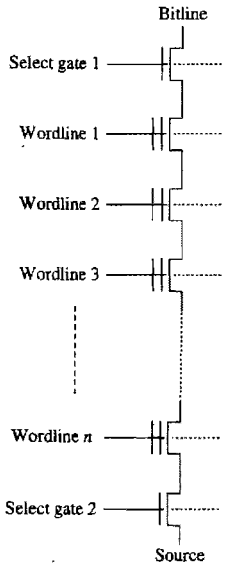


Figure 9.25. Schematic showing the serial connection of EEPROM devices in one bitline in a *NAND* array.

primarily for data storage where density is more important than access speed. *NOR* Flash is typically used in applications where access speed is important.

9.3.2.3 Endurance

The programming and erasure of an EEPROM device involve electric fields that are much higher than encountered in the normal operation of a MOSFET. As discussed in Sections 2.5 and 3.2.5, device degradation accompanies hot-carrier injection and electron tunneling into the gate oxide of a MOSFET. In the case of an EEPROM device, oxide degradation results in collapsing of the memory window (the threshold voltage difference between the programmed state and the erased state) as the device goes through many repetitive program and erase cycles. This is illustrated schematically in Fig. 9.26. The endurance of an EEPROM device is measured in terms of the number of program and erase cycles before the memory window is reduced to the point of inadequate margin. For a device that passes electrons through the same oxide location during both programming and erasure, the endurance is the lowest, often in the range of 10^3 to 10^4 cycles. The endurance can be improved by passing electrons in programming and in erasure through different oxide locations (e.g., programming through the drain and erasure through the source). Most Flash memories on the market have endurance specifications of 10^3 to 10^6 cycles.

It is also important to characterize the data retention of an EEPROM device after many program and erase cycles. As discussed in Section 2.5, when defects build up in an oxide

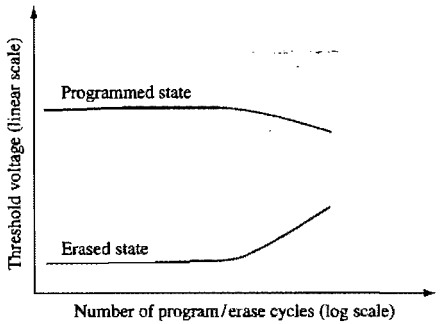


Figure 9.26. Schematic illustrating the collapse of the memory window as a function of the number of program and erase cycles.

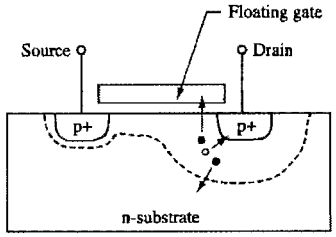


Figure 9.27. Schematic of a FAMOS device. The dotted line indicates the boundary of the depletion region.

layer, the oxide layer becomes leaky and its data retention characteristics degrade. Most Flash memories on the market have a data retention specification of ten years.

9.3.3 Floating-Gate Nonvolatile Memory Cells

Many versions of floating-gate nonvolatile devices have been developed. Judging from the level of research activities in the subject area, many more will be developed. Here we focus only on the devices that can be used to highlight the basic physics and operation principles of floating-gate nonvolatile devices.

9.3.3.1 UV-Erasable Floating-Gate Devices

One of the earliest successful floating-gate nonvolatile memory products was the *floating-gate avalanche-injection MOS* (FAMOS) device shown schematically in Fig. 9.27 (Frohman-Bentchkowsky, 1971). The device has no control gate. It is typically a p-channel device because, as discussed earlier in Section 9.3.1.1, the injection of hot electrons from silicon into gate insulator is much more efficient in a p-channel MOSFET than in an n-channel MOSFET. Programming is accomplished by avalanche hot electron injection. The device threshold voltage is designed such that the device is not conducting

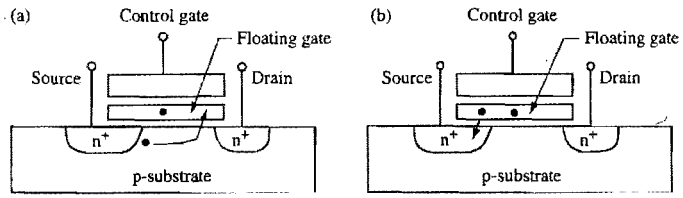


Figure 9.28. Schematic diagrams of a stacked-gate nonvolatile memory device. (a) Programming by channel hot electron injection. (b) Erasure by electron tunneling from floating gate to source.

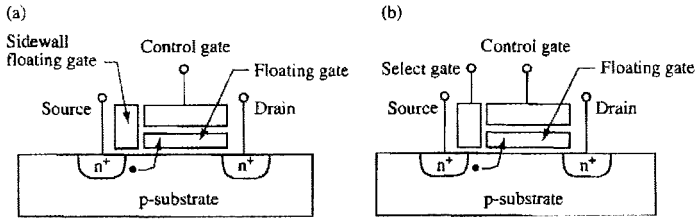


Figure 9.29. Schematics of floating-gate devices using source-side hot electron injection. (a) A stacked-gate device having a second floating gate at the source end. (b) A stacked-gate device having a second gate (select gate) at the source end.

when there are no electrons stored in the floating gate. When the device is programmed, the electrons in the floating gate induce an inversion channel of holes, thus making the device conducting. Erasure is typically accomplished by exposing the device to high-energy photons, such as ultra-violet light or X-rays.

9.3.3.2 Stacked-Gate Devices

To reduce the voltage needed for erasure in stacked-gate MOSFETs, designers sometimes use an insulator that is easier for electrons to tunnel through between the device channel and the floating gate, or between the floating gate and the control gate. For example, a silicon-rich oxide can be used to enhance tunneling (Hsu *et al.*, 1992). Another technique to enhance tunneling in a stacked-gate device is to use a thinner oxide in an extended overlap area between the gate stack and the source diffusion, as illustrated in Fig. 9.28. Programming is by channel hot electron injection near the drain region, and erasure is by electron tunneling from the floating gate to the source region.

9.3.3.3 Devices using Source-Side Injection

As shown earlier in Section 9.3.1.1, channel hot electron injection is a highly inefficient process, with only a tiny fraction of the channel carriers actually injected into the gate insulator. The injection efficiency can be greatly enhanced by using the device structures shown in Fig. 9.29. The structure in Fig. 9.29(a) has a sidewall floating gate at the source end (Wu *et al.*, 1986). The device is in effect two MOSFETs in series. The left (source side) device has a floating gate but no control gate, and the right (drain side) device is a

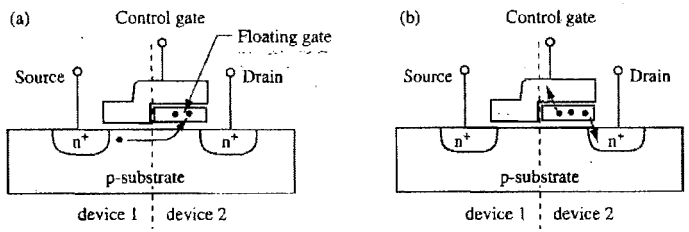


Figure 9.30. Schematic diagrams of a stacked-gate nonvolatile memory device having a split control gate. (a) Programming by channel hot electron injection. (b) Erasure can be accomplished by tunneling electrons from the floating gate to the drain region or to the control gate.

stacked-gate device. The sidewall floating gate is weakly coupled to the control gate. A positive control gate voltage causes a relatively strong surface inversion in the stacked-gate device and a relatively weak surface inversion in the floating-gate device. Being strongly inverted, the surface channel of the stacked-gate device behaves like an extended drain for the floating-gate device. When a large drain voltage is applied, the peak electric field in the silicon is located in the stacked-gate channel region but close to the sidewall floating-gate. Therefore, as the sidewall floating-gate device is turned on through coupling to the control gate, channel hot electrons are injected into the floating gate of the stacked-gate device at the source end. In the device structure shown in Fig. 9.29(b), instead of left floating, the sidewall gate is contacted to form a second control gate, called the select gate in the figure (Naruke *et al.*, 1989). The sidewall device can be turned on and off independently of the stacked-gate device, giving an additional degree of freedom in the operation of the memory device.

For a given drain voltage applied to a source-side injection device, the maximum electric field in the silicon channel increases with the control-gate voltage. Also, as the control-gate voltage increases, the electric field extracting the hot electrons from the channel to the floating gate increases. This is in contrast to the drain-side injection case discussed in Subsection 9.3.1.1 where the maximum channel field for hot carrier production decreases with gate voltage (see Fig. 3.34). Compared with a simple stacked-gate device with hot electron injection on the drain side, source-side injection devices have much better hot electron injection efficiency (Wu *et al.*, 1986; Naruke *et al.*, 1989).

9.3.3.4 Split- and Stacked-Gate Devices

If the select gate and the control gate in the device in Fig. 9.29(b) are electrically tied together, e.g., by forming them with the same polysilicon layer, then we have a *split-gate* EEPROM device shown schematically in Fig. 9.30. The device consists of a regular MOSFET (device 1) and a stacked-gate MOSFET (device 2) in series with a common gate electrode. Several variations of the split-gate EEPROM device have been reported, each having a different overlap coupling between the control gate and floating gate and between the floating gate and source–drain diffusions. In the version closest to that illustrated in Fig. 9.30 (Samachisa *et al.*, 1987), the channel region of device 1 behaves

like a source extension for device 2, and channel hot electrons can be injected into the floating gate near the drain region. Erasure can be done by tunneling the electrons from the floating gate to the drain region. In another version, the floating gate is purposely made to couple strongly to the drain diffusion so that the maximum electric field in the silicon is located close to the source end of the floating gate, thus achieving source-side channel hot electron injection (Kianian *et al.*, 1994). Furthermore, by intentionally shaping the floating gate edges to enhance the electric field locally where the control gate overlaps the floating-gate edge, erasure by electron tunneling from the floating gate to the control gate can be accomplished.

9.3.4 Nonvolatile Memory Cells with Charge Stored in Insulator

Several kinds of EEPROM devices with charge stored in the gate insulator have been reported, each using a different charge injection mechanism for programming and erasure. In this subsection, we discuss these devices and their basic operational principles.

9.3.4.1 MNOS Device Using Tunneling Injection

The simplest EEPROM device based on charge storage in the gate insulator region is one where programming and erasure are accomplished by tunneling electrons and/or holes into and out of a nitride layer. In the literature, such a device is often referred to as an MNOS (metal–nitride–oxide–semiconductor) device. Depending on whether it is a p-channel MOSFET or an n-channel MOSFET, electrons or holes can be injected from the device channel into the nitride layer for programming. In erasure, the stored charge can be driven to tunnel from the nitride layer either to the silicon or to the gate electrode. Since the charge is injected into the gate insulator region uniformly over the device channel, the effect on the surface potential is uniform over the entire channel region. Details of the transport of electrons and holes in MNOS devices can be found in the literature (e.g., Suzuki *et al.*, 1989, and the references therein.)

9.3.4.2 MNOS Device Using Channel Hot Electron Injection

If channel hot electrons are used to program an MNOS device, then the stored charge is localized near the drain. (Channel hot electron effect is discussed in Section 3.2.5.1.) This is illustrated in Fig. 9.31(a). The effect of stored charge on the surface potential in silicon is also localized. This makes the source and drain regions of the MOSFET asymmetric. Its threshold voltage then depends on the bias condition for the measurement. In the normal mode and with a high drain voltage, i.e., the same bias condition as that for programming, there is very little effect of the stored charge on the threshold voltage. This is because a MOSFET reaches threshold when the surface potential on the *source* side reaches $2\psi_B$ [Eq. (3.24)]. However, if the MOSFET is measured in the reversed source–drain mode, i.e., with the stored charge on the source side, the surface potential on the source side is retarded and there is a pronounced positive shift of the threshold voltage. These are schematically illustrated in Fig. 9.31(b) (Abbas and Dockerty, 1975; Ning *et al.*, 1977b). Thus, the memory effect is readily recognizable when the device is read in the reverse mode. The memory can be erased either by injecting hot holes to

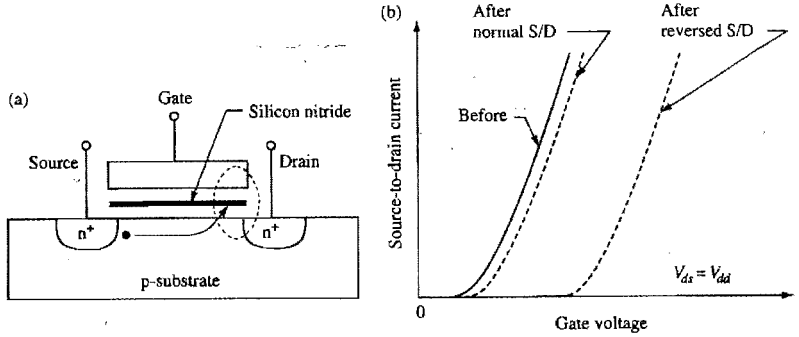


Figure 9.31. MNOS memory device using channel hot electron for programming. (a) Schematic of device cross section showing electrons being stored near the drain end after channel hot electron injection. (b) Schematic showing the current–voltage characteristics at large V_{ds} for three situations, namely, prior to programming, after programming with the device operated in the normal mode, and after programming with the device operated in the source–drain-reversed mode.

neutralize the stored electrons or by tunneling the stored electrons back to the device channel region or to the drain region (Chan *et al.*, 1987c).

9.3.4.3 MNOS Device Storing Two Bits per Cell

The source and the drain are structurally identical in an un-programmed MOSFET. Therefore, we can also inject electrons into the nitride layer near the source end by operating the MNOS device with the source and drain reversed. Furthermore, we can inject electrons into both the right (drain) end and the left (source) end of the nitride layer, thus creating two localized electron distributions, one at each end of the device channel. This is illustrated in Fig. 9.32(a). As long as the two charge distributions do not overlap much, the two localized charge distributions can be considered as distinctly separate hence can be used to represent two bits of memory information. In this way, two memory bits can be realized in one MNOS memory device (Eitan *et al.*, 2000). The expected current–voltage characteristics measured at large V_{ds} are illustrated schematically in Fig. 9.32(b). The key point is, at large V_{ds}, the measured device threshold voltage is determined primarily by the amount of charge stored in the gate insulator near the diffusion region used as the source for the measurement. At large V_{ds}, the charge stored near the diffusion region used as the drain for the measurement has relatively little influence on the measured device threshold voltage.

9.3.4.4 Devices with Other Charge Storage Material

More recently, other charge storage media besides silicon nitride have been explored. Among them, the most interesting one is to embed a thin region of nanocrystals of silicon within the gate insulator (Tiwari *et al.*, 1996). Charge is stored in the nanocrystals. The nanocrystals can enhance the tunneling of electrons into and out of the gate insulator. They can be used to implement the concept of two bits per device as well (Kim *et al.*,

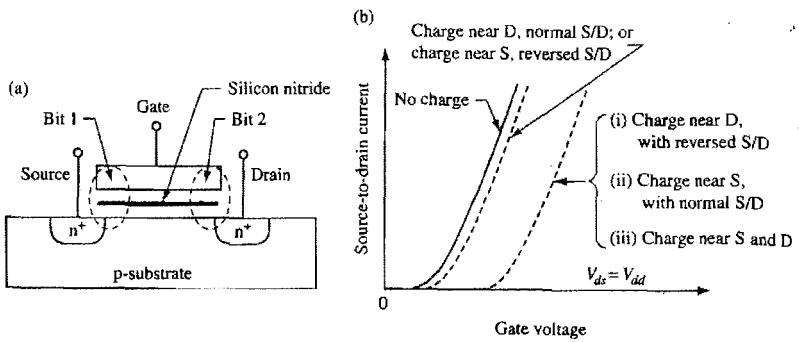


Figure 9.32. MNOS memory device using channel hot electron injection for programming. (a) Schematic of device cross section showing electrons being stored at both the drain and the source ends. (b) Schematic showing the expected current-voltage characteristics measured at large V_{ds}.

2003). No nonvolatile memory product employing nanocrystals for charge storage has yet been developed.

Exercise

9.1 Consider a bipolar transistor biased to operate in the saturation region. Ignoring parasitic resistances, the collector current is given by Eq. (9.5) and the base current is given by Eq. (9.6), namely,

$$I_C = I_{B0F} e^{qV_{BE}/kT}\left[\beta_0\left(1 - e^{-qV_{CE}/kT}\right) - \frac{I_{B0R}}{I_{B0F}} e^{-qV_{CE}/kT}\right],$$

and

$$I_B = I_{B0F} e^{qV_{BE}/kT}\left(1 + \frac{I_{B0R}}{I_{B0F}} e^{-qV_{CE}/kT}\right).$$

The current gain is

$$\beta \equiv \frac{I_C}{I_B} = \frac{\beta_0(1 - e^{-qV_{CE}/kT}) - \dfrac{I_{B0R}}{I_{B0F}} e^{-qV_{CE}/kT}}{1 + \dfrac{I_{B0R}}{I_{B0F}} e^{-qV_{CE}/kT}}.$$

Plot the current gain β as a function of qV_{CE}/kT for $qV_{CE}/kT = 0$ to $qV_{CE}/kT = 4$, using $\beta_0 = 100$ and $I_{B0R}/I_{B0F} = 1$. Repeat using $\beta_0 = 100$ and $I_{B0R}/I_{B0F} = 10$. This exercise shows that for practical transistors the current gain is less than 1 only for very small V_{CE} values.

10 Silicon-on-Insulator Devices

The last chapter of this book deals with silicon-on-insulator (SOI) devices, which include SOI CMOS (Lim and Fossum, 1983; Colinge *et al.*, 1986), SOI bipolar, and double-gate MOSFETs. They are not the mainstream VLSI technologies at present, but have the potential of playing a significant role in future generations.

There are three main types of SOI materials: SIMOX, BESOI, and Smart Cut® (Auberton-Hervé, 1996). SIMOX stands for synthesis by implanted oxygen. It is formed by first implanting a high dose of high energy oxygen ions into a silicon substrate. A high temperature anneal subsequently drives the chemical reaction which forms a stochiometric oxide layer buried in the silicon wafer. The anneal also regenerates the crystalline quality of the silicon layer remaining over the buried oxide. The main advantage of this technique is the thickness uniformity of the thin SOI layer. The main drawback is the high defect densities in the regrown silicon and in the buried oxide. BESOI stands for bond and etch back. It starts with two silicon wafers. After oxidation, the two wafers are bonded together by heating them to high temperatures. One of them is then etched back until only a thin film of silicon remains over the oxide and the other wafer. The crystalline quality of BESOI material is in principle as good as that of bulk silicon wafer. But there are usually significant thickness variations within the wafer. They can be tolerated for thick film SOI devices. For thin film SOI, some kind of etch stop technique needs to be employed in the etch back step to obtain better thickness uniformity. In the Smart Cut® process, both ion implantation and bonding are used. Before bonding, a high dose hydrogen implantation is made to wafer A to weaken the silicon bond strength at the implanted depth. After oxidation and bonding of wafers A and B, they are pulled apart mechanically. They break apart at the weakened cleavage plane, thus leaving a thin silicon film of A over the oxide and wafer B. The rest of A can be reused to save the cost. High temperature anneal and chemical-mechanical polish steps are done to the SOI wafer before device fabrication.

10.1 SOI CMOS

SOI CMOS involves building more or less conventional MOSFETs on a thin layer of crystalline silicon, as illustrated in Fig. 10.1. The thin layer of silicon is separated from the substrate by a thick layer (typically 100 nm or more) of buried SiO_2 film, thus

® Trade mark, SOITEC.

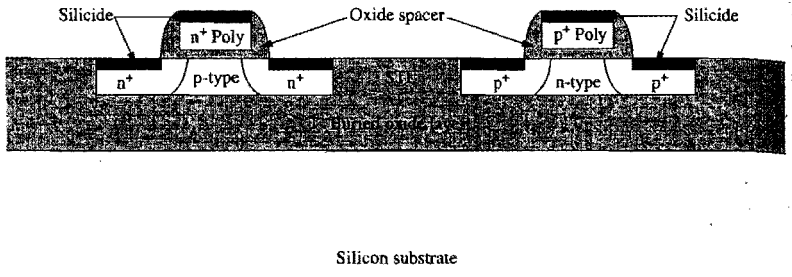


Figure 10.1. A schematic cross-section of SOI CMOS, with shallow trench isolation, dual polysilicon gates, and self-aligned silicide.

electrically isolating the devices from the underlying silicon substrate and from each other. SOI CMOS process can be readily developed due to the compatibility with established bulk processing technology.

The inherent advantages of SOI devices over bulk CMOS are listed below.

- *Very low junction capacitance*

  The source and drain junction capacitance is almost entirely eliminated in SOI MOSFETs. The capacitance through the thick buried oxide layer to the substrate is very small.
- *No body effect*

  The threshold voltage of stacked devices in SOI, e.g., transistor N1 in Fig. 5.10, is not degraded by the body effect since their body potential is not tied to the ground or V_{dd} but can rise to the same potential as the source (node V_x in Fig. 5.10).
- *Soft error immunity*

  In bulk devices, minority carriers are generated along the track of any high-energy particle or ionizing radiation that strikes through the silicon. If the collected charge of a junction node exceeds a certain threshold, it may cause an upset of the stored logic state. This is commonly referred to as a *soft error*. SOI devices offer a potential improvement in the soft-error rate since the presence of the buried oxide greatly reduces the volume susceptible to ionizing radiation.

10.1.1 Partially Depleted SOI MOSFETs

SOI MOSFETs are often distinguished as *partially depleted* (PD) when the silicon film is thicker than the maximum gate depletion width and the devices exhibit floating-body effect (Yoshimi *et al.*, 1989), and *fully depleted* (FD) when the silicon film is thin enough that the entire film is depleted before the threshold condition is reached.

In PD SOI MOSFETs, there is a quasineutral body region between the gate depletion boundary and the buried oxide. The short-channel effects of PD SOI MOSFETs are governed by the scale length theory just like bulk MOSFETs. However, there is a unique floating-body or kink effect which occurs when carriers of the same type as the body, generated by impact ionization near the drain and thereafter stored in the floating body,

affect the device threshold voltage (Yoshimi *et al.*, 1989). This effect is especially strong in nMOSFETs owing to the higher impact ionization rate of electrons (Fig. 2.59). Floating-body effects become rather complex in dynamic conditions. In a practical switching event, the charging-up of the floating body by impact ionization or junction leakage usually takes a longer time than the input transition. In fast gate ramps, the body potential tends to rise with the gate potential, which reduces V_t and increases the transient current. This phenomenon is referred to as the *drain current overshoot*. ***Even though floating-body effects tend to enhance circuit speed in certain conditions, the drain current overshoot (or undershoot) is history dependent*** (Gautier *et al.*, 1995; Sherony *et al.*, 1996). The floating-body potential depends on how recently and how often the device has been switched through its high impact ionization conditions. The worst case is when the device (usually the nMOSFET) has been in the on state for a long time and is then turned off, but is turned back on again before the body charge reaches the equilibrium state. This adds complexity to circuit design. Another undesirable consequence of the floating-body effect in a PD SOI is the higher off-current due to a forward-biased body-to-source junction when the drain voltage is high. Using body contacts can restore the device characteristics of SOI MOSFETs back to the bulk-MOSFET-like characteristics (Chen *et al.*, 1996). Body contact, however, carries a delay penalty and loses the body effect advantage of SOI devices.

The delay equation (5.39) can be used to assess the performance benefit of SOI over bulk CMOS. There is negligible improvement of R_{sw} from SOI as it uses the same channel length and oxide thickness as the bulk CMOS (assuming that the floating body effect is either negligible or neutral). A major part of the SOI advantage comes from its inherently low junction capacitance (Fig. 10.1). Assuming that the junction capacitance is zero in SOI devices (actually only the areal component is zero, a small perimeter component is still present in SOI), one can express the propagation delay of SOI CMOS with respect to that of bulk as

$$\frac{\tau_{soi}}{\tau_{bulk}} = 1 - \frac{C_J}{C_{out} + \mathrm{FO} \times C_{in} + C_L}, \tag{10.1}$$

where C_J is the junction capacitance contribution to C_{out} (Fig. 5.19, including both n- and pMOSFETs). Note that there is no junction component in C_{in} (Table 5.4). Clearly, the highest percentage of improvement occurs in unloaded circuits with $C_L = 0$ and FO = 1. With the folded layout in Fig. 5.14(b) in which the junction capacitance and therefore its contribution to C_{out} are effectively cut to half, there is less benefit of SOI over bulk. For heavily loaded circuits, $C_L >> C_{in}$ and C_{out}, the benefit from SOI is negligible.

For the example of 0.1 μm CMOS devices (Table 5.2) analyzed in Section 5.3, the C_{in} and C_{out} figures are given in Table 5.3 (non-folded layouts) with the percentage of junction capacitance in C_{out} listed in Table 5.4. Using Eq. (10.1), one can estimate the delay improvement of SOI devices over bulk CMOS for given fan-out and load capacitance. There is a substantial benefit of SOI over bulk, namely $\tau_{soi}/\tau_{bulk} \approx 0.69$ for FO = 1 and $C_L = 0$. The improvement becomes less at higher numbers of fan-outs. For example, $\tau_{soi}/\tau_{bulk} \approx 0.84$ for FO = 3 and $C_L = 0$. With a folded layout, the corresponding figures are $\tau_{soi}/\tau_{bulk} \approx 0.82$ for FO = 1 and $C_L = 0$, and $\tau_{soi}/\tau_{bulk} \approx 0.91$ for FO = 3 and $C_L = 0$.

Since the improvement factor of SOI over bulk varies widely depending on the circuit loading condition, extensive re-design is desirable to take full advantage of the SOI technology. For example, the device widths should be increased to lessen the effect of wire loading on circuit delays. In general, it is expected that a delay improvement somewhere between the intrinsic and the heavily loaded figures can be achieved at the chip level.

10.1.2 Fully Depleted SOI MOSFETs

Floating-body effect can be largely avoided in fully depleted (FD) SOI devices in which either the silicon film is thin enough or the doping is light enough that the entire film is depleted, i.e., there is no neutral region in the body. In fact, the entire silicon film can be undoped because FD SOI MOSFETs scale by the silicon film thickness, not by the gate depletion width (W_{dm}) as bulk and PD SOI CMOS do. The inverse subthreshold slope of a long-channel FD SOI MOSFET can be near the ideal 60 mV/decade number at 300 K. This is because the anchor point of potential in Fig. 3.5 extends all the way through the buried oxide (BOX) to the substrate. It is so far from the gate that the body effect coefficient $m = 1 + 3t_{ox}/W_{dm} \approx 1$. The steeper subthreshold slope permits a lower V_t for the same off-current, which in turn allows the devices to be used at lower supply voltages thereby attracting attention for low power operation. However, in a short-channel FD SOI MOSFET, the subthreshold current slope can be quickly degraded due to severe short-channel effects discussed below.

The problem is that ***in short-channel FD SOI MOSFETs, the thick buried-oxide region (equivalent to the gate depletion region) leaves them vulnerable to source–drain field penetration, which results in poor short-channel effect*** (Yan *et al.*, 1991; Su *et al.*, 1994). The scale length model discussed in Appendix 10 does not apply to FD SOI MOSFETs because the buried-oxide region, like the ε_3 region in Fig. A10.2, is so thick compared with the device dimension that linear interpolation of potential in the gaps between the source and drain and the substrate does not work.[1] In other words, the spacing between the source–drain and the bottom conductor is so large that no closed rectangle can be defined with known potential values on its boundary.

Figure 10.2 shows the 2-D field line plot of a simulated 50 nm FD-SOI MOSFET with a 200 nm thick buried oxide (Lu and Taur, 2006). The field lines in the buried oxide are highly two-dimensional within a ~25 nm depth from the silicon body. The penetration depth of the 2-D field is a function of channel length and is insensitive to BOX thickness. In fact, once the BOX is thick enough (much greater than the channel length), the 2-D field pattern and therefore the short-channel effect is independent of BOX thickness. Use of the three-region scale length model in Appendix 10 with t_3 (Fig. A10.2) being the BOX thickness would grossly overestimate short-channel effects in FD-SOI devices.

The 2-D numerical simulation is necessary to evaluate the minimum channel length of FD-SOI MOSFETs for given t_{ox} and t_{si} with a thick t_{BOX}. Figure 10.3 shows several examples of constant L_{min} contours in a t_{ox}–t_{si} plane (Lu and Taur, 2006). Here L_{min} is

[1] In double-gate like structures discussed in Section 10.3, the back oxide is as thin as the front oxide. Then the three-region scale length model is applicable, with more favorable short-channel scaling characteristics.

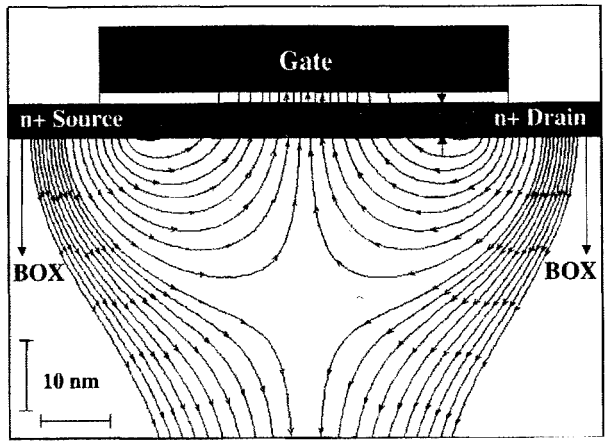


Figure 10.2. The 2-D electric field lines in a FD-SOI MOSFET with $L = 50$ nm, $t_{ox} = 1.5$ nm, $t_{si} = 5$ nm, and $t_{BOX} = 200$ nm (bottom not shown). The bias voltages are $V_{gs} = -0.3$ V (n^+ poly gate) and $V_{ds} = 0$.

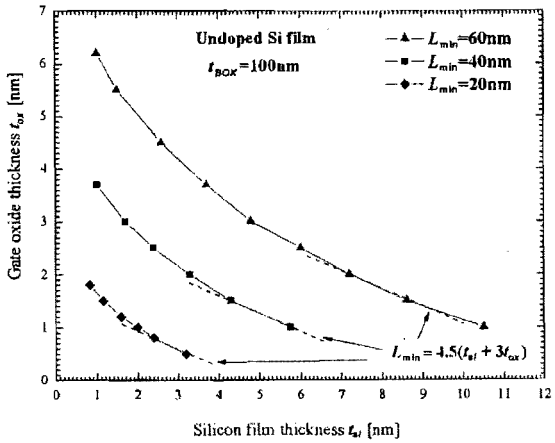


Figure 10.3. Contours of constant L_{min} in a t_{ox}–t_{si} plane for FD-SOI MOSFETs. The points are from 2-D numerical simulations. The dashed lines show the fitting of the data by Eq. (10.2).

defined as the channel length at which the high-drain-bias ($V_{ds} = 1$ V) threshold voltage rolloff equals 100 mV. As expected, if both t_{ox} and t_{si} are scaled by a common factor, L_{min} scales by that factor as well. A simple empirical relationship for the region where $t_{ox} \ll t_{si}$ is found to be

$$L_{min} \cong 4.5(t_{si} + 3t_{ox}). \tag{10.2}$$

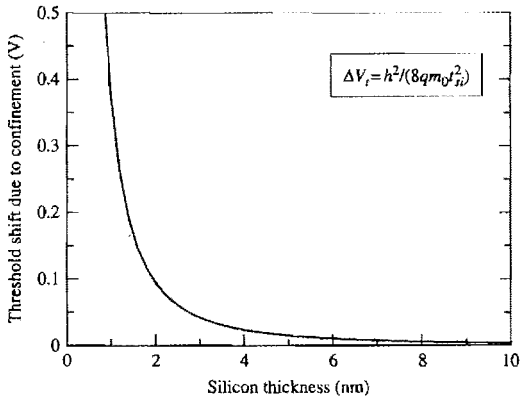


Figure 10.4. Increase of threshold voltage due to quantum confinement versus silicon film thickness in an ultra-thin SOI MOSFET.

This is roughly a factor of two longer than that of bulk MOSFETs with the same t_{ox} and a maximum depletion width W_{dm} the same as t_{si} (Section 3.2.1.4). Equation (10.2) can be generalized to high-κ gate insulators by replacing $3t_{ox}$ in the parentheses with $(\varepsilon_{si}/\varepsilon_i)t_i$ (Lu and Taur, 2006).

In principle, t_{si} in FD-SOI can be scaled further than W_{dm} in bulk MOSFETs since the latter is bounded by body effect considerations, $m = 1 + 3t_{ox}/W_{dm} \leq 1.4$, given a minimum t_{ox} limited by QM tunneling (Section 4.2.2.1). Quantum mechanical considerations also impose a practical limit on the silicon thickness. For very thin (<5 nm) silicon films, electrons are confined not only by the field (Appendix 12) but also by the physical thickness. Uncertainty principle requires that the momentum uncertainty must increase as $\sim h/t_{si}$ as t_{si} is scaled down. This translates into an electron ground-state energy significantly above the conduction band of silicon. Quantitatively, the solution of Schrödinger's equation for a particle in an infinite well of width t_{si} yields a ground-state energy of $E_0 = h^2/(8m^*t_{si}^2)$, where m^* is the electron effective mass. The device implication is that the threshold voltage of an ultra-thin SOI nMOSFET will increase by $\Delta V_t = E_0/q = h^2/(8qm^*t_{si}^2)$. Figure 10.4 plots such a threshold shift versus t_{si} assuming $m^* = m_0$, the free electron mass. For lighter effective masses, the quantum effect is even stronger. The threshold voltage increase becomes significant (>0.1 V) when the silicon thickness is thinner than 2 nm. Not only does this require a gate work function lower than that of n^+ silicon (4.1 eV) for realizing low nMOSFET threshold voltages, but the sensitivity of ΔV_t to silicon film thickness also poses a serious tolerance problem. Thickness control near atomic dimensions would be required. Recent experimental results (Uchida *et al.*, 2002) also suggest that lack of atomically flat surfaces could be the cause of severe mobility degradation in ultra-thin silicon films. With a minimum $t_{si} \approx 2$ nm, the above considerations project a channel length limit for FD SOI CMOS of ~20 nm with oxides and ~10 nm with high-κ insulators, comparable to the limits of bulk CMOS discussed in Section 4.2.2.1.

10.2 Thin-Silicon SOI Bipolar

As explained in Section 8.4.2.4, the large standby power dissipation of bipolar circuits has limited bipolar technology primarily to RF and analog applications, in which the number of transistors needed is relatively small and the characteristics of bipolar transistors, particularly SiGe-base bipolar transistors, are often preferred. Also, designers often prefer BiCMOS technology (bipolar and CMOS integrated on the same chip) for *mixed-signal* applications where the CMOS is used primarily for the digital logic functions and the bipolar is used primarily for the analog and RF functions. From a mixed-signal system perspective, SOI is attractive because an SOI substrate can provide good electrical isolation between the digital and the RF and analog components, particularly when a thick insulator or a high-resistivity substrate is used (see, e.g., Washio *et al.*, 2000) or when substrate engineering is applied (see, e.g., Burghartz *et al*, 2002). Reduction of substrate loss is required for the integration of high-quality-factor passive elements on chip.

A major challenge in building SOI BiCMOS integrated circuits arises from the fundamental structural difference between high-speed bipolar and high-speed SOI CMOS devices. As illustrated in Fig. 10.1, the silicon layers in high-speed SOI CMOS are sufficiently thin that the source and drain diffusion regions reach down to the buried oxide layer. For instance, the silicon layer thickness for 130-nm SOI CMOS is only about 120 nm. In vertical bipolar transistors, the subcollector layer alone is typically 1–2 μm thick. In the literature, there have been many reports of development or application of SOI BiCMOS technology where the silicon layers are thick enough to accommodate the subcollector layer of the vertical bipolar transistors. In such thick-silicon SOI BiCMOS, the silicon thickness is much larger than the depth of the source and drain diffusion regions so that the CMOS devices behave like regular bulk CMOS devices instead of high-speed SOI CMOS devices. When bipolar was still used to build high-speed mainframe computers, digital BiCMOS technology on thick-silicon SOI was developed to reduce soft-error susceptibility (Hiramoto *et al.*, 1992; Hashimoto *et al.*, 1998). In recent years, the development of thick-silicon SOI BiCMOS technology has shifted to mixed-signal applications. The focus is on integrating SiGe-base bipolar devices, bulk CMOS devices, and passive components all on the same chip (see, e.g., Washio *et al.*, 2000).

The development of an SOI SiGe-base bipolar transistor using silicon thickness compatible with high-speed SOI CMOS has been reported recently (Cai *et al.*, 2002a; Ouyang *et al.*, 2002; Cai *et al.*, 2003; Cai and Ning, 2006). This thin-silicon SOI SiGe-base transistor requires no subcollector layer. In theory, a thin-silicon SOI SiGe-base bipolar transistor can be integrated with SOI CMOS to give SiGe-base bipolar transistors and SOI CMOS on the same chip. In fact, in theory, a truly complementary SOI BiCMOS should be possible, with vertical SiGe-base n–p–n, vertical p–n–p, and SOI CMOS devices all on the same chip (Ning, 2003b).

A schematic of a thin-silicon SOI SiGe-base n–p–n bipolar transistor is shown in Fig. 10.5. It has the same emitter and base structure as a regular SiGe-base bipolar transistor, but there is no subcollector layer. The collector reach-through is only as deep

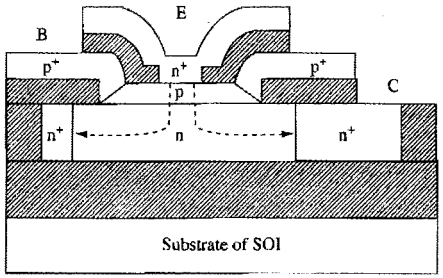


Figure 10.5. Schematic cross section of a thin-silicon SOI SiGe-base bipolar transistor. The dotted arrows indicate the path of electrons from the emitter to the collector reach-through. (After Cai *et al.*, 2003.)

as the silicon thickness, like the source/drain diffusion of an n-channel SOI MOSFET. In operation, instead of flowing from the emitter through the base to a subcollector, as in a regular SiGe-base bipolar transistor, the electrons make a more-or-less 90° turn below the base layer and flow towards the heavily doped collector reach-through region. The emitter–base diode and the intrinsic-base layer of a thin-silicon SOI bipolar transistor have the same characteristics as those of a regular vertical bipolar transistor. Only the collector region behaves differently.

To first order, the behavior of the collector region can be categorized into one of three modes: (i) the collector is fully depleted, (ii) the collector is partially depleted, and (iii) an accumulation layer is formed at the buried-oxide–silicon interface. The operation mode of the collector is determined by a combination of collector doping concentration N_C, silicon layer thickness t_{si}, substrate bias voltage V_S, and the separation L_{CE} between the collector reach-through edge and the emitter edge. The three modes for the collector region are depicted schematically in Fig. 10.6 and discussed below.

10.2.1 Fully Depleted Collector Mode

When the collector is fully depleted (Fig. 10.6(a)), the collector contains no n-type quasineutral region except adjacent to the heavily doped reach-through. After traversing the intrinsic-base layer, the electrons drift along the electric field lines in the collector space-charge region more-or-less laterally towards this quasineutral region. In a one-sided junction depletion approximation, the collector region directly underneath the intrinsic base is fully depleted if N_C is less than the value given in the equation [see Eq. (2.85)]

$$t_{si} = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi,CB} + V_{CB})}{qN_C}}, \tag{10.3}$$

where $\psi_{bi,CB}$ is the collector–base diode built-in potential and V_{CB} is the collector–base voltage. For example, if $t_{si} = 100$ nm and $V_{CB} = 2$ V, Eq. (10.3) suggests that the collector is in a fully depleted mode if N_C is less than 2×10^{17} cm^{-3}.

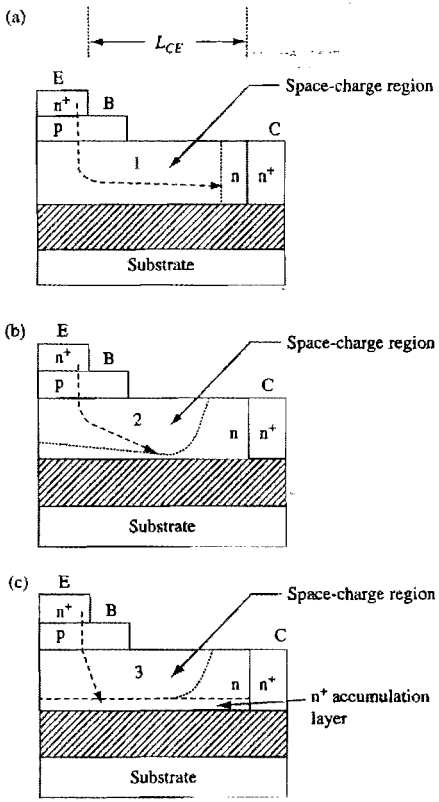


Figure 10.6. Schematics illustrating the electron flow paths in the collector space-charge region of a thin-silicon SOI bipolar transistor. (a) The collector region is fully depleted. After traversing the intrinsic-base layer, the electrons drift more-or-less laterally towards the collector reach-through, as indicated in flow path 1. (b) The collector is partially depleted. The electrons drift more or less diagonally from the base towards the collector reach-through, as indicated in path 2. (c) The collector is in accumulation. The electrons flow mostly vertically down from the base to the accumulation layer, as indicated in path 3. (After Cai *et al.*, 2003.)

In the fully depleted mode, the electron flow path in the base–collector space-charge region (path 1 in Fig. 10.6(a)), and hence the base–collector space-charge-region transit time t_{BC}, is a function of L_{CE} and the emitter-stripe width W_E. The dependence on W_E comes from the fact that an electron originating from the center of the emitter has a flow path that is approximately $W_E/2$ longer than an electron originating from the very edge of the emitter. As a result, the performance of the transistor improves as W_E and L_{CE} are reduced (Ouyang *et al.*, 2002).

It should be noted that Eq. (10.3) gives the condition for full depletion when the transistor is in the off state ($V_{BE} = 0$). As shown in Section 6.3.3.1, when a transistor is

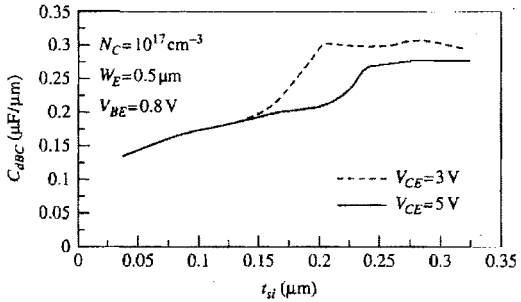


Figure 10.7. Simulated base–collector junction capacitance per emitter-stripe length as a function of silicon layer thickness for a typical thin-silicon SOI bipolar transistor. (After Ouyang *et al.*, 2002.)

turned on, its base–collector space-charge region widens because the mobile electrons in the space-charge region reduce the net charge in the region. Therefore, instead of Eq. (10.3), the condition for full-depletion operation when the transistor is on is given by

$$t_{si} = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi,CB} + V_{CB})}{q[N_C - \Delta n(V_{BE})]}}, \tag{10.4}$$

where $\Delta n(V_{BE})$ is the average electron concentration in the space-charge region when the emitter–base diode is forward biased at V_{BE}. In other words, for given N_C the value of t_{si} for a transistor to operate in the fully depleted collector mode can be substantially larger than suggested by Eq. (10.3).

The simulated collector–base junction capacitance per unit emitter-stripe length as a function of t_{si} is shown in Fig. 10.7 for a transistor biased with $V_{BE} = 0.8$ V (Ouyang *et al.*, 2002). The rapid drop in the collector–base junction capacitance as t_{si} is reduced indicates the transition from a "partial depletion" mode to a "full depletion" mode. The t_{si} value at which the transition starts is almost twice that suggested by Eq. (10.3). The reduced capacitance suggests that it is possible to have superior device performance at low power dissipation in the full depletion mode, provided that W_E and L_{CE} are sufficiently small so that t_{BC} is not the dominant delay component. Also, since the Early voltage is inversely proportional to the collector–base junction capacitance [see Eq. (6.72)], the smaller collector–base junction capacitance suggests that a transistor designed to operate in the full depletion mode should have a larger Early voltage as well.

10.2.2 Partially Depleted Collector Mode

When the collector doping concentration is larger than that given by Eq. (10.4) for a transistor, the collector is no longer in a fully depleted mode. There can be a finite quasineutral n-type region above the buried-oxide. After traversing the intrinsic base, the electrons will flow more-or-less vertically down to this quasineutral collector layer, just

like in a regular vertical bipolar transistor. As the electrons are conducted away laterally in this quasineutral layer towards the n^+ reach-through, they cause an *IR* drop in the quasineutral layer. The electrons originating from around the emitter edge see a smaller *IR* drop than the electrons originating from around the center of the emitter. The net result is that the average electron flow, as illustrated in path 2 in Fig. 10.6(b), has a path length that is shorter than the case of a fully depleted collector (path 1 in Fig. 10.6(a)). In many respects, a transistor in the partially depleted collector mode behaves like a regular vertical bipolar transistor having a high-resistance subcollector.

10.2.3 Accumulation Collector Mode

If a large positive voltage is applied to the substrate, an electron accumulation layer can be formed in the collector at the buried-oxide–silicon interface. If the collector is partially depleted when the substrate is zero biased, then the electron accumulation layer simply reduces the collector series resistance and has little effect on other device parameters. The reduced collector resistance leads to somewhat improved f_T [see Eq. (8.9)], which in turn leads to improved f_{max}.

On the other hand, if the collector is fully depleted at zero substrate bias, then forming an electron accumulation layer can have a rather dramatic effect on the device characteristics and performance (Cai *et al.*, 2003). When electron accumulation is strong, the transistor behaves like a regular vertical bipolar transistor, except that the collector layer is totally depleted and the thick subcollector layer is replaced by an ultra-thin electron layer. The collector–base junction capacitance becomes large, even larger than a regular vertical bipolar transistor because the collector–base space-charge layer thickness is smaller than in a regular bipolar transistor. The electron accumulation layer reduces the collector series resistance. The electron current flow path in the collector space-charge region, depicted as path 3 in Fig. 10.6(c), is short, being determined by the silicon layer thickness. That is, t_{BC} is the smallest in this mode. The reductions in r_c and t_{BC} combine to improve f_T. The improvement in peak f_T can be quite large. As for f_{max}, the improvement may not be as large because of the increase in collector–base junction capacitance. This is illustrated in Fig. 10.8 for an experimental thin-silicon SOI SiGe-base bipolar transistor (Cai *et al.*, 2003). Figure 10.9 is a plot showing the collector–base junction capacitance and the collector series resistance as a function of substrate–emitter bias voltage, extracted from the measured small-signal data for the same transistor. At zero substrate bias, the transistor operates in a fully depleted mode, showing small collector–base junction capacitance and large collector series resistance. As the substrate voltage is increased to form an electron accumulation layer, the collector–base junction capacitance increases and the collector series resistance decreases.

10.2.4 Discussion

More and more experimental and modeling studies of thin-SOI SiGe-base bipolar transistors are being reported (e.g., Avenier *et al.*, 2005; Chen *et al.*, 2005; Duvernay *et al.*, 2007). However, how this technology will be developed and exploited remains to

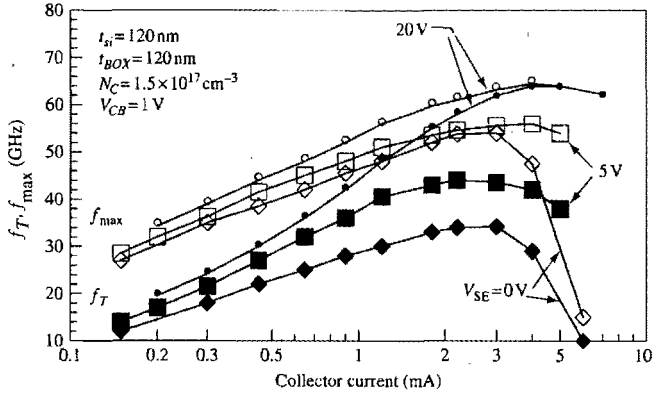


Figure 10.8. Measured f_T and f_{max} as a function of collector current for a thin-silicon SOI SiGe-base bipolar transistor for three values of substrate–emitter bias voltage. The transistor has an emitter area of 0.16 × 16 μm^2. (After Cai *et al.*, 2003.)

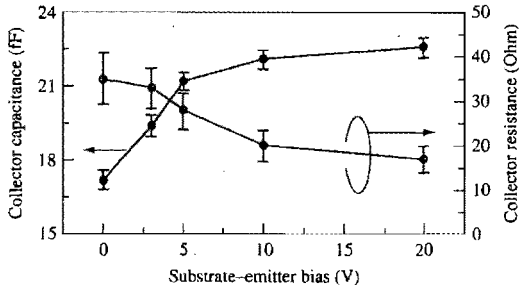


Figure 10.9. Collector capacitance and collector series resistance as a function of substrate–emitter bias voltage extracted from measured data for the same transistor as shown in Fig. 10.8. The data were taken at V_{CB} = 1 V, V_E = 0 V, and I_C = 1 mA. (After Cai *et al.*, 2003.)

be seen. At first glance, a thin-silicon SOI bipolar transistor appears more complex to design than a regular vertical bipolar transistor because of its two-dimensional current flow in the collector region and the dependence of its characteristics on silicon thickness and substrate voltage. However, this additional complexity also implies that a thin-silicon SOI bipolar transistor has additional degrees of freedom in design optimization. For instance, if the transistor is built in a thin-silicon SOI BiCMOS process, the silicon thickness is likely to be fixed by the CMOS requirements and hence may not be tunable. But the bipolar device characteristics can still be tuned by varying L_{CE} and N_C. If a contact to the SOI substrate is available, either as a common substrate contact for the entire chip or as an isolated contact for one or more transistors, a substrate bias can be used to modulate the device characteristics as well. As shown in Fig. 8.19, regular vertical SiGe-base bipolar transistors can have very high performance, but their BV_{CEO} values are

rather limited. It is quite likely that the additional freedom in the collector design of a thin-silicon SOI SiGe-base bipolar will enable SiGe-base bipolar transistors to have high performance and large BV_{CEO} values as well. Perhaps the most interesting feature of thin-silicon SOI bipolar technology is in the potential for integrating complementary SiGe-base bipolar and SOI CMOS on the same chip (Ning, 2003b; Duvernay *et al.*, 2007).

10.3 Double-Gate MOSFETs

It was discussed in Section 10.1 that fully depleted thin film SOI MOSFETs do not scale well to short channel lengths for lack of a conducting plane under the silicon region. If a conducting film is placed under the device, it needs to be only a thin insulator away from the silicon film to be effective. But in such a "ground-plane" configuration, the capacitance between the drain and the bottom conductor would be excessively high, resulting in poor switching performance. Ideally, a second gate placed under the silicon film, with minimal overlap to the source and drain regions, can serve as the bottom conductor. This forms a double gate (DG) MOSFET (Frank *et al.*, 1992). In the symmetric case, both gates have identical work function. When they switch together, two inversion channels are formed: one at the top of the silicon film near the top gate insulator, and the other at the bottom near the bottom insulator. The silicon film is usually lightly doped and fully depleted. This allows higher mobilities and avoids floating body effect associated with partially depleted bodies. Furthermore, since no doping is needed to confine the depletion region, discrete dopant number fluctuation is not a problem in DG MOSFETs. The threshold voltage is completely determined by the gate work function. Although this helps uniformity across a wafer, it could be a shortcoming when multiple threshold voltages are needed. Note that as far as the intrinsic delay is concerned, there is no inherent advantage of a symmetric double-gate MOSFET over a single-gate MOSFET of the same channel length because both the current and the inversion-charge capacitance (and the gate fringe capacitance for that matter) are doubled in the double-gate case. The performance advantage lies mainly in the ability of DG MOSFETs to scale to a shorter channel length.

DG CMOS is conceived as an ideal device structure that could in principle extend scaling beyond the limit of bulk CMOS. Section 10.3.1 describes an analytic potential model for the drain current of symmetric DG MOSFETs. In Section 10.3.2, the scale length and the scaling limit of DG MOSFETs are examined. Section 10.3.3 discusses the fabrication requirements and challenges of a DG MOSFET. Section 10.3.4 generalizes the concept to multiple-gate MOSFETs, in particular, a surrounding-gate or gate-all-around or nanowire MOSFET for which an analytic potential current model and a scale length model are also available.

10.3.1 An Analytic Drain Current Model for Symmetric DG MOSFETs

In Section 3.1.1, the drain current of a long-channel bulk MOSFET is expressed in terms of Pao-Sah's integral after combining 1-D Poisson's and current continuity equations under the gradual channel approximation. While Pao-Sah's integral uses the channel

quasi-Fermi potential and contains both the drift and diffusion current components hence is valid under all regions of MOSFET operation (subthreshold, linear, saturation), its mathematical form renders no general analytic solution. This had necessitated the charge sheet approximation which had led to further simplification into separate analytic current expressions for different bias regions. Such piecewise current solutions, while physical, would cause numerical convergence problems in circuit modeling applications. Furthermore, the charge sheet concept is physically inconsistent with a particular DG MOSFET operation mode in which "volume inversion," i.e., inversion of the entire silicon film in unison, takes place.

Fortunately, owing to the absence of the depletion charge term, 1-D Poisson's equation for an undoped DG MOS device is analytically integratable to yield a closed-form solution to the potential everywhere in the silicon film (Taur, 2000). By extending this approach, a continuous, analytic I–V model for double-gate MOSFETs can be derived directly from Pao-Sah's integral without the charge sheet approximation (Taur *et al.*, 2004). This analytic solution covers all regions of MOSFET operation, thus maintaining strong continuity while retaining the essential device physics.

Consider an undoped (or lightly doped), symmetric double-gate MOSFET, shown schematically in Fig. 10.10. The same voltage is applied to the two gates having the same work function. The band diagrams along a cut perpendicular to the silicon film and the two gates are shown in Fig. 10.11 for two gate voltages. At zero gate voltage below threshold, the bands are essentially flat throughout the silicon film as well as in the gate insulators because both the depletion charge and the inversion charge are negligible. ***Since there is no contact to the silicon body, it does not matter where the Fermi level of the lightly doped body exactly is in the band gap. The silicon bands will just float to the position dictated by the gate work function*** as shown. The energy levels are referenced to the electron Fermi level or the conduction band of the n^+ source, represented by the long

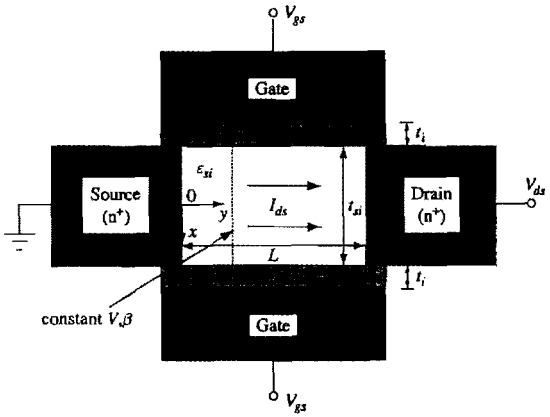


Figure 10.10. Schematic diagram of a double-gate MOSFET. $V(y)$ is the quasi-Fermi potential at a point in the channel. $V(0)=0$ at the source and $V(L)=V_{ds}$ at the drain. β is a function of V.

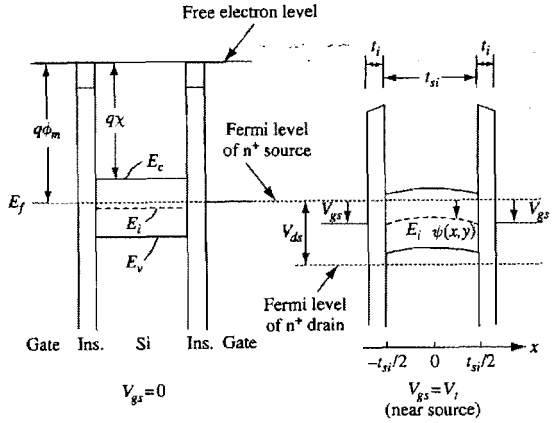


Figure 10.11. Band diagram along a vertical cut in Fig. 10.10 for two gate voltages. The gate work function in this example is slightly lower than that of intrinsic silicon. For $V_{gs}=V_t$ on the right, the cut is near the source where $V=0$ (Taur *et al.*, 2004).

dotted line in Fig. 10.11. As the gate voltage increases toward the threshold voltage in the right-hand part of Fig. 10.11, mobile charge or electron density becomes appreciable when the conduction band of the silicon body moves to near the Fermi level of the source.

Following Pao-Sah's gradual channel approach in Sections 3.1.1.1 and 3.1.1.2, Poisson's equation along a vertical cut perpendicular to the silicon film (Fig. 10.10) takes the following form with only the mobile charge (electrons) term:

$$\frac{d^2\psi}{dx^2} = \frac{q}{\varepsilon_{si}} n_i e^{q(\psi - V)/kT}, \tag{10.5}$$

where $\psi(x,y)$ is the electrostatic potential, defined as the intrinsic potential at (x,y) with respect to the Fermi potential at the source, as shown in Fig. 10.11. $V(y)$ is the electron quasi-Fermi potential at y with respect to that of the source. Here we consider an nMOSFET with $q\psi/kT >> 1$ so that the hole density is negligible.

Since the current flows predominantly from the source to the drain along the y-direction, the gradient of the electron quasi-Fermi potential is also in the y-direction. This justifies the gradual channel approximation that V is constant in the x-direction. Equation (10.5) can then be integrated twice to yield the solution

$$\psi(x) = V - \frac{2kT}{q} \ln\left[\frac{t_{si}}{2\beta}\sqrt{\frac{q^2 n_i}{2\varepsilon_{si}kT}}\cos\left(\frac{2\beta x}{t_{si}}\right)\right], \tag{10.6}$$

where β is a constant (independent of x) to be determined from the boundary condition

$$\varepsilon_i \frac{V_{gs} - \Delta\phi - \psi(x = \pm t_{si}/2)}{t_i} = \pm\varepsilon_{si}\left.\frac{d\psi}{dx}\right|_{x=\pm t_{si}/2}. \tag{10.7}$$

Here V_{gs} is the voltage applied to both gates, and $\Delta\phi$ is the work function of both the top and bottom gate electrodes with respect to that of the intrinsic silicon. In other words, $\Delta\phi = \phi_m - (\chi + E_g/2q)$. Substituting Eq. (10.6) into Eq. (10.7) leads to

$$\frac{q(V_{gs} - \Delta\phi - V)}{2kT} - \ln\left(\frac{2}{t_{si}}\sqrt{\frac{2\varepsilon_{si}kT}{q^2 n_i}}\right) = \ln\beta - \ln(\cos\beta) + \frac{2\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta\tan\beta. \tag{10.8}$$

For a given V_{gs}, β can be solved from the implicit equation (10.8) as a function of V. Along the channel direction (y), V varies from the source to the drain. So does β. The functional dependence of $V(y)$ and $\beta(y)$ is determined by the current continuity condition which requires the current $I_{ds} = \mu W Q_i dV/dy$ = constant, independent of V or y. Here μ is the effective mobility, W is the device width, and Q_i is the total mobile charge per unit area including both channels. Integrating $I_{ds}dy$ from the source to the drain and expressing dV/dy as $(dV/d\beta)(d\beta/dy)$, Pao-Sah's integral can be written as

$$I_{ds} = \mu\frac{W}{L}\int_0^{V_{ds}} Q_i(V)dV = \mu\frac{W}{L}\int_{\beta_s}^{\beta_d} Q_i(\beta)\frac{dV}{d\beta}d\beta, \tag{10.9}$$

where β_s, β_d are solutions to Eq. (10.8) corresponding to $V=0$ and $V=V_{ds}$ respectively. From Gauss's law, $Q_i = 2\varepsilon_{si}(d\psi/dx)_{x=t_{si}/2}$, which equals $2\varepsilon_{si}(2kT/q)(2\beta/t_{si})\tan\beta$ using Eq. (10.6). Note that $dV/d\beta$ can also be expressed as a function of β by differentiating Eq. (10.8). Substituting these factors in Eq. (10.9) and carrying out the integration analytically yield the drain current:

$$\begin{aligned} I_{ds} &= \mu\frac{W}{L}\frac{4\varepsilon_{si}}{t_{si}}\left(\frac{2kT}{q}\right)^2\int_{\beta_d}^{\beta_s}\left[\tan\beta + \beta\tan^2\beta + \frac{2\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta\tan\beta\frac{d}{d\beta}(\beta\tan\beta)\right]d\beta \\ &= \mu\frac{W}{L}\frac{4\varepsilon_{si}}{t_{si}}\left(\frac{2kT}{q}\right)^2\left[\beta\tan\beta - \frac{\beta^2}{2} + \frac{\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta^2\tan^2\beta\right]\Bigg|_{\beta_d}^{\beta_s}. \end{aligned} \tag{10.10}$$

The range of β_s, β_d is (0, π/2). ***MOSFET characteristics for all regions: linear, saturation, and subthreshold, can be generated from this continuous, analytic solution*** (Taur *et al.*, 2004). For example, in the linear region above threshold, the LHS of Eq. (10.8)>> 1 for both $V=0$ and V_{ds}, so β_s, $\beta_d \sim \pi/2$. The last terms on the RHS of Eqs. (10.8) and (10.10) dominate, therefore

$$I_{ds} = \mu\frac{\varepsilon_i}{t_i}\frac{W}{L}\left[(V_{gs} - V_t)^2 - (V_{gs} - V_t - V_{ds})^2\right] = 2\mu\frac{\varepsilon_i}{t_i}\frac{W}{L}(V_{gs} - V_t - V_{ds}/2)V_{ds}, \tag{10.11}$$

where

$$V_t = \Delta\phi + \frac{2kT}{q}\ln\left[\frac{2}{t_{si}}\sqrt{\frac{2\varepsilon_{si}kT}{q^2 n_i}}\right] + \frac{2kT}{q}\ln\left[\frac{q\varepsilon_i t_{si}(V_{gs} - \Delta\phi)}{2\pi\varepsilon_{si}t_i kT}\right]. \tag{10.12}$$

The last term is a second-order term coming from the ln[cos β] term in Eq. (10.8). It is kept here to show that the t_{si} factor cancels the $1/t_{si}$ factor in the previous term so V_t is independent of t_{si}.

In the saturation region, where $\beta_s \sim \pi/2$, but $\beta_d << 1$, one obtains

$$I_{ds} = \mu\frac{\varepsilon_i}{t_i}\frac{W}{L}\left[(V_{gs} - V_t)^2 - \frac{8\varepsilon_{si}t_i k^2T^2}{\varepsilon_i t_{si}q^2}e^{q(V_{gs} - V_t - V_{ds})/kT}\right]. \tag{10.13}$$

Note that in this continuous model, the current approaches the saturation value with a difference term exponentially decreasing with V_{ds}, in contrast to the common piecewise models in which the current is made to be constant in saturation.

In the subthreshold region, both β_s, $\beta_d << 1$, so the lnβ term dominates on the RHS of Eq. (10.8), and

$$I_{ds} = \mu\frac{W}{L}kTn_i t_{si}e^{q(V_{gs} - \Delta\phi)/kT}\left(1 - e^{-qV_{ds}/kT}\right), \tag{10.14}$$

as would be expected from the basic diffusion current, $J_{diff} = qDdn/dx$. Note that the subthreshold current is proportional to the silicon thickness, but independent of ε_i/t_i – a manifestation of "volume inversion" that the potential in silicon is near constant and directly follows the gate modulation when the mobile carrier density is low. In contrast, the currents above threshold, Eqs. (10.11) and (10.13), are proportional to ε_i/t_i, but independent of silicon thickness. This is the same as in a bulk MOSFET in that once it becomes energetically favorable for a large electron population in the conduction band, essentially all the gate-induced mobile charge appears at the surface, which electrostatically screens the interior of silicon from the gate.

While doping in the silicon film is neglected in the above analytic potential model, its first-order effect can be incorporated simply as a threshold voltage shift of magnitude $qN_a t_{si}/(2C_{ox})$ due to the depletion charge in silicon. P-type dopants shift the threshold positively and n-type negatively. It has been shown that too high a doping level has a negative impact on the device characteristics (Lu *et al.*, 2008).

10.3.2 The Scale Length of Double-Gate MOSFETs

How short a gate or channel length a DG MOSFET can be scaled to is a function of the silicon thickness, the gate insulator thickness, and their dielectric constants. The scale length can be obtained from the three-region model derived in Appendix 10. Referring to the DG device structure in Fig. 10.10, we replace ε_1, t_1 and ε_3, t_3 with ε_i, t_i, and ε_2, t_2 with ε_{si}, t_{si} in Eq. (A10.13) to obtain

$$\frac{\varepsilon_{si}}{\varepsilon_i^2}\tan^2\left(\frac{\pi t_i}{\lambda}\right)\tan\left(\frac{\pi t_{si}}{\lambda}\right) = \frac{2}{\varepsilon_i}\tan\left(\frac{\pi t_i}{\lambda}\right) + \frac{1}{\varepsilon_{si}}\tan\left(\frac{\pi t_{si}}{\lambda}\right). \tag{10.15}$$

Multiplying Eq. (10.15) by ε_{si} results in a quadratic equation in $(\varepsilon_{si}/\varepsilon_i)\tan(\pi t_i/\lambda)$ that can be solved in terms of $\tan(\pi t_{si}/\lambda)$. After further simplification using trigonometric identities, the scale length equation for a DG MOSFET takes the form of [2]

[2] There is a second eigenvalue equation that covers the asymmetric eigenfunctions in the case of asymmetric DG devices (Liang and Taur, 2004). But the lowest order (longest) solution of the scale length λ comes from Eq. (10.16) for symmetric eigenfunctions.

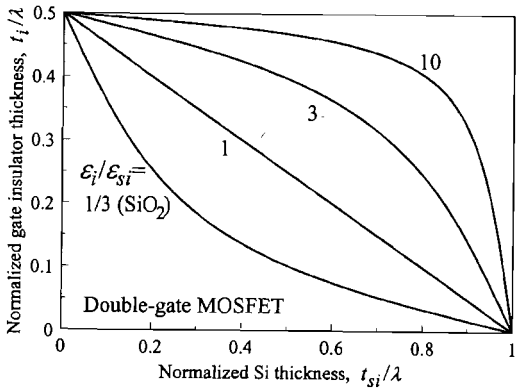


Figure 10.12. Numerical solutions to Eq. (10.16) for different values of $\varepsilon_i/\varepsilon_{si}$.

$$\tan\left(\frac{\pi t_i}{\lambda}\right)\tan\left(\frac{\pi t_{si}}{2\lambda}\right)=\frac{\varepsilon_i}{\varepsilon_{si}}. \tag{10.16}$$

The numerical solutions for the longest (lowest order) λ are plotted in Fig. 10.12 in normalized units with $\varepsilon_i/\varepsilon_{si}$ as a parameter. In the most straightforward case where $\varepsilon_i=\varepsilon_{si}$, the scale length is simply $\lambda=t_{si}+2t_i$, the physical height of the region between the two gates in Fig. 10.10. For $\varepsilon_i/\varepsilon_{si}<1$, $\lambda>t_{si}+2t_i$, and for $\varepsilon_i/\varepsilon_{si}>1$, $\lambda<t_{si}+2t_i$. But in no case can λ be smaller than t_{si} or $2t_i$, whichever is larger. This means that even extremely high-κ insulators would have to be physically thin to be useful. In the very high-κ limit, the scale length approaches $2t_i$, insensitive to the silicon thickness as long as $t_{si}<2t_i$. In the lower right corner of the curves where silicon is thick and insulator is thin, Eq. (10.16) can be approximated to $\lambda=t_{si}+2(\varepsilon_{si}/\varepsilon_i)t_i$, since $\pi t_i/\lambda<<1$ and $\pi t_{si}/2\lambda\approx\pi/2$. High-$\kappa$ gate dielectric is even more beneficial to DG than to bulk MOSFETs because of the double thickness. In the opposite, upper left corner, $\pi t_{si}/2\lambda<<1$ and $\pi t_i/\lambda\approx\pi/2$, and Eq. (10.16) can be approximated to $\lambda=2t_i+(\varepsilon_{si}/\varepsilon_i)t_{si}$. Once the scale length is determined, ***the scaling limit is given by*** $L_{min}\approx 1.5\lambda$ since the short-channel effect is proportional to $e^{-\pi L/2\lambda}$, just like in bulk MOSFETs.

Given that the high-κ gate insulator thickness t_i is limited by tunneling to ~2.5 nm (Section 4.2.2.1) and t_{si} is limited by threshold shift from quantum confinement to ~2 nm (Fig. 10.4), the minimum scale length of a DG MOSFET would be limited to 5 nm or so, implicating a channel length limit of 5–10 nm.

10.3.3 Fabrication Requirements and Challenges of DG MOSFETs

It is extremely challenging to find a manufacturable process for DG MOSFETs. Both gates must be self-aligned to the source–drain regions in the silicon film. Any misalignment could cause either excessive gate-to-drain overlap capacitance or underlapping in which an ungated gap would add large series resistance to the channel. The silicon film must be very thin (5–10 nm) in order for the scale length λ to go below 10 nm. Atomic

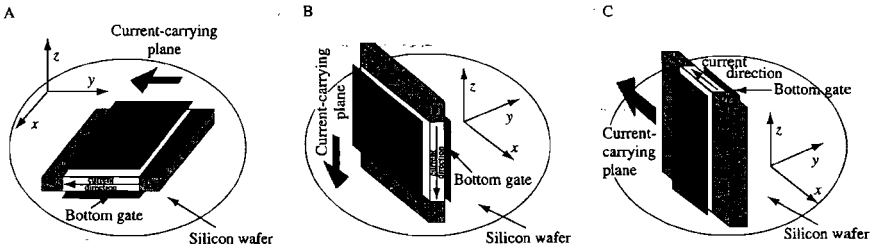


Figure 10.13. Three different orientations of double-gate MOSFETs (Wong *et al.*, 1999).

dimension tolerances would be required for the control of silicon film thickness. To reduce parasitic series resistance, source and drain fan-outs (raised source–drain) are needed to conduct current away from the thin channel region. The fan-out geometry may also be required for thermal reasons. Heat generated from the high drive current of the DG MOSFET is conducted away from the ~1 nm size hot spot mainly through the silicon film since any surrounding insulators have much poorer thermal conductivity. Ideally, such source–drain fan-out should also be self-aligned to the two gates. Finally, the threshold voltage of a symmetric DG MOSFET is only controlled by the gate work function. The standard n^+ silicon gate for nMOS and p^+ silicon gate for pMOS give near zero threshold voltages whereas a common midgap work function gate results in too high (magnitude) threshold voltages for both devices. New gate materials of work functions between the midgap and n^+ silicon and between the midgap and p^+ silicon need to be developed. Furthermore, modern CMOS technologies offer multiple threshold voltages for both n- and pMOSFETs in order for the circuit designers to deal with the tradeoff between off-current and performance discussed in Chapter 5. Possible ways to implement multiple threshold voltages include tunable (by implantation) work function gates and doping of the silicon film for limited adjustment (e.g., ±0.1 V) of threshold voltage.

Figure 10.13 shows the three possible orientations of a DG MOSFET on a silicon wafer (Wong *et al.*, 1999). In type A configuration, both the gates and the silicon film are parallel to the wafer plane. In terms of device and circuit layout, this geometry is most compatible to the conventional planar bulk CMOS technology. The key challenge is how to place the bottom gate under the active silicon film while self aligned to the top gate and source–drain regions. Attempts have been made to grow an epitaxial silicon film through a tunnel pre-formed between the sacrificial gates (Wong *et al.*, 1997). But the defect density turned out to be too high in the selective epi, especially near the corners. In type B configuration, both the gates and the silicon film are perpendicular to the wafer plane. This forms a vertical MOSFET with the two gates on both sides of the silicon film. It is easier to achieve self-alignment between the two gates. However, as discussed in Section 10.3.2, in order to scale the gate length of a DG MOSFET to L_{min}, a silicon film thickness of $\leq L_{min}/2$ is needed. This means that to scale below the bulk CMOS limit of 20 nm gate length, the silicon thickness must be ≤ 10 nm. It poses a severe challenge to lithographically pattern such a thin film edge on with a thickness tolerance on the order of 1 nm. In type C configuration, the gates and the silicon film are also perpendicular to the

wafer plane, but the current direction is in the wafer plane. This geometry is often referred to as FinFET (Hisamoto *et al.*, 2000). Conventional ion implantation process can be employed to form the source and drain regions on the same plane. The MOSFET width is the fin height in the vertical direction. Since the fin height is limited by structural stability, multiple fins in parallel are needed to build up the current drive. With the same edge-on film orientation, patterning of the film thickness is also a key issue as in configuration B.

10.3.4 Multiple-Gate MOSFETs

Driven by specific processes, other multiple-gate MOSFET structures have been investigated experimentally. These include triple gate, pi-gate, omega-gate, quadruple-gate, and surrounding gate, as shown schematically in Fig. 10.14 (Colinge, 2004).

With its cylindrical symmetry, an analytic long-channel *I–V* model similar to the planar DG MOSFET in Section 10.3.1 has also been developed for the surrounding-gate or nanowire MOSFET (Jiménez *et al.*, 2004). The result is

$$I_{ds} = \mu \frac{4\pi\varepsilon_{si}}{L}\left(\frac{2kT}{q}\right)^2\left[\frac{\varepsilon_{si}}{\varepsilon_i}\left(\frac{1-2\beta}{\beta^2}\right)\ln\left(1+\frac{t_i}{r_{si}}\right)+\frac{1}{\beta}+\frac{\ln\beta}{2}\right]\Bigg|_{\beta_d}^{\beta_s}, \tag{10.17}$$

where r_{si} is the silicon radius. The rest of the parameters are the same as in Eq. (10.10). β_s and β_d are solutions to the implicit equation,

$$\frac{q(V_{gs}-\Delta\phi-V)}{kT}-\ln\left(\frac{8\varepsilon_{si}kT}{q^2n_ir_{si}^2}\right)=\ln(1-\beta)-\ln(\beta^2)+\frac{4\varepsilon_{si}}{\varepsilon_i}\frac{1-\beta}{\beta}\ln\left(1+\frac{t_i}{r_{si}}\right), \tag{10.18}$$

for $V=0$ and V_{ds}, respectively.

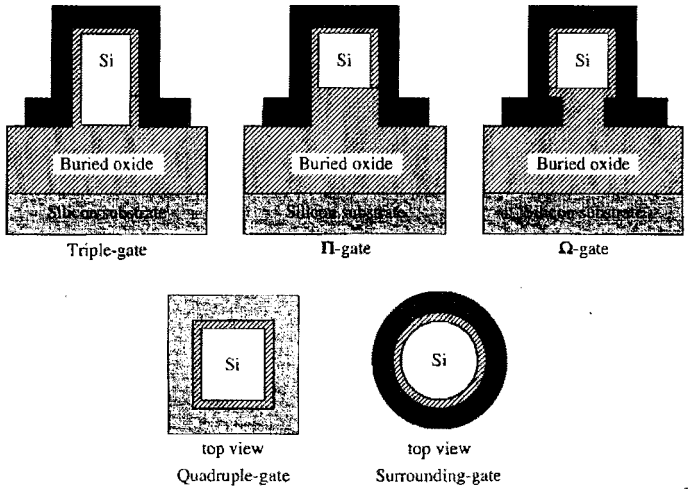


Figure 10.14. Various types of multiple-gate MOSFETs (Yu *et al.*, 2008).

The scale length equation for surrounding-gate MOSFETs (Oh *et al.*, 2000) has been derived from 2-D Poisson's equation in cylindrical coordinates following a similar approach as in Appendix 10,

$$\frac{Y_0'(\pi r_{si}/\lambda)}{J_0'(\pi r_{si}/\lambda)}=\frac{\varepsilon_{si}}{\varepsilon_i}\frac{Y_0(\pi r_{si}/\lambda)}{J_0(\pi r_{si}/\lambda)}+\left(1-\frac{\varepsilon_{si}}{\varepsilon_i}\right)\frac{Y_0[\pi(r_{si}+t_i)/\lambda]}{J_0[\pi(r_{si}+t_i)/\lambda]}, \tag{10.19}$$

where J_0 and Y_0 are Bessel and Neumann functions of the zeroth order and J_0', Y_0' their derivatives. The ultimate scaling limits of all these multi-gate structures are set by quantum mechanical considerations of silicon and insulator thickness limits as discussed in Section 10.3.2.

Exercise

10.1 Consider a symmetric double-gate nMOSFET with $t_{ox}=2$ nm and $t_{si}=10$ nm. The gate work functions are $q\phi_m = 4.33$ eV (half way between n^+ Si and intrinsic Si).
(a) Plot I_{ds} versus V_{ds} for the range of $0 < V_{ds} < 2$ V, with $V_{gs} = 0, 0.5, 1, 1.5, 2$ V. Assume an effective mobility of 200 cm^2/V-s, $W = 1$ μm and $L = 1$ μm.
(b) Derive expressions for g_m and g_{ds} (defined in Exercise 4.8) as functions of β_s and β_d.
(c) What channel length can this device be scaled to without severe short-channel effects?

Appendix 1 CMOS Process Flow

A more or less generic CMOS process flow is described below. It features shallow trench isolation (STI) (Davari *et al.*, 1988, 1989), dual n^+/p^+ polysilicon gates (Wong *et al.*, 1988; Sun *et al.*, 1989), and self-aligned silicide (Ting *et al.*, 1982). The front-end-of-the-line process consists of six or seven masking levels and is suitable for sub-0.5-μm generations of VLSI logic and SRAM technology.

- Starting material p-type substrate or p^- epi on p^+ substrate for latch-up prevention (Taur *et al.*, 1984).
- Grow pad oxide. Deposit CVD (Chemical Vapor Deposition) nitride. (See Fig. A1.1.)
- Lithography to cover the active region with photoresist.
- Reactive ion etching (RIE) nitride and oxide in the field region.
- RIE shallow trench in silicon. (See Fig. A1.2.)
- Grow pad oxide. Deposit thick CVD oxide. (See Fig. A1.3.)
- Chemical–mechanical polishing planarization. (See Fig. A1.4.)
- n-well lithography and implant (also channel doping).
- p-well lithography and implant (also channel doping). (See Fig. A1.5.)
- Grow gate oxide.
- Deposit polysilicon film. (See Fig. A1.6.)
- Gate lithography.
- RIE polysilicon gate. (See Fig. A1.7.)
- Sidewall reoxidation.
- n^+ source–drain lithography and implant (also dope n^+ polysilicon gate).
- p^+ source–drain lithography and implant (also dope p^+ polysilicon gate). (See Fig. A1.8.)
- Oxide (or nitride) spacer formation by CVD and RIE.
- Source–drain anneal. (See Fig. A1.9.)
- Self-aligned silicide process. (See Fig. A1.10.)
- Back-end-of-the-line process.

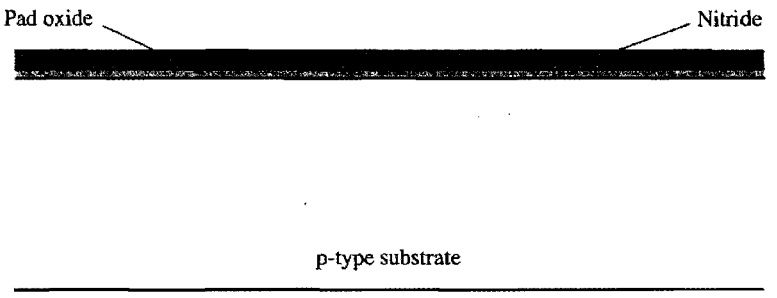


Figure A1.1.

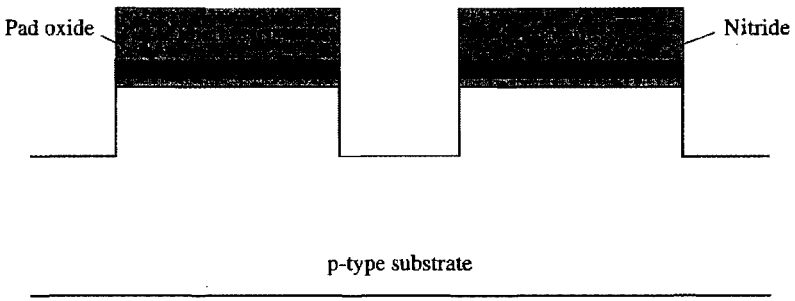


Figure A1.2.

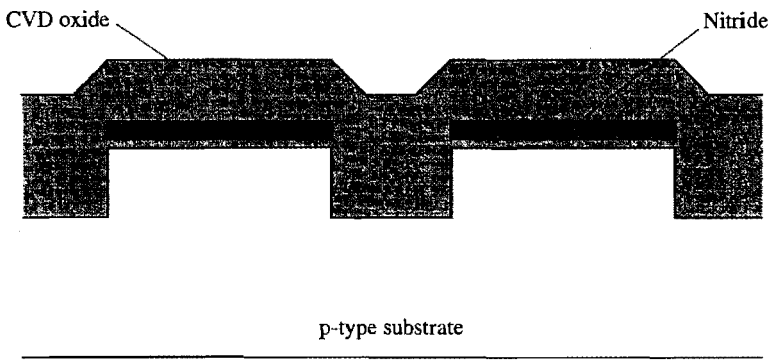


Figure A1.3.

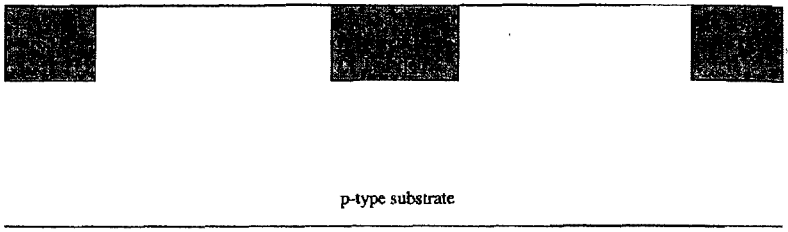


Figure A1.4.

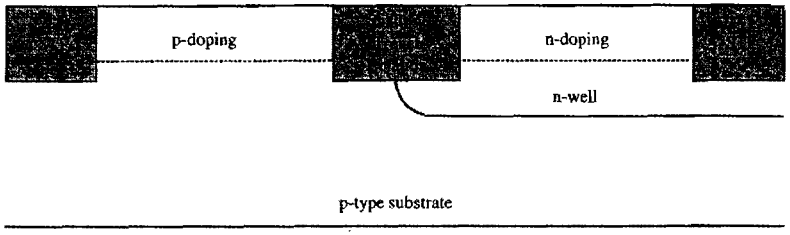


Figure A1.5.

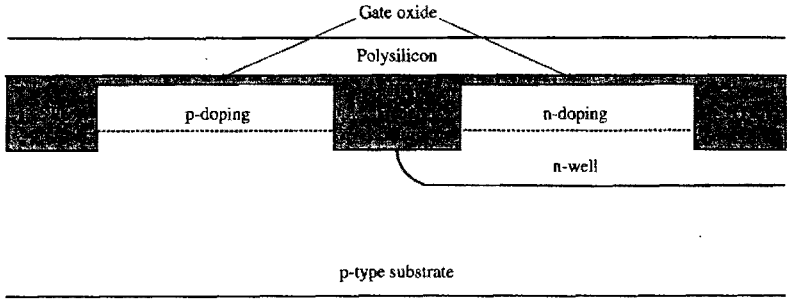
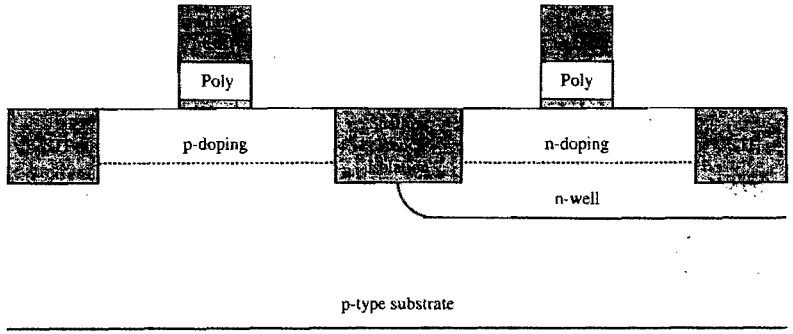


Figure A1.6.



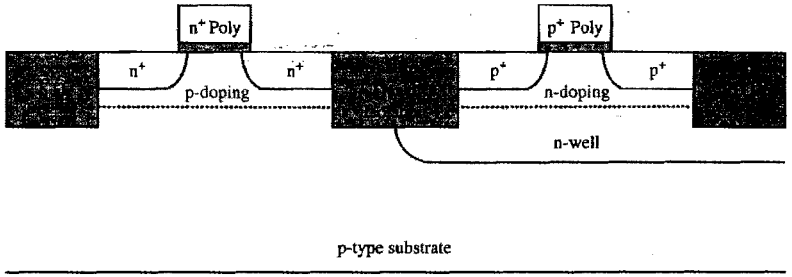


Figure A1.8.

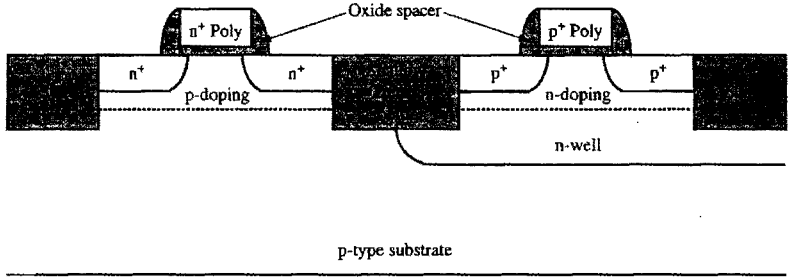


Figure A1.9.

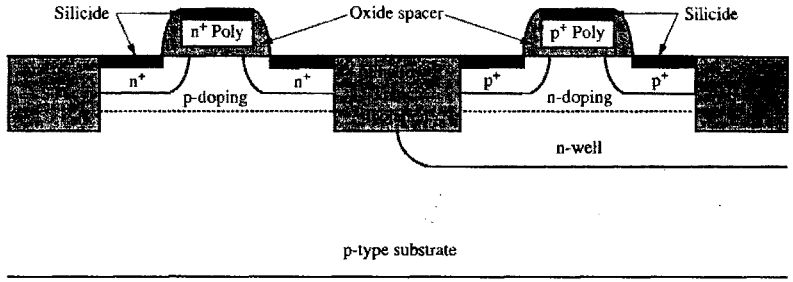
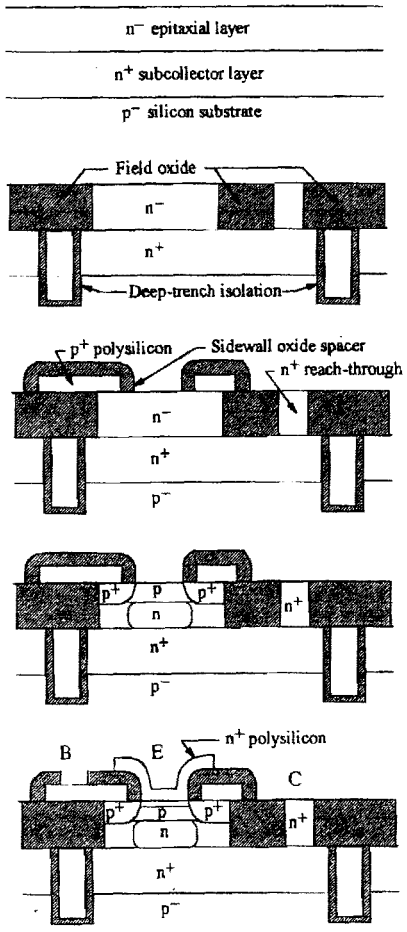


Figure A1.10.

Appendix 2 Outline of a Process for Fabricating Modern n–p–n Bipolar Transistors



- Starting wafer: p^- silicon
- n^+ subcollector layer
- n^- silicon epitaxial layer

- Polysilicon-filled deep-trench isolation
- Shallow-trench field oxide

- n^+ reach-through
- p^+ polysilicon layer for base contact
- Sidewall oxide spacer

- Diffusion from p^+ polysilicon
- n-type pedestal collector
- p-type intrinsic base

- n^+ polysilicon layer for emitter
- Contact windows
- Metal (not shown)

Figure A2.1.

Appendix 3 Einstein Relations

It was stated in Section 2.1.3 that carrier motion in silicon consists of *drift* in the presence of an electric field and *diffusion* in the presence of a concentration gradient. Drift is characterized by the *mobility* defined in Eq. (2.26), and diffusion by the *diffusion coefficient* defined in Eqs. (2.36) and (2.37). Since both mechanisms are closely tied to the random thermal motion of electrons (or holes), the diffusion coefficient and the mobility are related by the Einstein relations, Eqs. (2.38) and (2.39). In this appendix, we briefly describe the physical picture of the drift and diffusion processes, leading to the basic concept behind the Einstein relations. Note that MKS units are used throughout this appendix (i.e., length must be in meters, not centimeters).

A3.1 Drift

Under thermal equilibrium, electrons possess an average kinetic energy proportional to kT. They move in random directions through the silicon crystal with an average thermal velocity v_{th}. At room temperature, v_{th} is of the order of 10^7 cm/s. Electrons scatter frequently with the lattice (phonons) and ionized impurity atoms. The average distance electrons travel between collisions is called the *mean free path* l, and the average time between collisions is called the *mean free time* $\tau = l/v_{th}$. Typically, $l \approx 10$ nm and $\tau \approx 0.1$ ps.

In the absence of electric field, the net velocity of electrons in any particular direction is zero, since the thermal motion is completely random. When a small electric field $\mathscr{E}$ is applied, however, electrons are accelerated in the direction opposite to the field during the time between collisions. If the effective mass of electrons is m^*, the acceleration is given by $-q\mathscr{E}/m^*$. The drift velocity the electrons gain during a mean free time τ between collisions is thus

$$v_d = -q\mathscr{E}\tau/m^*. \tag{A3.1}$$

After a collision event, electron velocities become randomized again, which means that the average drift velocity is reset to zero and the process starts over again. Since the mobility μ is defined by Eq. (2.26) as the proportionality of the drift velocity to the electric field, one obtains

$$\mu = \frac{q\tau}{m^*} = \frac{ql}{m^* v_{th}}. \tag{A3.2}$$

The last step follows from $l = v_{th}\tau$. The drift current density is given by

$$J_{drift} = qnv_d = qn\mu\mathscr{E}, \tag{A3.3}$$

as indicated by Eq. (2.28).

A3.2 Diffusion

Diffusion current in silicon arises when there is a spatial variation of carrier concentration in the material, that is, the carriers tend to move from a region of high concentration to a region of low concentration. To illustrate the diffusion process, let us assume a one-dimensional case shown in Fig. A3.1, in which the electron density n varies in the x-direction (Muller and Kamins, 1977). We consider the number of electrons crossing the plane at x per unit area per unit time. The electrons are moving at thermal velocity v_{th} either in the left or in the right direction and are scattered each time after they travel, on the average, a distance equal to the mean free path l. Therefore, the electrons crossing the plane at x from the left in each collision time start at approximately $x - l$, i.e., one mean free path away on the left side of x. Since electrons have equal chances of moving left or right, half of the electrons at $x - l$ will move across the plane at x before the next collision takes place. The current density per unit area resulting from the motion of these carriers is then

$$J_- = \frac{1}{2}qn(x-l)v_{th}. \tag{A3.4}$$

This current is negative, since it consists of negative charge moving in the positive direction. Similarly, half of the electrons at $x + l$, one mean free path away on the right side of x, will move across the plane x from the right, resulting in a current density in the positive direction:

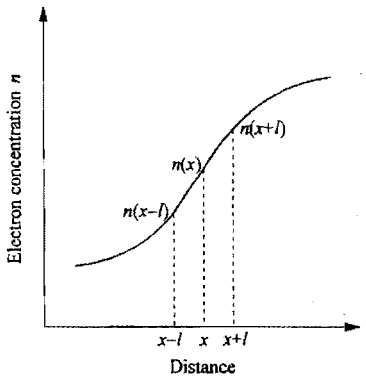


Figure A3.1. Schematic diagram of electron concentration versus distance in a one-dimensional case. (After Muller and Kamins, 1977.)

$$J_+ = \frac{1}{2}qn(x+l)v_{th}. \tag{A3.5}$$

The net diffusion current density at x is then

$$J_{diff} = J_+ - J_- = \frac{1}{2}qv_{th}[n(x+l) - n(x-l)]. \tag{A3.6}$$

Using a Taylor series expansion of $n(x + l)$ and $n(x - l)$, and keeping only the first-order terms, one obtains

$$J_{diff} = \frac{1}{2}qv_{th}\left[\left(n(x) + l\frac{dn}{dx}\right) - \left(n(x) - l\frac{dn}{dx}\right)\right] = qv_{th}l\frac{dn}{dx}. \tag{A3.7}$$

We thus see that the diffusion current density is proportional to the spatial derivative of the electron concentration. In other words, diffusion current arises because of the random thermal motion of charged carriers in a concentration gradient. Equation (A3.7) can be written in the form of Eq. (2.36) with the diffusion coefficient defined as

$$D \equiv v_{th}l. \tag{A3.8}$$

It is now straightforward to derive the Einstein relations. Taking the ratio of Eq. (A3.8) to Eq. (A3.2), one obtains

$$\frac{D}{\mu} = \frac{m^* v_{th}^2}{q}. \tag{A3.9}$$

From the theorem for the equipartition of energy in a one-dimensional case [see Exercise 2.3(a) for the proof of a three-dimensional case],

$$\frac{1}{2}m^* v_{th}^2 = \frac{1}{2}kT. \tag{A3.10}$$

Therefore,

$$\frac{D}{\mu} = \frac{kT}{q}, \tag{A3.11}$$

which is the Einstein relations Eqs. (2.38) and (2.39).

Appendix 4 Spatial Variation of Quasi-Fermi Potentials

In this appendix, we examine the spatial variation of the minority-carrier quasi-Fermi potentials in the quasineutral regions of a p–n diode in the low-injection approximation, i.e., when the injected minority-carrier densities are small compared to the majority-carrier densities. We also examine the change of the quasi-Fermi potentials across the space-charge region as a function of the applied voltage. A detailed model for the spatial variation of the quasi-Fermi potentials in a symmetrical step p–n junction can be found in a paper by Sah (1966).

A4.1 Spatial Variation of Minority-Carrier Quasi-Fermi Potentials

Specially, we want to examine the spatial variation of $\phi_n(x)$ in the p-type region of an n^+–p diode illustrated schematically in Fig. A4.1. For simplicity, we assume the p-type base has a uniform doping concentration N_a. The p-type quasineutral region is assumed to extend from the space-charge-region edge at $x=0$ to the p-contact at $x=W_p$. At the p-contact, the electron density is equal to the equilibrium minority-electron density n_{p0}, i.e., $n_p(W_p)=n_{p0}$.

The transport of electrons in the p-type region is governed by the diffusion equation [see Eq. (2.115)]

$$\frac{d^2n_p(x)}{dx^2}-\frac{n_p(x)-n_{p0}}{L_n^2}=0, \tag{A4.1}$$

where

$$L_n\equiv\sqrt{\tau_n D_n}=\sqrt{\frac{kT\mu_n\tau_n}{q}} \tag{A4.2}$$

is the electron diffusion length in the p-region. Equation (A4.1) gives the minority-electron distribution as [see Eq. (2.119)]

$$n_p(x)-n_p(W_p)=[n_p(0)-n_p(W_p)]\frac{\sinh[(W_p-x)/L_n]}{\sinh(W_p/L_n)}. \tag{A4.3}$$

Equation (A4.3) applies to both a forward-biased diode, where $n_p(0)/n_p(W_p)>1$, and a reverse-biased diode, where $n_p(0)/n_p(W_p)<1$.

In the p-region, $\phi_n(x)$ is given by [see Eq. (2.65)]

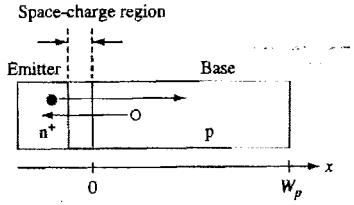


Figure A4.1. Schematic of an n^+–p diode.

$$\phi_n(x)=\psi_i(x)-\frac{kT}{q}\ln\left[\frac{n_p(x)}{n_i}\right], \tag{A4.4}$$

where $\psi_i(x)=-E_i/q$ is the electrostatic potential. Within the uniformly doped p-region, ψ_i is independent of x. Therefore,

$$\begin{aligned}\phi_n(x)-\phi_n(W_p)&=\psi_i(x)-\psi_i(W_p)-\frac{kT}{q}\ln\left[\frac{n_p(x)}{n_i}\right]+\frac{kT}{q}\ln\left[\frac{n_p(W_p)}{n_i}\right]\\&=-\frac{kT}{q}\ln\left[\frac{n_p(x)}{n_p(W_p)}\right]\\&=-\frac{kT}{q}\ln\left\{\left[\frac{n_p(0)}{n_p(W_p)}-1\right]\frac{\sinh[(W_p-x)/L_n]}{\sinh(W_p/L_n)}+1\right\},\end{aligned} \tag{A4.5}$$

where we have used Eq. (A4.3) for the ratio $n_p(x)/n_p(W_p)$.

The diffusion current density in the p-region due to electrons is

$$\begin{aligned}J_n(x)&=qD_n\frac{dn_p(x)}{dx}\\&=-\frac{kT\mu_n}{L_n}[n_p(0)-n_p(W_p)]\frac{\cosh[(W_p-x)/L_n]}{\sinh(W_p/L_n)},\end{aligned} \tag{A4.6}$$

where we have used the Einstein relationship $D_n/\mu_n=kT/q$. Therefore, in the p-region we have

$$\begin{aligned}\frac{d\phi_n(x)}{dx}&=-\frac{J_n(x)}{qn_p(x)\mu_n}\\&=\frac{kT}{qL_n}\left\{\frac{\dfrac{\cosh[(W_p-x)/L_n]}{\sinh(W_p/L_n)}}{\dfrac{\sinh[(W_p-x)/L_n]}{\sinh(W_p/L_n)}+\dfrac{n_p(W_p)}{n_p(0)-n_p(W_p)}}\right\}.\end{aligned} \tag{A4.7}$$

Equation (A4.7) gives

$$\left.\frac{d\phi_n(x)}{dx}\right|_{x=0}=\frac{kT}{qL_n\tanh(W_p/L_n)}\left[1-\frac{n_p(W_p)}{n_p(0)}\right] \tag{A4.8}$$

in the p-region at the space-charge-layer boundary.

Figure A4.2 is a plot of $\phi_n(x)-\phi_n(W_p)$, i.e., Eq. (A4.5), in units of kT/q as a function of x/W_p for a wide ($W_p/L_n=10$) p-region, and Fig. A4.3 is a similar plot for a narrow

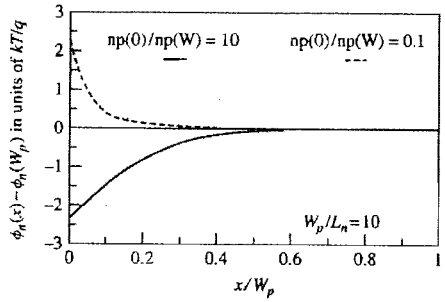


Figure A4.2. $\phi_n(x) - \phi_n(W_p)$ in units of kT/q as a function of x/W_p for $W_p/L_n = 10$.

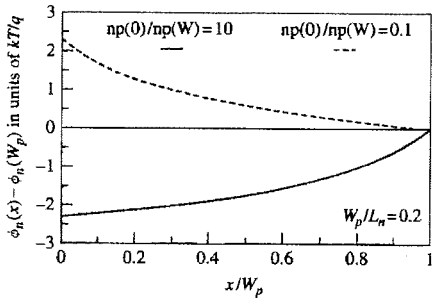


Figure A4.3. $\phi_n(x) - \phi_n(W_p)$ in units of kT/q as a function of x/W_p for $W_p/L_n = 0.2$.

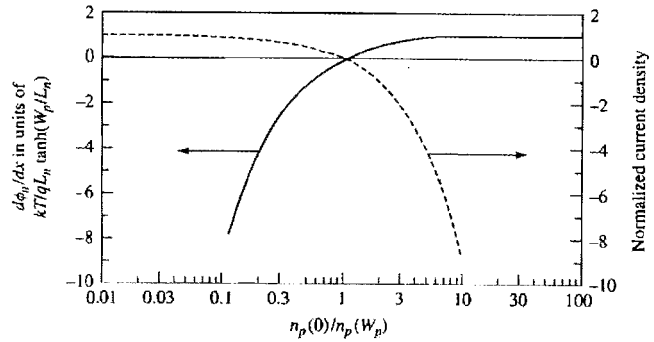


Figure A4.4. Plot of $d\phi_n(x)/dx$ at $x = 0$ in units of $[kT/qL_n \tanh(W_p/L_n)]$ and plot of the normalized electron current density, i.e., $J_n(x)$ divided by $qD_n n_p(W_p) \cosh(W_p - x)/L_n \sinh(W_p/L_n)$, as a function of the ratio $n_p(0)/n_p(W_p)$. $n_p(0)/n_p(W_p) > 1$ is for forward bias, and $n_p(0)/n_p(W_p) < 1$ is for reverse bias.

$(W_p/L_n = 0.2)$ p-region. The curves for $n_p(0)/n_p(W_p) = 10$ are for a forward bias while the curves for $n_p(0)/n_p(W_p) = 0.1$ are for a reverse bias.

Figure A4.4 is a plot of $d\phi_n(x)/dx$ at $x = 0$, i.e. Eq. (A4.8), in units of $kT/qL_n \tanh(W_p/L_n)$ as a function of the ratio $n_p(0)/n_p(W_p)$. Also plotted in the figure is the normalized current density from Eq. (A4.6). In forward bias, the magnitude of the current density increases as $n_p(0)/n_p(W_p)$ increases, but the magnitude of $d\phi_n(x)/dx$ saturates rapidly towards a value of $kT/qL_n \tanh(W_p/L_n)$ which is independent of $n_p(0)/n_p(W_p)$. However, in reverse bias, the opposite happens, with the magnitude of the current density saturating rapidly to a value independent of $n_p(0)/n_p(W_p)$ while the magnitude of $d\phi_n(x)/dx$ keeps on increasing as the ratio $n_p(0)/n_p(W_p)$ decreases.

A4.2 Change of the Quasi-Fermi Potentials Across the Space-Charge Region

In this section, we want to estimate the change of a quasi-Fermi potential across the space-charge region of a diode and its dependence on the diode bias voltage. The energy bands and quasi-Fermi potentials for a reverse-biased p–n diode are sketched schematically in Fig. A4.5, where W_d is width of the space-charge region. We want to evaluate the difference $\phi_n(-W_d) - \phi_n(0)$ as a function of V_{app}.

From Eqs. (2.73) and (2.81), we have

$$\Delta\psi_i \equiv \psi_i(-W_d) - \psi_i(0) = \psi_{bi} - V_{app} = \frac{kT}{q}\ln\left(\frac{N_d N_a}{n_i^2}\right) - V_{app}. \tag{A4.9}$$

If we ignore generation–recombination current, then the electron current density in the space-charge region is constant and can be obtained from Eq. (A4.6). It is

$$J_n = J_n(0) = -\frac{kT\mu_n}{L_n \tanh(W_p/L_n)}[n_p(0) - n_{p0}]. \tag{A4.10}$$

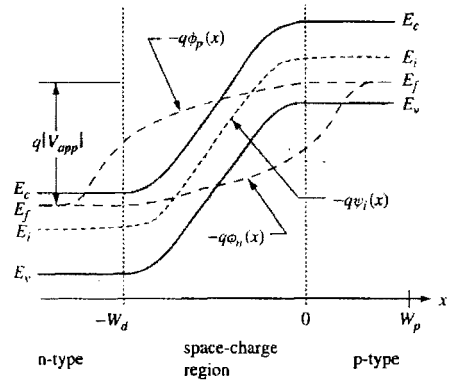


Figure A4.5. Schematic diagram illustrating the energy bands, the electrostatic potential and the quasi-Fermi potentials in a reverse-biased p–n diode.

Using the relationships [see Eqs. (2.63) and (2.65)]

$$J_n(x) = -q\mu_n n(x)\frac{d\phi_n(x)}{dx} \tag{A4.11}$$

and

$$n(x) = n_i e^{q[\psi_i(x)-\phi_n(x)]/kT}, \tag{A4.12}$$

we can write

$$e^{-q\phi_n(x)/k/T}d\phi_n = \frac{-J_n(x)}{q\mu_n n_i}e^{-q\psi_i(x)/kT}dx \tag{A4.13}$$

which can be integrated from $x=-W_d$ to $x=0$, where J_n is constant, to give

$$\begin{aligned} e^{-q\phi_n(0)/kT} - e^{-q\phi_n(-W_d)/kT} &= \frac{J_n}{kT\mu_n n_i}\int_{-W_d}^{0} e^{-q\psi_i(x)/kT}dx \\ &= -\frac{\delta[n_p(0)-n_{p0}]e^{-q\psi_i(0)/kT}}{n_i}, \end{aligned} \tag{A4.14}$$

where

$$\delta \equiv \frac{1}{L_n \tanh(W_p/L_n)}\int_{-W_d}^{0} e^{q[\psi_i(0)-\psi_i(x)]/kT}dx. \tag{A4.15}$$

Using Eq. (A4.12), Eq. (A4.14) can be rearranged to give

$$\begin{aligned} -\frac{\delta[n_p(0)-n_{p0}]}{n_i} &= e^{q[\psi_i(0)-\phi_n(0)]/kT} - e^{q[\psi_i(0)-\phi_n(-W_d)]/kT} \\ &= \frac{n_p(0)}{n_i} - \frac{N_d}{n_i}e^{-q\Delta\psi_i/kT}. \end{aligned} \tag{A4.16}$$

Equation (A4.9) gives

$$e^{-q\Delta\psi_i/kT} = \frac{n_i^2}{N_d N_a}e^{qV_{app}/kT} = \frac{n_{p0}}{N_d}e^{qV_{app}/kT}. \tag{A4.17}$$

Therefore, Eq. (A4.16) can be reduced to

$$\frac{n_p(0)}{n_{p0}} = \frac{e^{qV_{app}/kT}+\delta}{1+\delta}. \tag{A4.18}$$

Using Eqs. (A4.12), (A4.17) and (A4.18), we have

$$\begin{aligned} \phi_n(-W_d) - \phi_n(0) &= \Delta\psi_i - \frac{kT}{q}\ln\left[\frac{N_d}{n_p(0)}\right] \\ &= \frac{kT}{q}\ln\left[\frac{1+\delta e^{-qV_{app}/kT}}{1+\delta}\right]. \end{aligned} \tag{A4.19}$$

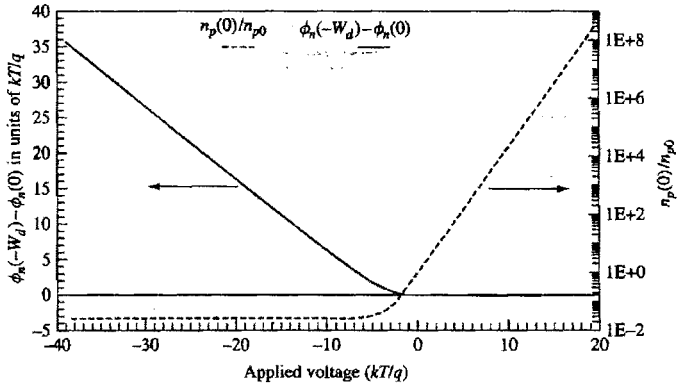


Figure A4.6. The change of $\phi_n(x)$ across the space-charge region, i.e., Eq. (A4.19), and the normalized minority-electron density from Eq. (A4.18) as a function of the applied voltage across an n^+–p diode. $V_{app}>$ is for forward bias and $V_{app}<0$ is for reverse bias. The assumptions are: $N_d = 1\times 10^{20}$ cm^{-3} for the n^+-region, and $N_a = 1\times 10^{18}$ cm^{-3}, $L_n = 20$ μm, and $W_p = 1$ μm for the p-region. In this plot, the forward bias is limited to $20kT/q$ to be consistent with the low-injection approximation used.

Equations (A4.15), (A4.18) and (A4.19) are applicable to forward bias ($V_{app}>0$) and reverse bias ($V_{app}<0$). The parameter δ can be obtained by integrating Eq. (A4.15), numerically if necessary, if we know the relationship among $\psi_i(x)$, W_d, and V_{app}. It can be inferred readily from Fig. A4.5 that $\psi_i(0) - \psi_i(x)$ is negative within the space-charge region so that the integral in Eq. (A4.15) is less than W_d. That is, $\delta < W_d/L_n\tanh(W_p/L_n)$.

As a specific example, let us consider an n^+–p diode with a uniformly doped p-region. In this case, the space-charge region resides primarily in the p-side (see Section 2.2.2.3), and the diode current is carried primarily by electrons injected from the n^+ emitter into the p-region in the case of forward bias or by the back diffusion of electrons from the p-region towards the n^+ emitter in the case of reverse bias (see Sections 2.2.4.3 and 2.2.4.4). The electrostatic potential in the space-charge region is governed by the Poisson equation [Eq. (2.44)]:

$$\begin{aligned} \frac{d\psi_i^2}{dx^2} &= \frac{q}{\varepsilon_{si}}[N_a + n(x)] \\ &\approx \frac{q}{\varepsilon_{si}}N_a, \end{aligned} \tag{A4.20}$$

where we have neglected the mobile electron density compared to the ionized acceptor density, which is valid in the low-injection approximation. Equation (A4.20) can be integrated twice to give

$$\psi_i(x) = \frac{qN_a}{2\varepsilon_{si}}x^2 + \psi_i(0). \tag{A4.21}$$

From Eq. (2.85), we have

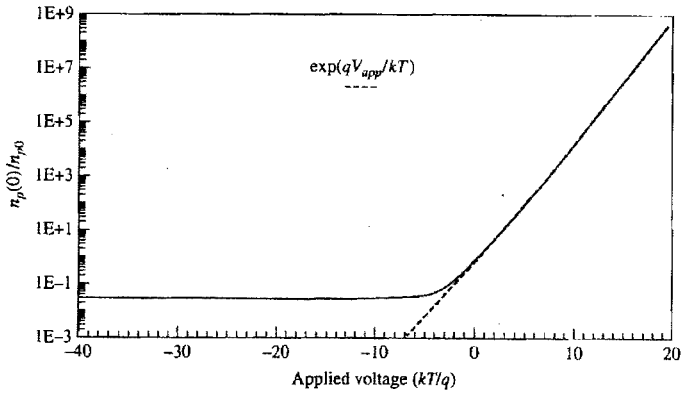


Figure A4.7. Comparison of the ratio $n_p(0)/n_p0$ as given in Eq. (A4.18) with $\exp(qV_{app}/kT)$ for the diode in Fig. A4.6.

$$W_d = \sqrt{\frac{2\varepsilon_{si}\Delta\psi_i}{qN_a}}, \tag{A4.22}$$

where $\Delta\psi_i$ is given by Eq. (A4.9). Using Eqs. (A4.9), (A4.21) and (A4.22), Eq. (A4.15) can be integrated numerically to obtain the parameter δ as a function of V_{app}, which in turn can be used in Eqs. (A4.18) and (A4.19) to obtain the ratio $n_p(0)/n_{p0}$ and the difference $\phi_n(-W_d) - \phi_n(0)$ as a function of V_{app}.

Figure A4.6 is a plot of $n_p(0)/n_{p0}$ and $\phi_n(-W_d) - \phi_n(0)$ as a function of V_{app}. It shows that the difference $\phi_n(W_d) - \phi_n(0)$ as a function of V_{app} has two distinct regions. For forward bias and for small reverse bias ($|V_{app}|$ less than about $4kT/q$), $\phi_n(W_d) - \phi_n(0)$ is small compared to kT/q. For larger reverse bias, the difference $\phi_n(W_d) - \phi_n(0)$ increases with increasing reverse bias, varying approximately linearly with $|V_{app}|$. When $\phi_n(W_d) - \phi_n(0)$ is small compared to kT/q, the ratio $n_p(0)/n_{p0}$ varies exponentially with V_{app}. In the reverse-bias region where $\phi_n(W_d) - \phi_n(0)$ increases approximately linearly with $|V_{app}|$, the ratio $n_p(0)/n_{p0}$ has a saturated value that is practically independent of V_{app}.

Figure A4.7 is a plot comparing the ratio $n_p(0)/n_{p0}$ to $\exp(qV_{app}/kT)$. It shows that $n_p(0)/n_{p0} \sim \exp(qV_{app}/kT)$ in forward bias and for small reverse bias ($|V_{app}|$ less than about $4kT/q$). For larger reverse bias, $\exp(qV_{app}/kT)$ overestimates the degree of minority-electron depletion in the p-region.

Appendix 5 Generation and Recombination Processes and Space-Charge-Region Current

Electron and hole generation and recombination processes play an important role in the operation of many silicon devices and in determining their current–voltage characteristics. Electron and hole generation and recombination can take place directly between the valence band and the conduction band, or indirectly via trap centers in the energy gap. Direct transitions involve energies larger than the bandgap energy E_g. Indirect transitions via trap centers involve energies smaller than E_g. As a result, indirect transitions are much more efficient than direct transitions. In this appendix, we derive the expressions for the generation and recombination rates due to indirect transitions via trap centers and examine the characteristics of generation and recombination currents.

A5.1 Capture and Emission at a Trap Center

Consider a piece of silicon having in it a concentration of N_t trap centers per unit volume. For simplicity, we assume all of the trap centers to be identical and located at energy E_t in the bandgap. Also, we assume that each trap center can exist in one of two charge states, namely neutral when it is not occupied by an electron and negatively charged when it is occupied by an electron. Each unoccupied center can capture an electron from the conduction band (electron capture). An electron in an occupied center can be emitted into the conduction band (electron emission). Similarly, each occupied center can capture a hole from the valence band (hole capture), and an unoccupied center can emit a hole into the valence band (hole emission). These four capture and emission processes are illustrated in Fig. A5.1. Note that in hole capture, the center turns from a negatively charged (occupied) state into a neutral (unoccupied) state. Hole capture from the valence band is equivalent to electron emission into the valence band. In hole emission, the center turns from a neutral (unoccupied) state into a negatively charged (occupied) state. Hole emission is equivalent to electron capture from the valence band.

Let N_t^0 and N_t^- denote the steady-state densities of unoccupied and occupied trap centers, respectively. Let $\bar{N}_t^0$ and $\bar{N}_t^-$ denote the corresponding quantities at thermal equilibrium. We have

$$N_t = N_t^0 + N_t^- = \bar{N}_t^0 + \bar{N}_t^-. \tag{A5.1}$$

The electron capture rate, R_n, is the density of electrons being captured per second. R_n is proportional to the electron density in the conduction band and the density of unoccupied

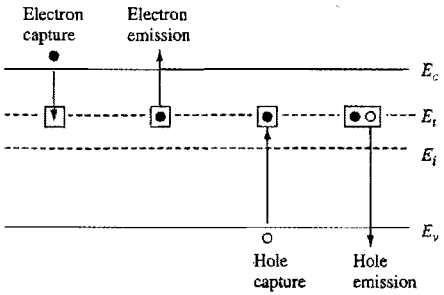


Figure A5.1. Schematic illustrating electron and hole capture and emission processes at a trap center located at every level E_t. E_i is the intrinsic Fermi level. The Fermi level E_f, which lies close to the conduction-band edge in n-type silicon and close to the valence-band edge in p-type silicon, is not shown because the physical mechanisms apply equally to both n-type and p-type silicon.

trap centers. Following the literature (Sah *et al.*, 1957; Hall, 1952; Shockley and Read, 1952), we write

$$R_n = \sigma_n v_{th} n N_t^0, \tag{A5.2}$$

where σ_n is the electron capture cross section and v_{th} is the thermal velocity of electrons. The electron emission rate, G_n, is the density of electrons being emitted per second. Except for very heavily doped n-type silicon, the electron density is small compared with the density of states in the conduction band (see Table 2.1). That is, in practical devices where generation and recombination via trap centers are important, there are plenty of unoccupied states in the conduction band to accept an emitted electron. As a result, the electron emission rate is simply proportional to the concentration of occupied trap centers, i.e.,

$$G_n = \text{constant} \times N_t^-. \tag{A5.3}$$

The *principle of detailed balance* requires the capture and emission rates to be equal at thermal equilibrium. Using this relationship in Eqs. (A5.2) and (A5.3), the constant in Eq. (A5.3) can be obtained, and we have

$$G_n = \sigma_n v_{th} n_0 \left(\frac{\bar{N}_t^0}{\bar{N}_t^-}\right) N_t^-. \tag{A5.4}$$

Similarly, we have

$$R_p = \sigma_p v_{th} p N_t^- \tag{A5.5}$$

and

$$G_p = \sigma_p v_{th} p_0 \left(\frac{\bar{N}_t^-}{\bar{N}_t^0}\right) N_t^0 \tag{A5.6}$$

for the hole capture rate R_p and the hole emission rate G_p, respectively.

A5.2 Steady-State Trap Center Occupation

In a steady state, if an electron in the conduction band is captured and then re-emitted before it recombines with a captured hole, there is no change in the electron density as a result of this capture-followed-by-emission process. An electron is said to have recombined with a hole only after the captured electron has subsequently been emitted into the valence band, i.e., only after a hole has been captured by the occupied center. That is, in a steady state, the net electron capture rate is equal to the net hole capture rate. Mathematically, we have

$$R_n - G_n = R_p - G_p \tag{A5.7}$$

in a steady state. Substituting Eqs. (A5.2) to (A5.6) into Eqs. (A5.7), we obtain

$$\frac{N_t^0}{N_t} = \frac{\sigma_n n_0 \left(\frac{\bar{N}_t^0}{\bar{N}_t^-}\right) + \sigma_p p}{\sigma_n \left[n + n_0 \left(\frac{\bar{N}_t^0}{\bar{N}_t^-}\right)\right] + \sigma_p \left[p + p_0 \left(\frac{\bar{N}_t^-}{\bar{N}_t^0}\right)\right]} \tag{A5.8}$$

and

$$\frac{N_t^-}{N_t} = \frac{\sigma_p p_0 \left(\frac{\bar{N}_t^-}{\bar{N}_t^0}\right) + \sigma_n n}{\sigma_n \left[n + n_0 \left(\frac{\bar{N}_t^0}{\bar{N}_t^-}\right)\right] + \sigma_p \left[p + p_0 \left(\frac{\bar{N}_t^-}{\bar{N}_t^0}\right)\right]}. \tag{A5.9}$$

Equations (A5.8) and (A5.9) give the probabilities of a trap center being unoccupied and occupied, respectively, in a steady-state recombination process.

The probability that an electronic state at energy E is occupied by an electron is given by the Fermi–Dirac distribution $f_D(E) = \{1 + \exp[(E - E_f)/kT]\}^{-1}$, where E_f is the Fermi level [cf. Eq. (2.4)]. Therefore, we have

$$\frac{\bar{N}_t^-}{N_t} = \frac{\bar{N}_t^-}{\bar{N}_t^0 + \bar{N}_t^-} = \frac{1}{1 + e^{(E_t - E_f)/kT}}, \tag{A5.10}$$

which gives

$$\frac{\bar{N}_t^0}{\bar{N}_t^-} = e^{(E_t - E_f)/kT}. \tag{A5.11}$$

From Eqs. (2.49) and (2.50), we have

$$n_0 = n_i e^{(E_f - E_i)/kT} \tag{A5.12}$$

and

$$p_0 = n_i e^{(E_i - E_f)/kT}. \tag{A5.13}$$

Substituting Eqs. (A5.11) to (A5.13) into Eqs. (A5.8) and (A5.9), we obtain

$$\frac{N_t^0}{N_t} = \frac{\sigma_n n_i e^{(E_t - E_i)/kT} + \sigma_p p}{\sigma_n (n + n_i e^{(E_t - E_i)/kT}) + \sigma_p (p + n_i e^{(E_i - E_t)/kT})}, \tag{A5.14}$$

and

$$\frac{N_t^-}{N_t} = \frac{\sigma_p n_i e^{(E_i - E_t)/kT} + \sigma_n n}{\sigma_n (n + n_i e^{(E_t - E_i)/kT}) + \sigma_p (p + n_i e^{(E_i - E_t)/kT})}. \tag{A5.15}$$

The physical interpretation of the above mathematical equations is as follows. A trap center is characterized by its energy E_t and its electron and hole capture cross sections σ_n and σ_p. E_t alone determines the probability that a trap is occupied at thermal equilibrium. However, in a steady-state condition, the probability that a trap is occupied is determined by the properties of the trap center (E_t, σ_n and σ_p) and by the electron density n and hole density p.

A5.3 Net Recombination Rate

Substituting Eqs. (A5.11) to (A5.15) into Eqs. (A5.2) and (A5.4), we can write the *net recombination rate* U (in units of number per unit volume per second) as

$$U \equiv R_n - G_n = R_p - G_p = \frac{\sigma_n \sigma_p v_{th} (np - n_i^2) N_t}{\sigma_n (n + n_i e^{(E_t - E_i)/kT}) + \sigma_p (p + n_i e^{(E_i - E_t)/kT})}. \tag{A5.16}$$

[The notation can be quite confusing. In the literature, U is often called simply the recombination rate. Also, R_n and R_p are often called the electron and hole recombination rates, respectively. To minimize confusion, we will call U the *net* recombination rate, to accurately reflect its physical meaning and to avoid confusion.]

A5.4 Effective Generation and Recombination Centers

Equation (A5.16) shows that the net recombination rate is a function of the electron and hole densities, the electron and hole capture cross sections, and the energy of the trap center, i.e., $U = U(n, p, \sigma_n, \sigma_p, E_t)$. If we assume the electron and hole capture cross sections to be independent of E_t, then we find from $\partial U/\partial E_t = 0$ that the recombination rate is maximum when $E_t = E_{t,\max}$ where

$$E_{t,\max} = E_i + \frac{kT}{2} \ln\left(\frac{\sigma_p}{\sigma_n}\right). \tag{A5.17}$$

Even for a $|\sigma_n/\sigma_p|$ ratio as large as 1000, $|E_{t,\max} - E_i|$ is less than $3.5kT$. That is, ***effective net recombination centers are mid gap states, or states with energy levels close to E_i.*** States with energy levels close to either the conduction band or the valence band do not have comparable electron and hole capture rates to function efficiently as net recombination centers.

Substituting Eq. (A5.17) into Eq. (A5.16), we have

$$U(n, p, \sigma_n, \sigma_p, E_t = E_{t,\max}) = \frac{\sigma_n \sigma_p v_{th} (np - n_i^2) N_t}{\sigma_n n + \sigma_p p + 2n_i \sqrt{\sigma_n \sigma_p}}. \tag{A5.18}$$

Also, Eq. (A5.16) gives

$$U(n, p, \sigma_n, \sigma_p, E_t = E_i) = \frac{\sigma_n \sigma_p v_{th} (np - n_i^2) N_t}{\sigma_n (n + n_i) + \sigma_p (p + n_i)}. \tag{A5.19}$$

So far we have assumed the capture cross sections to be independent of the trap energy. Physically, we expect capture processes involving smaller energies to be more efficient than those involving larger energies. That is, it is reasonable to expect a connection between capture cross section and trap energy. We expect σ_n to be small and σ_p to be large if E_t is close to the valence band, and σ_n to be large and σ_p to be small if E_t is close to the conduction band. Since only traps with energy close to mid gap are efficient net recombination centers, ***it is physically reasonable and convenient to simply assume $E_t = E_i$ and characterize generation and recombination centers by only their capture cross sections***. It is equivalent to using Eq. (A5.19), instead of Eq. (A5.16), for the net recombination rate. We shall do the same here.

Equations (A5.18) and (A5.19) are equal if $\sigma_n = \sigma_p$. That is, the net recombination rate as given by Eq. (A5.19) is maximum for centers having $\sigma_n = \sigma_p$. As we shall see from Eqs. (A5.23) and (A5.24) below, $\sigma_n = \sigma_p$ means that the electron and hole lifetimes are equal.

A5.5 Minority-Carrier Lifetime

Consider a quasineutral silicon region where the carrier concentrations are perturbed by some small amounts, i.e., $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$. We want to derive the lifetimes of the excess minority carriers determined by mid gap trap centers. From Eq. (A5.19), we have

$$U = \frac{\sigma_n \sigma_p v_{th} [(n_0 + \Delta n)(p_0 + \Delta p) - n_i^2] N_t}{\sigma_n (n_0 + \Delta n + n_i) + \sigma_p (p_0 + \Delta p + n_i)}. \tag{A5.20}$$

For n-type silicon, we have $n_0 >> n_i >> p_0$. Quasineutrality means that $\Delta n = \Delta p$. For small perturbations, i.e., for $\Delta n = \Delta p << n_0$, Eq. (A5.20) is reduced to

$$U(\text{n-region}) \approx \sigma_p v_{th} N_t \Delta p. \tag{A5.21}$$

The minority hole lifetime τ_p is defined by

$$\tau_p \equiv \frac{\Delta p}{U(\text{n-region})}. \tag{A5.22}$$

Comparison of Eqs. (A5.21) and (A5.22) gives

$$\tau_p = \frac{1}{\sigma_p v_{th} N_t}. \tag{A5.23}$$

Similarly, for p-type silicon, the minority electron lifetime τ_n is equal to

$$\tau_n = \frac{1}{\sigma_n v_{th} N_t}. \tag{A5.24}$$

A5.6 Generation Rate in a Depletion Region

In a region depleted of mobile carriers, we have $n = 0$ and $p = 0$, and Eq. (A5.19) gives

$$U(n = p = 0) = \frac{-\sigma_n \sigma_p v_{th} n_i N_t}{\sigma_n + \sigma_p}. \tag{A5.25}$$

The net recombination rate in a depletion region is negative because there are no capture events, only emission events. There is a net generation of electrons and holes, hence a negative net recombination rate. It can readily be shown by setting $n = p = 0$ in Eqs. (A5.4) and (A5.6) that the emission rates are

$$G_n(n = p = 0) = G_p(n = p = 0) = \frac{\sigma_n \sigma_p v_{th} n_i N_t}{\sigma_n + \sigma_p}. \tag{A5.26}$$

That is, $G_n = G_p = -U$ in a depletion region, as expected. The generated electrons and holes are swept away in opposite directions by the electric field in the depletion region, giving rise to a leakage current. [The emission rates G_n and G_p are sometimes called generation rates. Here we also use these terms interchangeably.] The generation rates are often expressed in terms of the minority-carrier lifetimes. Using Eqs. (A5.23) and (A5.24) in Eq. (A5.26), we have

$$G_n(n = p = 0) = G_p(n = p = 0) = \frac{n_i}{\tau_n + \tau_p}. \tag{A5.27}$$

A5.7 Net Recombination Rate in a Space-Charge Region

When a p–n diode is forward biased, electrons from the n-side and holes from the p-side are injected into the space-charge region. It is shown in Section 2.2.3.2 that the electron quasi-Fermi levels ϕ_n and the hole quasi-Fermi level ϕ_p are approximately constant inside a space-charge region. However, the electrostatic potential $\psi_i(x) \equiv -E_i(x)/q$ is not spatially constant. This is illustrated in Fig. A5.2. The $n(x)p(x)$ product is given by

$$n(x)p(x) = n_i^2 e^{q(\phi_p - \phi_n)/kT} = n_i^2 e^{qV_{app}/kT}, \tag{A5.28}$$

where V_{app} is the forward bias voltage [cf. Eqs. (2.67) and (2.106)]. That is, $n(x)p(x)$ is constant in the space-charge region. From Eqs. (2.61) and (2.62), we have

$$n(x) = n_i e^{[-q\phi_n + q\psi_i(x)]/kT} \tag{A5.29}$$

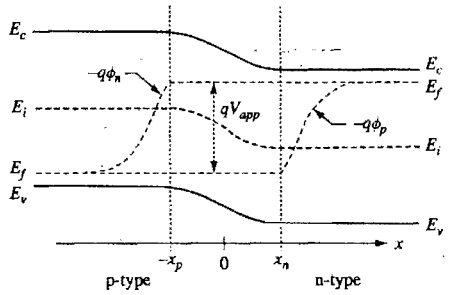


Figure A5.2. Schematic showing the spatial dependence of the quasi-Fermi levels and the intrinsic Fermi level in the space-charge region of a forward-biased p–n diode.

and

$$p(x) = n_i e^{[q\phi_p - q\psi_i(x)]/kT}, \tag{A5.30}$$

which implies that the electron and hole densities have an implicit spatial dependence through $\psi_i(x)$. The net recombination rate, i.e. Eq. (A5.19), in turn has an implicit spatial dependence through $n(x)$ and $p(x)$. Notice that the net recombination rate is also a function of the applied voltage because $\psi_i(x)$ depends on the applied voltage. For a given applied voltage, we want to determine the maximum net recombination rate inside the space-charge region.

Using Eq. (A5.28), we can rewrite Eq. (A5.19) as a function of V_{app} and n, i.e.,

$$U = U(V_{app}, x) = U[V_{app}, n(x)] = \frac{\sigma_n \sigma_p v_{th} n_i^2 \left(e^{qV_{app}/kT} - 1\right) N_t}{\sigma_n (n + n_i) + \sigma_p \left(\dfrac{n_i^2 e^{qV_{app}/kT}}{n} + n_i\right)}. \tag{A5.31}$$

Let us denote by x_m the location where the net recombination rate is maximum. x_m is determined by the equation

$$\left.\frac{\partial U[V_{app}, n(x)]}{\partial n(x)}\right|_{x = x_m} = 0, \tag{A5.32}$$

which gives

$$n(x_m) = \sqrt{\frac{\sigma_p}{\sigma_n}} n_i e^{qV_{app}/2kT}, \tag{A5.33}$$

and

$$p(x_m) = \sqrt{\frac{\sigma_n}{\sigma_p}} n_i e^{qV_{app}/2kT}. \tag{A5.34}$$

Notice that $\sigma_n n(x_m) = \sigma_p p(x_m)$. That is, the net recombination rate is maximum at locations where the probability of an available trap center capturing an electron is equal to the probability of an available trap center capturing a hole. Elsewhere, the net recombination

rate is less because the electron and hole capture events are not balanced. Substituting Eq. (A5.33) into Eq. (A5.31), we have

$$U_{\max} \equiv U(V_{app}, x_m) = \frac{\sigma_n \sigma_p v_{th} n_i^2 \left(e^{qV_{app}/kT} - 1\right) N_t}{2\sqrt{\sigma_n \sigma_p} n_i e^{qV_{app}/2kT} + n_i(\sigma_n + \sigma_p)}, \tag{A5.35}$$

which is the maximum net recombination rate in the space-charge region of a p–n diode forward-biased at a voltage V_{app}. For moderate to high forward biases, Eq. (A5.35) reduces to

$$\begin{aligned} U_{\max} &\approx \frac{\sigma_n \sigma_p v_{th} n_i^2 e^{qV_{app}/kT} N_t}{2\sqrt{\sigma_n \sigma_p} n_i e^{qV_{app}/2kT}} \\ &= \frac{1}{2}\sqrt{\sigma_n \sigma_p} v_{th} n_i e^{qV_{app}/2kT} N_t . \end{aligned} \tag{A5.36}$$

A5.8 Generation–Recombination Current from the Space-Charge Region

The diode current density due to generation and recombination from the space-charge region is a function of the applied voltage, and is given by

$$J_{SC}(V_{app}) = q \int_{-x_p}^{x_n} U(V_{app}, x) dx. \tag{A5.37}$$

We know that at zero bias, $np = n_0 p_0$ and $U(x) = 0$. We also know that for large reverse bias, U is given by Eq. (A5.25). That is

$$J_{SC}(V_{app} < 0) = \frac{-qn_i(x_n + x_p)}{\tau_n + \tau_p} = \frac{-qn_i W_d}{\tau_n + \tau_p}. \tag{A5.38}$$

The reverse-bias space-charge-region current is negative because the generated electrons flow toward the n-side of the diode (+x direction). The magnitude of the reverse current is proportional to the depletion layer width W_d.

When the diode is forward biased ($V_{app} > 0$), we expect the space-charge-region current to be some fraction of the product of the maximum net recombination rate and the space-charge-layer thickness, i.e., we expect

$$J_{SC}(V_{app} > 0) \sim qU_{\max} W_d, \tag{A5.39}$$

where $U_{\max}$ is given by Eq. (A5.35). Using Eq. (A5.36) for $U_{\max}$, we have for large forward biases

$$J_{SC}(qV_{app}/kT \gg 1) \sim q\sqrt{\sigma_n \sigma_p} v_{th} n_i e^{qV_{app}/2kT} N_t W_d. \tag{A5.40}$$

That is, at large forward bias, ***the recombination current in the space-charge region has an approximately exp(qV$_{app}$/2kT) dependence.*** Sah (Sah *et al.*, 1957; Sah, 1991) proposed writing the space-charge-region current in the form

$$J_{SC}(V_{app}) = J_{SC0}(e^{qV_{app}/2kT} - 1), \tag{A5.41}$$

with

$$J_{SC0} = \frac{qn_i W_d}{\tau_n + \tau_p}. \tag{A5.42}$$

Equation (A5.41) shows that the space-charge-region current is zero at zero bias, as required. It reduces to Eq. (A5.38) at large reverse bias and has the correct dependence on V_{app} at large forward bias. In the literature, Eq. (A5.41) is referred to as the *Sah–Noyce–Shockley diode equation* (Sah *et al.*, 1957; Sah, 1991). It gives a good description of the observed space-charge-region currents in practical silicon diodes.

Appendix 6 Diffusion Capacitance of a p–n Diode

In a quasisteady approximation, the mobile carriers in a forward-biased diode are assumed to follow the applied voltage instantaneously. This assumption is a good one for calculating the depletion-layer capacitance where the majority carriers are able to respond to the applied voltage virtually instantaneously. (See Section 2.1.4.8 for a discussion on majority-carrier response time.) However, the redistribution of minority carriers is through diffusion and recombination processes. These processes do not occur instantaneously, but on a time scale related to the minority-carrier lifetime or transit time. As a consequence, when a small signal is applied to a diode, the changes in the minority-carrier densities at different locations in the diode have different phases and cannot be lumped together and treated as a single entity. In this appendix, the diffusion capacitance is derived from a small-signal analysis of the current through a diode starting from the differential equations governing the transport of minority carriers (Shockley, 1949; Lindmayer and Wrigley 1965; Pritchard, 1967). The diffusion capacitance can also be obtained from a transmission-line analysis of a diode equivalent circuit (Bulucea, 1968).

Consider an n^+–p diode with a time-dependent forward-bias voltage $v_{BE}(t)$ applied across it. The emitter is assumed to be wide, i.e., $L_{pE} << W_E$, where L_{pE} is the hole diffusion length in the emitter and W_E is the thickness of the emitter. The base is assumed to be narrow, i.e., $L_{nB} >> W_B$, where L_{nB} is the electron diffusion length in the base and W_B is the base width. (This diode represents the emitter–base diode of an n–p–n bipolar transistor.) We assume that $v_{BE}(t)$ consists of a small-signal voltage $v_{be}(t)$ in series with a dc voltage V_{BE}, i.e., $v_{BE}(t) = V_{BE} + v_{be}(t)$. For simplicity, we assume parasitic resistances are negligible so that $v_{BE}(t)$ is the same as the emitter–base junction voltage. The current flows, including the displacement current, are shown schematically in Figure A6.1, where $i_E(t)$ is the emitter terminal current, $i_B(t)$ is the base terminal current, and $C_{dBE,tot}$ is the depletion-layer capacitance of the emitter–base junction. Overall charge neutrality or Kirchoff's law requires that

$$i_E(t) + i_B(t) = 0. \tag{A6.1}$$

A6.1 Small-Signal Electron and Hole Current Components

Let us first consider the electron current in the p-type base. From Eq. (2.110), the continuity equation for the minority electrons in the base is

$$A_{diode}\frac{\partial n_p(x,t)}{\partial t} = \frac{1}{q}\frac{\partial i_n(x,t)}{\partial x} - A_{diode}\frac{n_p(x,t) - n_{p0}}{\tau_{nB}}, \tag{A6.2}$$

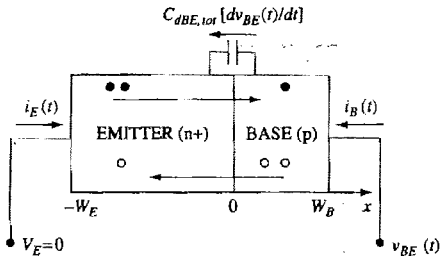


Figure A6.1. Schematic of an n^+–p diode showing the current components when a time-dependent voltage $v_{BE}(t)$ is applied.

where τ_{nB} is the electron lifetime in the base and A_{diode} is the cross-sectional area of the diode. From Eq. (2.54), the electron current is

$$i_n(x,t) = A_{diode}qn_p(x,t)\mu_{nB}\mathscr{E} + A_{diode}qD_{nB}\frac{\partial n_p(x,t)}{\partial x}, \tag{A6.3}$$

where μ_{nB} and D_{nB} are the electron mobility and electron diffusion coefficient, respectively, in the base, and $\mathscr{E}$ is the electric field. If we assume the p-type base to be uniformly doped, so that at low electron injection currents the hole density is uniform in the p-region, then both the electric field and the field gradient are negligible in the p-region, and Eq. (A6.3) becomes

$$\frac{\partial i_n(x,t)}{\partial x} = A_{diode}qD_{nB}\frac{\partial^2 n_p(x,t)}{\partial x^2}, \tag{A6.4}$$

and (A6.2) becomes

$$\frac{\partial n_p(x,t)}{\partial t} = D_{nB}\frac{\partial^2 n_p(x,t)}{\partial x^2} - \frac{n_p(x,t) - n_{p0}}{\tau_{nB}}. \tag{A6.5}$$

If we assume the small signals to have a time dependence $e^{j\omega t}$, i.e., $v_{be}(t) = v_{be}\,e^{j\omega t}$, and the electron concentration has a form

$$n_p(x,t) = n_p(x) + \Delta n_p(x)e^{j\omega t}, \tag{A6.6}$$

then Eq. (A6.5) gives

$$\Delta n_p(x)j\omega e^{j\omega t} = D_{nB}\frac{d^2 n_p(x)}{dx^2} - \frac{n_p(x) - n_{p0}}{\tau_{nB}} + \left(D_{nB}\frac{d^2\Delta n_p(x)}{dx^2} - \frac{\Delta n_p(x)}{\tau_{nB}}\right)e^{j\omega t}. \tag{A6.7}$$

For Eq. (A6.7) to be valid at all times, it must be valid for the time-independent part and the time-dependent part separately, i.e.,

$$D_{nB}\frac{d^2 n_p(x)}{dx^2} - \frac{n_p(x) - n_{p0}}{\tau_{nB}} = 0, \tag{A6.8}$$

and

$$\Delta n_p(x)j\omega = D_{nB}\frac{d^2\Delta n_p(x)}{dx^2} - \frac{\Delta n_p(x)}{\tau_{nB}}. \tag{A6.9}$$

Equation (A6.8) is simply the equation governing the quasisteady state charge distribution of a forward-biased diode, which has been solved in Section 2.2.4.1. The solution is given in Eq. (2.119), i.e.,

$$\begin{aligned} n_p(x) - n_{p0} &= n_{p0}\left[\exp\left(\frac{qV_{BE}}{kT}\right) - 1\right]\frac{\sinh[(W_B - x)/L_{nB}]}{\sinh(W_B/L_{nB})} \\ &\approx n_{p0}\exp\left(\frac{qV_{BE}}{kT}\right)\frac{\sinh[(W_B - x)/L_{nB}]}{\sinh(W_B/L_{nB})}, \end{aligned} \tag{A6.10}$$

where the diffusion length is related to the diffusion coefficient and the lifetime through

$$L_{nB} = \sqrt{\tau_{nB}D_{nB}}. \tag{A6.11}$$

Equation (A6.9) can be re-arranged to give

$$\frac{d^2\Delta n_p(x)}{dx^2} - \frac{\Delta n_p(x)}{L_{nB}^2} = 0, \tag{A6.12}$$

where

$$L'_{nB} \equiv \sqrt{\frac{\tau_{nB}D_{nB}}{1 + j\omega\tau_{nB}}} = \frac{L_{nB}}{\sqrt{1 + j\omega\tau_{nB}}}. \tag{A6.13}$$

Equation (A6.12) has the same form as Eq. (A6.8), and hence can be solved in the same manner. The general solution to (A6.12) is

$$\Delta n_p(x) = A\exp(x/L'_{nB}) + B\exp(-x/L'_{nB}), \tag{A6.14}$$

where A and B can be determined from the boundary conditions for $\Delta n_p(x)$.

Expanding the Shockley diode equation, Eq. (2.107), and keeping only the first order term in v_{be}, we have

$$n_p(0,t) = n_{p0}\exp(qv_{BE}(t)/kT) \approx n_{p0}\exp(qV_{BE}/kT)\left[1 + \frac{q}{kT}v_{be}e^{j\omega t}\right]. \tag{A6.15}$$

Therefore, the boundary condition for $\Delta n_p(x)$ at $x = 0$ is

$$\Delta n_p(0) = n_{p0}\exp(qV_{BE}/kT)\frac{qv_{be}}{kT}. \tag{A6.16}$$

At the ohmic base contact, there are no excess electrons. Therefore the boundary condition for $\Delta n_p(x)$ at $x = W_B$ is

$$\Delta n_p(W_B) = 0. \tag{A6.17}$$

Applying these boundary conditions to Eq. (A6.12), we obtain

$$\Delta n_p(x) = \frac{n_{p0}qv_{be}}{kT}\exp\left(\frac{qV_{BE}}{kT}\right)\frac{\sinh[(W_B - x)/L'_{nB}]}{\sinh(W_B/L'_{nB})}. \tag{A6.18}$$

The electron current entering the p-type base is [see Eq. (2.120)]

$$\begin{aligned} i_n(0,t) &= A_{diode}qD_{nB}\left(\frac{\partial n_p(x,t)}{\partial x}\right)_{x=0} \\ &= A_{diode}qD_{nB}\left(\frac{dn_p(x)}{dx}\right)_{x=0} + A_{diode}qD_{nB}e^{j\omega t}\left(\frac{d\Delta n_p(x)}{dx}\right)_{x=0} \\ &= -I_n + \Delta i_n(0,t), \end{aligned} \tag{A6.19}$$

where

$$-I_n = A_{diode}qD_{nB}\left(\frac{dn_p(x)}{dx}\right)_{x=0} = -A_{diode}\frac{qD_{nB}n_{p0}\exp(qV_{BE}/kT)}{L_{nB}\tanh(W_B/L_{nB})} \tag{A6.20}$$

is the steady-state electron current, and

$$\begin{aligned} \Delta i_n(0,t) &= A_{diode}qD_{nB}e^{j\omega t}\left(\frac{d\Delta n_p(x)}{dx}\right)_{x=0} \\ &= -A_{diode}\frac{qD_{nB}n_{p0}\exp(qV_{BE}/kT)}{L'_{nB}\tanh(W_B/L'_{nB})}\frac{qv_{be}(t)}{kT} \\ &= -\left(\frac{L_{nB}}{L'_{nB}}\right)\frac{\tanh(W_B/L_{nB})}{\tanh(W_B/L'_{nB})}\frac{qI_n}{kT}v_{be}(t) \end{aligned} \tag{A6.21}$$

is the small-signal electron current entering the p-type base. (Note that i_n is negative because electrons flowing in the x direction give rise to a negative current. I_n is the magnitude of the steady-state electron current, a positive quantity.)

The small-signal hole current entering the emitter can be derived in a similar manner (see Exercise 2.16). It is

$$\Delta i_p(0,t) = -\left(\frac{L_{pE}}{L'_{pE}}\right)\frac{\tanh(W_E/L_{pE})}{\tanh(W_E/L'_{pE})}\frac{qI_p}{kT}v_{be}(t), \tag{A6.22}$$

where

$$L_{pE} = \sqrt{\tau_{pE}D_{pE}} \tag{A6.23}$$

and

$$L'_{pE} \equiv \sqrt{\frac{\tau_{pE}D_{pE}}{1 + j\omega\tau_{pE}}} = \frac{L_{pE}}{\sqrt{1 + j\omega\tau_{pE}}} \tag{A6.24}$$

are the corresponding parameters for holes in the n^+ emitter. I_p is the magnitude of the steady-state hole current flowing from the base into the emitter. (Note that i_p is negative, since holes flowing in the $-x$ direction give a negative current flow.)

Since the base is assumed to be narrow, i.e., $W_B/L_{nB} << 1$, we have $\tanh(W_B/L_{nB}) \approx W_B/L_{nB}$, and Eq. (A6.21) gives

$$\Delta i_n(0,t) \approx -\left(\frac{W_B}{L'_{nB}}\right)\frac{1}{\tanh(W_B/L'_{nB})}\frac{qI_n}{kT}v_{be}(t) \quad \text{(narrow base)}. \tag{A6.25}$$

As the emitter is assumed to be wide, we have $W_E/L_{pE} >> 1$ and $\tanh(W_E/L_{pE}) \approx 1$. It also implies $\left(W_E/L'_{pE}\right) = \left(W_E/L_{pE}\right)\sqrt{1+j\omega\tau_{pE}} >> 1$. Therefore, $\tanh(W_E/L'_{pE}) \approx 1$, and Eq. (A6.22) gives

$$\Delta i_p(0,t) \approx -\left(\frac{L_{pE}}{L'_{pE}}\right)\frac{qI_p}{kT}v_{be}(t) = -\sqrt{1+j\omega\tau_{pE}}\frac{qI_p}{kT}v_{be}(t) \quad \text{(wide emitter)}. \tag{A6.26}$$

A6.2 Small-Signal Base Terminal Current

From the current components shown in Fig. A6.1, the base current is

$$\begin{aligned} i_B(t) &= -i_n(0,t) - i_p(0,t) + C_{dBE,tot}\frac{dv_{BE}(t)}{dt} \\ &= I_n + I_p - \Delta i_n(0,t) - \Delta i_p(0,t) + j\omega C_{dBE,tot}v_{be}(t) \\ &= I_B + i_b(t), \end{aligned} \tag{A6.27}$$

where

$$I_B = I_n + I_p \tag{A6.28}$$

is the steady-state base terminal current and

$$i_b(t) = -\Delta i_n(0,t) - \Delta i_p(0,t) + j\omega C_{dBE,tot}v_{be}(t) \tag{A6.29}$$

is the small-signal base terminal current.

As discussed in Section 2.2.5.1, for charging and discharging a wide-emitter region, the time constant is the lifetime τ_E for the emitter. For charging and discharging a narrow-base region, the time constant is the base transit time t_B. The lifetime in the emitter is typically smaller than 10^{-9} s [see Fig. 2.24(b)] while the base transit time of a modern bipolar transistor is typically smaller than 10^{-11} s (see Section 2.2.5.2). Therefore, a low-frequency assumption for the wide emitter, i.e., $\omega\tau_{pE} < 1$, also means low-frequency for the narrow base, i.e., $\omega t_B < 1$. With t_B smaller than 10^{-11} s for a modern bipolar transistor, the condition of $\omega t_B > 1$ is rarely reached. That is, in considering the diffusion capacitance of a modern bipolar transistor, we need to consider the cases of $\omega\tau_{pE} < 1$ and $\omega\tau_{pE} > 1$. In both cases, we have $\omega t_B < 1$.

A6.3 Low-Frequency [$\omega\tau_{pE} < 1$ and $\omega t_B < 1$] Diffusion Capacitance

In the low-frequency approximation, Eq. (A6.25) can be expanded to keep only the lowest terms in W_B/L_{nB} and ωt_B, i.e.,

$$\begin{aligned} \Delta i_n(0,t) &\approx -\left[1 + \frac{W_B^2}{3L_{nB}^2}(1+j\omega\tau_{nB}) + \ldots\right]\frac{qI_n}{kT}v_{be}(t) \quad \text{(narrow base)} \\ &\approx -\left(1 + j\omega\frac{W_B^2\tau_{nB}}{3L_{nB}^2}\right)\frac{qI_n}{kT}v_{be}(t) \\ &= -\left(1 + j\frac{2\omega t_B}{3}\right)\frac{qI_n}{kT}v_{be}(t), \end{aligned} \tag{A6.30}$$

where we have used Eq. (2.147) for the base transit time t_B. Similarly, Eq. (A6.26) gives

$$\Delta i_p(0,t) \approx -\left(1 + j\frac{\omega\tau_{pE}}{2}\right)\frac{qI_p}{kT}v_{be}(t) \tag{A6.31}$$

Substituting Eqs. (A6.30) and (A6.31) into Eq. (A6.29), we have

$$\begin{aligned} i_b(t) &= \frac{q}{kT}(I_n + I_p)v_{be}(t) + j\omega\left(\frac{2qI_nt_B}{3kT} + \frac{qI_p\tau_{pE}}{2kT}\right)v_{be}(t) + j\omega C_{dBE,tot}v_{be}(t) \\ &= \frac{1}{r_d}v_{be}(t) + j\omega\left(C_{Dn} + C_{Dp} + C_{dBE,tot}\right)v_{be}(t), \end{aligned} \tag{A6.32}$$

where

$$\frac{1}{r_d} = \frac{q}{kT}(I_n + I_p) = \frac{qI_B}{kT} \tag{A6.33}$$

is the ac conductance of the diode,

$$C_{Dn} = \frac{2qI_nt_B}{3kT} \quad \text{(narrow base)} \tag{A6.34}$$

is the diffusion capacitance due to the minority electrons in the narrow base, and

$$C_{Dp} = \frac{qI_p\tau_{pE}}{2kT} \quad \text{(wide emitter)} \tag{A6.35}$$

is the diffusion capacitance due to the minority holes in the wide emitter. These diffusion capacitance results were stated without proof in Section 2.2.6. Also, as discussed in

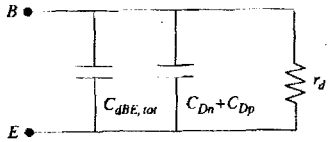


Figure A6.2. Small-signal equivalent circuit for a forward-biased diode. Parasitic resistances are ignored.

Section 2.2.6, the diffusion capacitance component from the emitter is usually small compared to that from the base, i.e., $C_{Dp} < C_{Dn}$ in our n^+–p diode.

Note that, even in the low-frequency approximation, the diffusion capacitance due to minority carriers cannot be calculated quasistatically. Quasistatic calculation gives a diffusion capacitance in a wide emitter that is twice as large as it should be, and a diffusion capacitance in a narrow base that is 3/2 as large as it should be. (See the discussion in Section 2.2.6.)

The small-signal equivalent circuit for a forward-biased diode can be obtained from Eq. (A6.32). It is shown in Fig. A6.2 in the form of a lumped-component model. It should be noted that in the low-frequency approximation, the diffusion capacitances are independent of frequency.

A6.4 Diffusion Capacitance at High-Frequencies [$\omega\tau_{pE} > 1$]

In this case, Eq. (A6.26) gives

$$\begin{aligned}\Delta i_p(0,t) &\approx -\sqrt{j\omega\tau_{pE}}\frac{qI_p}{kT}v_{be}(t)\\ &= -\left(\sqrt{\frac{\omega\tau_{pE}}{2}}+j\omega\sqrt{\frac{\tau_{pE}}{2\omega}}\right)\frac{qI_p}{kT}v_{be}(t),\end{aligned} \tag{A6.36}$$

which implies that the diffusion capacitance due to the minority holes in the emitter is

$$C_{Dp} = \frac{qI_p\tau_{pE}}{kT\sqrt{2\omega\tau_{pE}}} \qquad \text{(wide emitter)}. \tag{A6.37}$$

Comparison of Eqs. (A6.37) and (A6.35) suggests that the emitter component is approximately independent of frequency until about $\omega\tau_{pE}\sim 2$, and then decreases as $1/\sqrt{\omega\tau_{pE}}$ at higher frequencies. As discussed earlier, Eq. (A6.34) is valid for the base component in this high-frequency régime.

Appendix 7 Image-Force-Induced Barrier Lowering

Consider a metal surface exposed to free space, as illustrated in Figure A7.1. The metal work function $q\phi_m$ is the energy difference between the Fermi level and the free electron level. In the absence of any externally applied electric field, an electron at the Fermi level must gain a kinetic energy equal to $q\phi_m$ in order to break loose from the crystal field of the metal.

Let us now consider an electron at a distance x from the metal surface. This electron induces a positive image charge located at $-x$ in the metal. The attractive force exerted by the positive image charge on the electron is equal to

$$F_{image}(x) = \frac{-q^2}{4\pi\varepsilon_0(2x)^2} = \frac{-q^2}{16\pi\varepsilon_0 x^2}, \tag{A7.1}$$

where ε_0 is the vacuum permittivity. The image force has a negative sign because it pulls the electron towards the metal surface. The electric field (electrostatic force per unit charge) due to this image force is

$$\mathscr{E}_{image}(x) = \frac{F_{image}(x)}{-q} = \frac{q}{16\pi\varepsilon_0 x^2}. \tag{A7.2}$$

The corresponding electrostatic potential $\psi_{image}(x)$ is given by the equation

$$\mathscr{E}_{image}(x) = -\frac{d\psi_{image}(x)}{dx}, \tag{A7.3}$$

which can be integrated, subject to the boundary condition of $\psi_{image}(\infty)=0$, to give

$$\psi_{image}(x) = \int_x^\infty \mathscr{E}_{image}(x)dx = \frac{q}{16\pi\varepsilon_0 x}. \tag{A7.4}$$

If a voltage is applied to pull an electron away from the metal surface, this voltage establishes an electric field $\mathscr{E}(x)$. (Note that in the metal–free-space system being considered, $\mathscr{E}$ is independent of x. Also, $\mathscr{E}$ is negative because it is in a direction of driving an electron away from the metal surface.) The electrostatic potential $\psi_{field}(x)$ associated with this field is given by

$$\mathscr{E}(x) = -\frac{d\psi_{field}(x)}{dx}, \tag{A7.5}$$

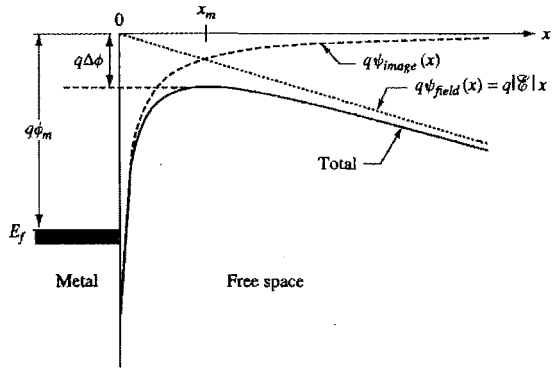


Figure A7.1. Schematic showing the energy-band diagram between a metal and a free space appropriate for electron emission from the metal into the free space. The metal work function is $q\phi_m$. However, the energy barrier for electron emission is $q(\phi_m - \Delta\phi)$.

which can be integrated, subject to the boundary condition of $\psi_{field}(0) = 0$, to give

$$\psi_{field}(x) = -\int_0^x \mathscr{E}(x)dx = |\mathscr{E}|x. \tag{A7.6}$$

Therefore, the total electrostatic potential as seen by an electron being pulled away from a metal surface by an electric field $\mathscr{E}$ is $\psi_{image}(x) + \psi_{field}(x)$. The electron energy associated with this potential is

$$PE(x) = -q\left[\psi_{image}(x) + \psi_{field}(x)\right] = -\left(\frac{q^2}{16\pi\varepsilon_0 x} + q|\mathscr{E}|x\right). \tag{A7.7}$$

The electron energies associated with $\psi_{image}(x)$, $\psi_{field}(x)$, as well as the total electrostatic potential are illustrated in Fig. A7.1. Note that $PE(x)$ has a peak that lies below the free electron level ($E_f + q\phi_m$). That is, the combination of an electric field and the image force lowers the energy barrier for electron emission from a metal surface into the free space. This is known as *Schottky effect* or *Schottky barrier lowering*. The amount of energy barrier lowering $q\Delta\phi$ and the location x_m of the potential energy peak are indicated in Fig. A7.1. They can be obtained from $dPE(x)/dx = 0$. The results are:

$$x_m = \sqrt{\frac{q}{16\pi\varepsilon_0|\mathscr{E}|}}, \tag{A7.8}$$

and

$$q\Delta\phi = \sqrt{\frac{q^3|\mathscr{E}|}{4\pi\varepsilon_0}}. \tag{A7.9}$$

Schottky barrier lowering also applies to a metal–semiconductor system (Schottky barrier diode). In this case, ε_0 in Eq. (A7.1) is replaced by the permittivity of the

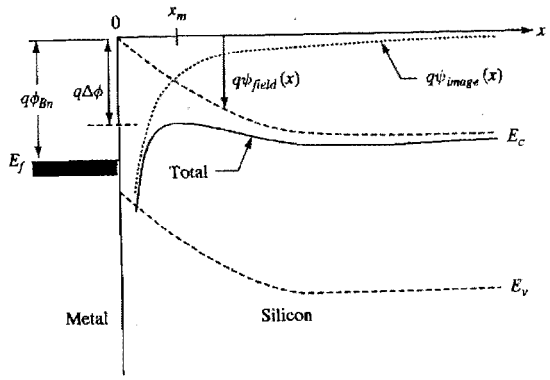


Figure A7.2. Schematic showing the energy-band diagram of a metal–silicon Schottky barrier diode. Also shown are the electrostatic potentials involved in determining the energy barrier for electron emission. E_c is the silicon conduction-band edge. Ignoring Schottky barrier lowering, the energy barrier for electrons is $q\phi_{Bn}$. The image force lowers this energy barrier by $q\Delta\phi$.

semiconductor (Sze *et al.*, 1964). The energy-band diagram of a metal–silicon Schottky barrier diode for illustrating Schottky effect is shown schematically in Fig. A7.2. There is an electric field associated with the band bending near the surface. The electric field is not constant because of the charge in the depletion region. The electrostatic potential associated with this field is $q\psi_{field}(x)$. The maximum electric field is located at the metal–silicon interface, and has an absolute value of

$$\mathscr{E}_m = \frac{qN_dW_d}{\varepsilon_{si}}, \tag{A7.10}$$

where N_d is the dopant concentration, ε_{si} is the permittivity of silicon, and W_d is the width of the space-charge layer [see Eq. (2.78)]. It is left as an exercise (Ex. A7.1) to the reader to show that for typical Schottky barrier diodes, the location of the potential barrier peak, x_m, is small compared with W_d. With $x_m << W_d$, we have

$$x_m \approx \sqrt{\frac{q}{16\pi\varepsilon_{si}\mathscr{E}_m}}, \tag{A7.11}$$

and

$$q\Delta\phi \approx \sqrt{\frac{q^3\mathscr{E}_m}{4\pi\varepsilon_{si}}}. \tag{A7.12}$$

That is, for a metal–semiconductor system, it is the maximum electric field at the surface that determines the amount of Schottky barrier lowering. For a maximum electric field of $\mathscr{E}_m = 1 \times 10^5$ V/cm, $q\Delta\phi = 35$ meV for a silicon Schottky diode.

In the literature, the Schottky effect has also been invoked in the studies of electron injection from a metal or semiconductor into an insulator (Berglund and Powell, 1971) by

replacing ϵ_0 in Eqs. (A7.8) and (A7.9) by the permittivity of the insulator. However, there are also publications questioning the validity of the concept of image force at the interface between a semiconductor and an insulator. Interested readers are referred to the publications for more detailed discussions (see e.g. Fischetti *et al.*, 1995, and the references therein).

Exercise

A7.1 The energy-band diagram of a metal–silicon Schottky barrier diode and the electrostatic potentials involved in determining the energy barrier for electron emission is shown schematically in Fig. A7.2. Derive the equation for the total potential energy $PE(x) = -q[\psi_{image}(x) + \psi_{field}(x)]$ in terms of the dopant concentration N_d and the depletion layer thickness W_d. Show that the location x_m of the peak of $PE(x)$ is small compared with the depletion layer thickness W_d in the silicon. Show that x_m is given by

$$x_m \approx \sqrt{\frac{q}{16\pi\varepsilon_{si}\mathscr{E}_m}}$$

and the image-force-induced barrier lowering $q\Delta\phi$ is given by

$$q\Delta\phi \approx \sqrt{\frac{q^3\mathscr{E}_m}{4\pi\varepsilon_{si}}},$$

where $\mathscr{E}_m$ is absolute value of the electric field in the silicon at the metal–silicon interface.

Appendix 8 Electron-Initiated and Hole-Initiated Avalanche Breakdown

In Section 2.5.1, we showed that the multiplication factor for hole-initiated impact ionization is

$$\frac{1}{M_p} = \exp\left(-\int_0^W (\alpha_p - \alpha_n)dx\right) - \int_0^W \alpha_n \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx, \tag{A8.1}$$

and that for electron-initiated impact ionization is

$$\frac{1}{M_n} = \exp\left(-\int_0^W (\alpha_n - \alpha_p)dx\right) - \int_0^W \alpha_p \exp\left(-\int_x^W (\alpha_n - \alpha_p)dx'\right)dx. \tag{A8.2}$$

It can be shown (see Exercise 2.12) that in general

$$\int_0^W f(x)\exp\left(-\int_0^x f(x')dx'\right)dx = 1 - \exp\left(-\int_0^W f(x)dx\right) \tag{A8.3}$$

and

$$\int_0^W f(x)\exp\left(-\int_x^W f(x')dx'\right)dx = 1 - \exp\left(-\int_0^W f(x)dx\right). \tag{A8.4}$$

Applying Eq. (A8.3) to Eq. (A8.1), we obtain

$$1 - \frac{1}{M_p} = \int_0^W \alpha_p \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx. \tag{A8.5}$$

Similarly, applying Eq. (A8.4) to Eq. (A8.2), we have

$$1 - \frac{1}{M_n} = \int_0^W \alpha_n \exp\left(-\int_x^W (\alpha_n - \alpha_p)dx'\right)dx. \tag{A8.6}$$

Avalanche breakdown occurs when carrier multiplication by impact ionization runs away, i.e., when the multiplication factor becomes infinite. Thus, Eq. (A8.5) gives the hole-initiated breakdown condition as

$$\int_0^W \alpha_p \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx = 1. \tag{A8.7}$$

Using Eq. (A8.3) the left-hand side of Eq. (A8.7) can be rearranged to give

$$\int_0^W \alpha_p \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx = 1 - \exp\left(-\int_0^W (\alpha_p - \alpha_n)dx\right) + \int_0^W \alpha_n \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx. \tag{A8.8}$$

Substituting Eq. (A8.7) into Eq. (A8.8) gives

$$\int_0^W \alpha_n \exp\left(-\int_0^x (\alpha_p - \alpha_n)dx'\right)dx = \exp\left(-\int_0^W (\alpha_p - \alpha_n)dx\right). \tag{A8.9}$$

Dividing both sides of Eq. (A8.9) by its RHS gives

$$\int_0^W \alpha_n \exp\left(-\int_x^W (\alpha_n - \alpha_p)dx'\right)dx = 1, \tag{A8.10}$$

which, according to Eq. (A8.6), is simply the condition for electron-initiated breakdown. Thus ***the condition for avalanche breakdown is the same whether the breakdown process is initiated by electrons or by holes.***

Appendix 9 An Analytical Solution for the Short-Channel Effect in Subthreshold

In this appendix, we outline the mathematical approach that leads to the analytical expression, Eq. (3.67) in Section 3.2.1, for the short-channel threshold rolloff. The short-channel effect (SCE) is a very complex mathematical problem involving the solution of an irregular 2-D boundary-value problem. It is impractical to derive an exact analytical solution applicable to all general cases. Numerical simulations running on a finite-element program should be used to obtain accurate solutions for specific device geometries and doping conditions. Nevertheless, an analytical expression, even an approximate one, goes a long way in providing valuable insights into the fundamentals of the short-channel effect and its controlling parameters. The approach here essentially follows the Ph.D. thesis of Thao N. Nguyen published in 1984.

A9.1 Defining the Problem with Simplified Boundary Conditions

To simplify the 2-D boundary-value problem to a manageable level, we make a number of approximations so as to retain only the most basic aspects of the short-channel effect. A simplified short-channel MOSFET geometry is shown in Fig. A9.1 (Nguyen, 1984). The x-axis is along the vertical direction, the y-axis along the horizontal direction, and the origin at point A. As in Section 2.3.2, $\psi(x, y) = \psi_i(x, y) - \psi_i(x = \infty)$ is defined as the intrinsic potential at a point (x, y) with respect to the intrinsic potential of the p-type substrate. The substrate is assumed to be uniformly doped with a concentration N_a. In the oxide region $AFGH$, Poisson's equation becomes a homogeneous (Laplace) equation,

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = 0. \tag{A9.1}$$

In the depletion region in silicon, the concentrations of both types of mobile carriers are negligible under subthreshold conditions. Therefore, in $ABEF$, Poisson's equation is approximated by

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = \frac{qN_a}{\varepsilon_{si}}. \tag{A9.2}$$

The length of the silicon region is equal to the channel length L. The depth is given by the depletion layer width, W_d, to be determined later.

As described by Eq. (2.47), the normal component of the electric field changes by a factor of $\varepsilon_{si}/\varepsilon_{ox} \approx 3$ across the silicon–oxide boundary AF. To eliminate this boundary condition so that ψ and its derivatives are continuous, the oxide is replaced by an equivalent region of the same dielectric constant as silicon, but with a thickness equal to $3t_{ox}$, as was done in Fig. 3.5. This preserves the capacitance and allows the entire rectangular region to be treated as a homogeneous material of dielectric constant ε_{si}. The drawback is that it may cause some errors in the tangential field, whose magnitude does not change across the silicon–oxide boundary (Eq. (2.46)). In the equivalent structure, the tangential field apparently experiences a thicker-than-actual oxide. The errors are expected to be small when the gate oxide is thin compared to the silicon depletion depth W_d so that the oxide field is dominated by its normal component.

If we assume that the source and drain junctions are abrupt and deeper than W_d,[1] we can write down the following set of simplified boundary conditions:

$$\psi(-3t_{ox}, y) = V_{gs} - V_{fb} \qquad \text{along } GH, \tag{A9.3}$$

$$\psi(x, 0) = \psi_{bi} \qquad \text{along } AB, \tag{A9.4}$$

$$\psi(x, L) = \psi_{bi} + V_{ds} \qquad \text{along } EF, \tag{A9.5}$$

$$\psi(W_d, y) = 0 \qquad \text{along } CD, \tag{A9.6}$$

where V_{gs} and V_{ds} are the gate and source–drain voltages defined in Section 3.1.1, V_{fb} is the flatband voltage of the gate electrode, and ψ_{bi} is the built-in potential of the source or drain-to-substrate junction. For an abrupt n^+–p junction, $\psi_{bi} = E_g/2q + \psi_B$, where ψ_B is given by Eq. (2.48). If there is a substrate bias V_{bs}, then ψ_{bi} should be replaced by $\psi_{bi} - V_{bs}$ in Eqs. (A9.4) and (A9.5), and V_{gs} by $V_{gs} - V_{bs}$ in Eq. (A9.3). The bottom boundary is actually a movable one, as W_d is subject to change with the gate voltage V_{gs}. The distance BC is approximately given by the source junction depletion width,

$$W_S = \sqrt{\frac{2\varepsilon_{si}\psi_{bi}}{qN_a}}. \tag{A9.7}$$

Similarly, DE is given by the drain junction depletion width,

$$W_D = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} + V_{ds})}{qN_a}}. \tag{A9.8}$$

The boundary conditions for the potential along FG and HA are assumed to vary linearly between the end-point values, while those along BC and DE are assumed to vary parabolically between the end points.

[1] This is the technologically more relevant case. The case where the source and drain junction depths are shallower than the gate depletion width is discussed toward the end of this appendix.

A9.2 Solution Techniques

The solution technique makes use of the superposition principle and breaks the electrostatic potential into the following terms:

$$\psi(x, y) = v(x, y) + u_L(x, y) + u_R(x, y) + u_B(x, y). \tag{A9.9}$$

Here $v(x, y)$ is a solution to the inhomogeneous (Poisson's) equation and satisfies the top boundary condition, Eq. (A9.3). u_L, u_R, u_B are all solutions to the homogeneous (Laplace) equation, and are chosen in order for $\psi(x, y)$ to satisfy the rest of the boundary conditions, namely, on the left, the right, and the bottom of the rectangular box in Fig. A9.1. For example, u_L is zero on the top, bottom, and right boundaries, but $v + u_L$ satisfies the left boundary condition, Eq. (A9.4). Likewise, u_R is zero on the top, bottom, and left boundaries, but $v + u_R$ satisfies the right boundary condition, Eq. (A9.5), and so on.

A natural choice for $v(x, y)$ [actually $v(x)$] is the long-channel, 1-D MOS solution employing the depletion approximation discussed in Section 2.3.2:

$$v(x, y) = \psi_s - \frac{V_{gs} - V_{fb} - \psi_s}{3t_{ox}} x \quad \text{for the oxide region,} \quad -3t_{ox} \le x \le 0, \tag{A9.10}$$

and

$$v(x, y) = \psi_s \left(1 - \frac{x}{W_d}\right)^2 \quad \text{for the silicon region,} \quad 0 \le x \le W_d. \tag{A9.11}$$

Here ψ_s is the long-channel surface potential which varies with V_{gs}. It is related to the depletion region depth W_d by

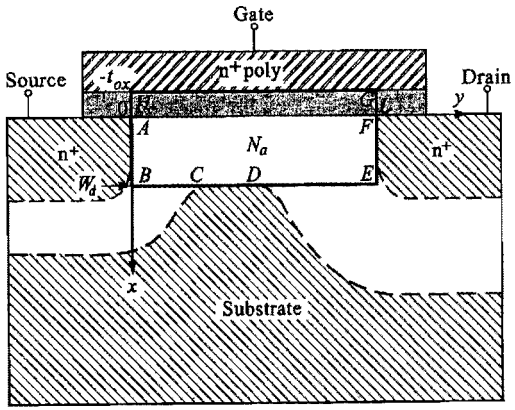


Figure A9.1. Simplified geometry for analytically solving Poisson's equation in a short-channel MOSFET. The hashed areas represent conductor-like regions of constant potential. (After Nguyen, 1984.)

$$\psi_s = \frac{qN_a W_d^2}{2\varepsilon_{si}} \tag{A9.12}$$

for Eq. (A9.11) to satisfy the differential equation (A9.2). Note that Eqs. (A9.10) and (A9.11) satisfy both the top and the CD (Fig. A9.1) part of the bottom boundary conditions, Eqs. (A9.3) and (A9.6), and are continuous at $x=0$. The requirement that $\partial v/\partial x$ be continuous at $x=0$ gives the relationship between ψ_s and V_{gs}:

$$\frac{V_{gs} - V_{fb} - \psi_s}{3t_{ox}} = \frac{2\psi_s}{W_d} = \sqrt{\frac{2qN_a\psi_s}{\varepsilon_{si}}}. \tag{A9.13}$$

The last step made use of Eq. (A9.12). Note that Eq. (A9.13) is just the gate bias equation in subthreshold, Eq. (3.38).

The rest of the solutions are of the following forms (Nguyen, 1984):

$$u_L(x,y) = \sum_{n=1}^{\infty} b_n \frac{\sinh\left(\frac{n\pi(L-y)}{W_d+3t_{ox}}\right)}{\sinh\left(\frac{n\pi L}{W_d+3t_{ox}}\right)} \sin\left(\frac{n\pi(x+3t_{ox})}{W_d+3t_{ox}}\right), \tag{A9.14}$$

$$u_R(x,y) = \sum_{n=1}^{\infty} c_n \frac{\sinh\left(\frac{n\pi y}{W_d+3t_{ox}}\right)}{\sinh\left(\frac{n\pi L}{W_d+3t_{ox}}\right)} \sin\left(\frac{n\pi(x+3t_{ox})}{W_d+3t_{ox}}\right), \tag{A9.15}$$

and

$$u_B(x,y) = \sum_{n=1}^{\infty} d_n \frac{\sinh\left(\frac{n\pi(x+3t_{ox})}{L}\right)}{\sinh\left(\frac{n\pi(W_d+3t_{ox})}{L}\right)} \sin\left(\frac{n\pi y}{L}\right). \tag{A9.16}$$

The coefficients are determined by requiring $\psi(x,y)$ to satisfy the boundary conditions. For example, on the left boundary,

$$\psi(x,0) = v(x,0) + u_L(x,0) = v(x,0) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi(x+3t_{ox})}{W_d+3t_{ox}}\right). \tag{A9.17}$$

To evaluate the individual coefficients, Eq. (A9.17) is multiplied by the orthogonal eigenfunctions and integrated over $(-3t_{ox}, W_d)$,

$$b_n = \frac{2}{W_d+3t_{ox}} \int_{-3t_{ox}}^{W_d} [\psi(x,0) - v(x,0)] \sin\left(\frac{n\pi(x+3t_{ox})}{W_d+3t_{ox}}\right) dx. \tag{A9.18}$$

$v(x, 0)$ is given by Eqs. (A9.10) and (A9.11). But the boundary condition, Eq. (A9.4), only specifies $\psi(x, 0)=\psi_{bi}$ over $0 \le x \le W_d$. To fill the $\psi(x,0)$ values for the unspecified gap along HA in Fig. A9.1, linear interpolation between $\psi(-3t_{ox},0)=V_{gs} - V_{fb}$ and $\psi(0, 0)=\psi_{bi}$ is used for $-3t_{ox} \le x \le 0$. This is a good approximation if the gap width, $3t_{ox}$, is much smaller than W_d.

In the middle of the device, $y \approx L/2$, the terms in u_L and u_R vary as $\exp\{-n\pi L/[2(W_d+3t_{ox})]\}$. If the channel length L is not too short, the higher-order terms in both series may be neglected. Carrying out the integration in Eq. (A9.18) for $n=1$, one obtains

$$b_1 = \frac{2}{\pi}\left[\left(1+\frac{\sin\theta}{\theta}\right)\psi_{bi} - \left(\frac{\sin\theta}{\theta} + \frac{2\sin\theta}{\pi-\theta} - \frac{2(1+\cos\theta)}{(\pi-\theta)^2}\right)\psi_s\right], \tag{A9.19}$$

where $\theta \equiv \pi(3t_{ox})/(W_d+3t_{ox})$. Since $3t_{ox}$ is thin compared with W_d+3t_{ox}, it is a good approximation to let $\sin\theta \approx \theta$ so that

$$b_1 \approx \frac{4}{\pi}[\psi_{bi} - a\psi_s], \tag{A9.20}$$

where $a \approx 0.4$ for $15° \le \theta \le 45°$. Similarly, the lowest-order coefficient in the u_R series is obtained from the right boundary condition,

$$c_1 \approx \frac{4}{\pi}[(\psi_{bi} + V_{ds}) - a\psi_s]. \tag{A9.21}$$

The third series, u_B, can be neglected altogether because the boundary condition, Eq. (A9.6), over a major part of the bottom boundary (CD in Fig. A9.1) is already satisfied by $v(x, y)$, hence by $\psi(x, y)$. The remaining contributions from segments BC and DE to the coefficients d_n are much smaller than either b_1 or c_1.

An approximate analytical solution under the subthreshold condition is then:

$$\psi(x,y) = \psi_s\left(1-\frac{x}{W_d}\right)^2 + \frac{b_1 \sinh\left(\frac{\pi(L-y)}{W_d+3t_{ox}}\right) + c_1 \sinh\left(\frac{\pi y}{W_d+3t_{ox}}\right)}{\sinh\left(\frac{\pi L}{W_d+3t_{ox}}\right)} \sin\left(\frac{\pi(x+3t_{ox})}{W_d+3t_{ox}}\right) \tag{A9.22}$$

for the silicon region, $0 \le x \le W_d$. It is straightforward to evaluate $\mathscr{E}_y = -\partial\psi/\partial y$ from the above equation and show that it behaves as depicted in Fig. 3.23(a) and (b). The characteristic length of the exponential decay is $(W_d+3t_{ox})/\pi$, which scales with the vertical depth of the rectangular region in Fig. A9.1.

A9.3 Short-Channel Threshold Voltage

To find the threshold voltage of a short-channel device, we consider the potential at the silicon surface, $\psi(0, y)$. It has a minimum value at $y = y_c$, determined by solving $\partial\psi(0, y)/\partial y = 0$ with the approximation $\sinh z \approx e^z/2$ for $z > 0$ (Nguyen, 1984):

$$y_c = \frac{L}{2} - \frac{W_d+3t_{ox}}{2\pi}\ln\left(\frac{c_1}{b_1}\right) \approx \frac{L}{2} - \frac{W_d+3t_{ox}}{2\pi}\ln\left(1+\frac{V_{ds}}{\psi_{bi}-a\psi_s}\right). \tag{A9.23}$$

This point corresponds to the point of maximum barrier height in Fig. 3.21. It is close to the midpoint of the channel when the drain voltage is low. When the drain voltage is high, the point of highest barrier moves closer to the source, as depicted in Fig. 3.20(b).

Under subthreshold conditions, the inversion charge density varies exponentially with the potential at that point. Current conduction is controlled by the lowest surface potential along the channel, i.e., at $y=y_c$.[2] Substituting Eq. (A9.23) into Eq. (A9.22) and letting $x=0$, we obtain

$$\psi(0,y_c)=\psi_s+2\sqrt{b_1c_1}e^{-\frac{\pi L/2}{W_d+3t_{ox}}}\sin\left(\frac{\pi(3t_{ox})}{W_d+3t_{ox}}\right)$$
$$\approx\psi_s+\frac{24t_{ox}}{W_d+3t_{ox}}\left(\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})}-a\psi_s\right)e^{-\frac{\pi L/2}{W_d+3t_{ox}}}. \tag{A9.24}$$

The last result was obtained by substituting Eqs. (A9.20), (A9.21) and applying approximations $\sqrt{(\psi_{bi}-a\psi_s)(\psi_{bi}+V_{ds}-a\psi_s)}\approx\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})}-a\psi_s$ and $\sin[3\pi t_{ox}/(W_d+3t_{ox})]\approx 3\pi t_{ox}/(W_d+3t_{ox})$. In Eq. (A9.24), the first term is the long-channel surface potential. The second term stems from the source-drain boundary conditions. It raises the surface potential and helps a short-channel device reach the threshold criterion, $\psi(0,y_c)=2\psi_B$, with a ψ_s less than the long-channel threshold value, $\psi_s=2\psi_B$. Such a ψ_s reduction, $\Delta\psi_s$, can be translated into a V_t reduction, ΔV_t, using $\Delta V_t/\Delta\psi_s=m=1+3t_{ox}/W_{dm}$ (Eq. (3.27)), where W_{dm} is the maximum depletion width given by Eq. (2.190). The threshold voltage lowering in a short-channel device is therefore,[2]

$$\Delta V_t=\frac{24t_{ox}}{W_{dm}}\left[\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})}-a(2\psi_B)\right]e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}. \tag{A9.25}$$

Note that $a\approx 0.4$ as stated before. The V_{ds} dependence of ΔV_t leads to the DIBL effect.

A9.4 Short-Channel Subthreshold Slope and Substrate Sensitivity

The $-a\psi_s$ term in Eq. (A9.24) degrades the subthreshold slope of a short-channel device, since it reduces the sensitivity of $\psi(0,y_c)$ to variations in ψ_s and therefore to variations in gate voltage. In other words,

$$\frac{d\psi(0,y_c)}{d\psi_s}=1-\frac{24at_{ox}}{W_d+3t_{ox}}e^{-\frac{\pi L/2}{W_d+3t_{ox}}}. \tag{A9.26}$$

Following Eq. (3.41), one can write the inverse subthreshold slope of a short-channel device as

$$S\approx 2.3\frac{mkT}{q}\left[1+\frac{24at_{ox}}{W_{dm}+3t_{ox}}e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}\right]. \tag{A9.27}$$

[2] Rigorously, the current is calculated by integrating $n(x,y)=(n_i^2/N_a)\exp[q\psi(x,y)/kT]$ over x to obtain the sheet density $Q_i(y)/q$, then integrating $1/Q_i(y)$ from $y=0$ to L for the total resistance (Liang and Taur, 2004). The resulting threshold rolloff is more accurate, but rather tedious.

There is no significant degradation of the subthreshold slope from its long-channel value as long as $L\geq 1.5(W_{dm}+3t_{ox})$.

When there is a substrate bias V_{bs} present, one can use the same scheme as in Fig. 3.13 to replace ψ_{bi} and V_{gs} with $\psi_{bi}-V_{bs}$ and $V_{gs}-V_{bs}$, and the threshold condition $2\psi_B$ with $2\psi_B-V_{bs}$. The short-channel threshold rolloff, ΔV_t of Eq. (A9.25), is then expressed as a function of the substrate bias, from which the sensitivity, $d(\Delta V_t)/d(-V_{bs})$, can be evaluated. After linearization, the approximate result is

$$\frac{d(\Delta V_t)}{d(-V_{bs})}\approx\frac{24t_{ox}}{W_{dm}}(1-a)e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}. \tag{A9.28}$$

Note that a reverse substrate bias $V_{bs}<0$ aggravates the short-channel V_t rolloff. Combining Eq. (A9.28) with the long channel substrate sensitivity, Eq. (3.45), we obtain the substrate sensitivity of a short-channel device at $V_{bs}=0$,

$$\frac{dV_t}{d(-V_{bs})}=\frac{3t_{ox}}{W_{dm}}\left[1-8(1-a)e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}\right], \tag{A9.29}$$

which is slightly lower than the long-channel value, $m-1=3t_{ox}/W_{dm}$. This is because some of the substrate depletion charge is coupled to the source and drain instead of to the gate in a short-channel device.

A9.5 Extreme Retrograde Doped (Ground-Plane) MOSFET

A similar analysis can be carried out for an extreme retrograde or ground-plane doping profile instead of the previously assumed uniform doping. Such a doping profile is discussed in Subsection 4.2.3.6, for which the depletion depth equals the undoped layer thickness, independent of gate voltage. The analysis is easier than the uniformly doped case because $v(x,y)$ is simply a continuous linear function of x for the entire region, $-3t_{ox}\leq x\leq W_d$. It is also more accurate because the entire bottom boundary condition, $\psi(W_d,y)=0$ along BE in Fig. A9.1, is satisfied by $v(x,y)$ without the need for the u_B series. The result is similar to Eq. (A9.25) with the parameter a replaced by $m/2$. Since in general, $m/2>a\approx 0.4$, the ground-plane MOSFET has slightly better short-channel effect than the uniformly doped case for the same W_{dm} and t_{ox}.

Appendix 10 Generalized MOSFET Scale Length Model

The scale length model described in Appendix 9 is called the "one-region" model. It replaces the gate oxide with an equivalent region of the same dielectric constant as silicon, but with a thickness equal to $(\varepsilon_{si}/\varepsilon_{ox})t_{ox}$ or $3t_{ox}$. As pointed out earlier, this treats the normal field ($\mathscr{E}_x$) correctly, but the tangential field ($\mathscr{E}_y$) incorrectly. The one-region approximation is valid only if the gate oxide is much thinner than the scale length λ, in which case the oxide field is dominated by its normal component. In 1998, a generalized scale length model was published (Frank *et al.*, 1998) which extended the one-region model to two- and three-regions with arbitrary dielectric constants and thicknesses. It considers the different boundary conditions of the normal and tangential fields separately at the dielectric interfaces. These relations then lead to an eigenvalue equation that can be solved for the scale length λ for such general structures. The generalized scale length model is particularly important for high-κ gate dielectrics which can be physically thick (Section 3.2.1.5), as well as for SOI and double-gate MOSFETs (Section 10.3.2).

A10.1 Two-Region Scale Length Equation

In this appendix, the derivation of a generalized MOSFET scale length is described. Consider the two-region MOSFET model depicted in Fig. A10.1. The gate insulator region is assumed to have a permittivity ε_1 and thickness t_1. The depletion region in the semiconductor has a permittivity ε_2 and thickness t_2. Note that the bottom boundary of the depletion region is simplified to a straight line in the same manner as in Fig. A9.1. In subthreshold, there are negligible mobile carriers in the channel. The electric potential is solved from the 2-D Poisson equation applied to the rectangular region (lightly shaded) in Fig. A10.1. This is a boundary value problem in which the potential on the four conductor sides of the rectangle, left (source), top (gate), right (drain), bottom (body), is specified. There are actually two small gaps not enclosed by conductors: on the top left between the gate and source and on the top right between the gate and drain. When these gaps are not excessively large (compared with, e.g., λ), it is a good approximation to assign potential values by linear interpolation between the gate and source potentials for the left gap and between the gate and drain potentials for the right gap.

With the same superposition approach as in Appendix 9, the potential solution inside the enclosed region, $\psi_1(x, y)$ for region 1 and $\psi_2(x, y)$ for region 2, can be decomposed into several components. The first component is the solution to the 1-D Poisson equation

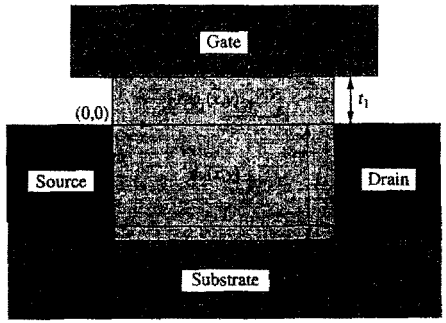


Figure A10.1. Schematic MOSFET diagram for the two-region scale length model.

in the x-direction, which satisfies the inhomogeneous (long channel) equation and the top and bottom boundary conditions. The other two components are solutions to the 2-D homogeneous or Laplace equation (A9.1) and are chosen in order to satisfy the left (source) and right (drain) boundary conditions, respectively. It is shown in Appendix 9 that the latter terms are responsible for short-channel effects. The left (source) and right (drain) components of $\psi_1(x, y)$ can be written as

$$u_{L1}(x,y) = \sum_{n=1}^{\infty} b_{n1} \frac{\sinh\left(\frac{\pi(L-y)}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)} \sin\left(\frac{\pi(x+t_1)}{\lambda_n}\right) \tag{A10.1}$$

and

$$u_{R1}(x,y) = \sum_{n=1}^{\infty} c_{n1} \frac{\sinh\left(\frac{\pi y}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)} \sin\left(\frac{\pi(x+t_1)}{\lambda_n}\right). \tag{A10.2}$$

Note that every term of the u_{L1} and u_{R1} series satisfies Laplace's equation, e.g., $\partial^2 u_{L1}/\partial x^2 + \partial^2 u_{L1}/\partial y^2 = 0$, for any λ_n. Also note that u_{L1} vanishes on the top ($x = -t_1$) and the right boundaries ($y = L$) while u_{R1} vanishes on the top and the left ($y = 0$) boundaries. Likewise, the left (source) and right (drain) components of $\psi_2(x, y)$ are

$$u_{L2}(x,y) = \sum_{n=1}^{\infty} b_{n2} \frac{\sinh\left(\frac{\pi(L-y)}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)} \sin\left(\frac{\pi(x-t_2)}{\lambda_n}\right) \tag{A10.3}$$

and

$$u_{R2}(x,y) = \sum_{n=1}^{\infty} c_{n2} \frac{\sinh\left(\frac{\pi y}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)} \sin\left(\frac{\pi(x-t_2)}{\lambda_n}\right), \tag{A10.4}$$

so that u_{L2} vanishes on the bottom ($x = t_2$) and the right boundaries ($y = L$) while u_{R2} vanishes on the bottom and the left ($y = 0$) boundaries.

At the common boundary shared by both $\psi_1(x, y)$ and $\psi_2(x, y)$, $x = 0$, the normal displacement, $\varepsilon\partial\psi/\partial x$, as well as the potential ψ (hence the tangential field, $\partial\psi/\partial y$) must be continuous from one dielectric medium to the other (Section 2.1.4.2). Because of the different functional forms in y, every term in u_{L1} and $\varepsilon_1\partial u_{L1}/\partial x$ must equal its counterpart in u_{L2} and $\varepsilon_2\partial u_{L2}/\partial x$ at $x = 0$. Therefore, one obtains

$$b_{n1}\sin\left(\frac{\pi t_1}{\lambda_n}\right) = b_{n2}\sin\left(\frac{-\pi t_2}{\lambda_n}\right) \tag{A10.5}$$

and

$$b_{n1}\varepsilon_1\cos\left(\frac{\pi t_1}{\lambda_n}\right) = b_{n2}\varepsilon_2\cos\left(\frac{-\pi t_2}{\lambda_n}\right). \tag{A10.6}$$

Similar relations in terms of c_{n1} and c_{n2} are obtained from u_{R1}, u_{R2}, and their derivatives. For nonzero solutions of b_{n1} and b_{n2}, Eqs. (A10.5) and (A10.6) can be divided to yield an eigenvalue equation for λ_n:

$$\frac{1}{\varepsilon_1}\tan\left(\frac{\pi t_1}{\lambda}\right) + \frac{1}{\varepsilon_2}\tan\left(\frac{\pi t_2}{\lambda}\right) = 0. \tag{A10.7}$$

Here, the eigenvalue is designated as λ which has a discrete series of solutions, $\lambda_1, \lambda_2, \ldots, \lambda_n$, etc., in descending order. The largest of the eigenvalues is of most importance because each term of the short-channel contributions to the potential is proportional to $\exp(-\pi L/2\lambda_n)$ [see Appendix 9, e.g., Eq. (A9.24)]. The entire short-channel part of the potential is dominated by the leading term with the longest λ, i.e., λ_1. Higher order terms in Eqs. (A10.1)–(A10.4) are negligible provided that the channel length L is not comparable to or shorter than λ_1.

For given ε_1, ε_2, t_1, t_2, Eq. (A10.7) can be solved for the scale length λ (used interchangeably with λ_1 throughout the book), which is the two-region equivalent of $W_{dm} + 3t_{ox}$. ***Short channel effects are tolerable if $L/2\lambda > 1$. In other words, $L_{min} \approx 2\lambda$.*** Two special cases are worth discussing. First, for $t_1 << t_2$, approximations $t_1 << \lambda$ and $t_2 \approx \lambda$ can be applied to Eq. (A10.7) to yield $\lambda \approx t_2 + (\varepsilon_2/\varepsilon_1)t_1$, which is of course the result of the one-region approximation in Appendix 9. Second, if $\varepsilon_1 = \varepsilon_2$, the longest λ solution to Eq. (A10.7) is simply $\lambda = t_1 + t_2$, the total physical height of the rectangular box in Fig. A10.1, as might be expected. In general, it helps conceptually to interpret λ as the dielectric equivalent height of the two-region rectangle, so that control of short-channel effect is equivalent to keeping a low aspect ratio of the same rectangle. More about the general solutions to the scale length equation (A10.7) can be found in Section 3.2.1.5 of the main text.

A10.2 Three-Region Scale Length Equation

The two-region scale length model can be further extended to three-region MOSFET structures shown in Fig. A10.2. Following a similar approach as above, the left (source)

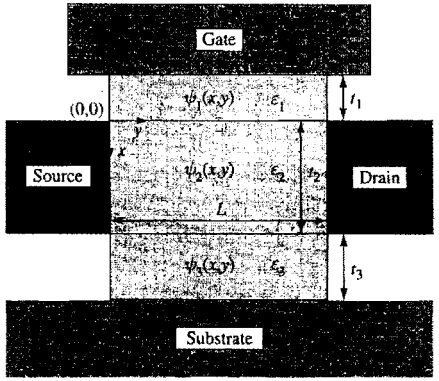


Figure A10.2. Schematic MOSFET diagram for the three-region scale length model.

components of the homogeneous part of the solutions $\psi_1(x, y)$, $\psi_2(x, y)$, $\psi_3(x, y)$ can be written as

$$u_{L1}(x,y) = \sum_{n=1}^{\infty} b_{n1}\frac{\sinh\left(\frac{\pi(L-y)}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)}\sin\left(\frac{\pi(x+t_1)}{\lambda_n}\right) \tag{A10.8}$$

$$u_{L2}(x,y) = \sum_{n=1}^{\infty} b_{n2}\frac{\sinh\left(\frac{\pi(L-y)}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)}\sin\left(\frac{\pi x}{\lambda_n}+\theta\right) \tag{A10.9}$$

$$u_{L3}(x,y) = \sum_{n=1}^{\infty} b_{n3}\frac{\sinh\left(\frac{\pi(L-y)}{\lambda_n}\right)}{\sinh\left(\frac{\pi L}{\lambda_n}\right)}\sin\left(\frac{\pi(x-t_2-t_3)}{\lambda_n}\right). \tag{A10.10}$$

Note that the phases of the sine functions are chosen such that u_{L1} vanishes at the top boundary, and u_{L3} vanishes at the bottom boundary. The phase in u_{L2} consists of an additional variable θ, to be determined later by the boundary conditions. The drain components of the homogeneous solution can be expressed in a form similar to Eqs. (A10.2) and (A10.4). They give redundant results as far as the eigenvalue equation is concerned.

At the boundary between ε_1 and ε_2, $u_{L1}(0, y) = u_{L2}(0, y)$ and $\varepsilon_1\partial u_{L1}/\partial x(0, y) = \varepsilon_2\partial u_{L2}/\partial x(0, y)$. These equalities apply to every pair of the corresponding terms in the two series. The ratio of them then gives

$$\frac{1}{\varepsilon_1}\tan\frac{\pi t_1}{\lambda_n} = \frac{1}{\varepsilon_2}\tan\theta. \tag{A10.11}$$

At the boundary between ε_2 and ε_3, $u_{L2}(t_2, y) = u_{L3}(t_2, y)$ and $\varepsilon_2\partial u_{L2}/\partial x(t_2, y) = \varepsilon_3\partial u_{L3}/\partial x(t_2, y)$. The ratio of the corresponding terms yields

$$\frac{1}{\varepsilon_2}\tan\left(\frac{\pi t_2}{\lambda_n}+\theta\right)=\frac{1}{\varepsilon_3}\tan\left(\frac{-\pi t_3}{\lambda_n}\right). \tag{A10.12}$$

Eliminating θ from Eqs. (A10.11) and (A10.12) results in a three-region eigenvalue equation for λ_n (or λ):

$$\frac{\varepsilon_2}{\varepsilon_1\varepsilon_3}\tan\left(\frac{\pi t_1}{\lambda}\right)\tan\left(\frac{\pi t_2}{\lambda}\right)\tan\left(\frac{\pi t_3}{\lambda}\right)$$
$$=\frac{1}{\varepsilon_1}\tan\left(\frac{\pi t_1}{\lambda}\right)+\frac{1}{\varepsilon_2}\tan\left(\frac{\pi t_2}{\lambda}\right)+\frac{1}{\varepsilon_3}\tan\left(\frac{\pi t_3}{\lambda}\right). \tag{A10.13}$$

The solutions are $\lambda_1, \lambda_2, \ldots, \lambda_n$, etc., in descending order. The longest eigenvalue λ (used interchangeably with λ_1) is the equivalent scale length for the three-region MOSFET structure in Fig. A10.2. In the special case where $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$, it can be shown that the fundamental solution is $\lambda = t_1 + t_2 + t_3$, as expected.

The three-region scale length model is applicable to, e.g., double-gate MOSFETs and SOI devices with very thin buried oxides. It can be applied to other cases as well, noting that while the source and drain are assumed to be in ε_2 in Fig. A10.2, Eq. (A10.13) is equally valid if the source and drain are in ε_3. For example, layer 3 can be the depletion region of bulk silicon where the source and drain also reside and layers 1, 2 represent a composite gate insulator with different dielectric constants. Alternatively, layer 2 can be the sole gate insulator and layer 1 the depletion region of the polysilicon gate. Various applications of the two-region and three-region scale length models scatter throughout the main text.

A10.3 Orthogonality among the Piecewise Eigenfunctions

To evaluate the coefficients $b_{n1}, b_{n2}, c_{n1}, c_{n2}, \ldots$, etc., in the $u_{L1}, u_{L2}, u_{R1}, u_{R2}, \ldots$ series, an orthogonality relationship among the eigenfunctions is needed. For the one-region model in Appendix 9, it is straightforward as the continuous sinusoidal functions in Eq. (A9.17) are self-orthogonal. For the boundary-value problem of a multi-layer dielectric structure like those in Figs A10.1 and A10.2, the eigenfunctions are piecewise continuous and the orthogonality relationship takes a different form. It was discussed in Section 2.1.4.2 that the rigorous form of a 2-D Poisson equation is

$$\frac{\partial}{\partial x}\left(\varepsilon(x)\frac{\partial\psi}{\partial x}\right)+\frac{\partial}{\partial y}\left(\varepsilon(x)\frac{\partial\psi}{\partial y}\right)=0 \tag{A10.14}$$

for zero charge density in all regions. Here $\varepsilon(x)$ is a step function defined by the different permittivities in different regions. For example, for the two-region device model in Fig. A10.1, $\varepsilon(x)=\varepsilon_1$ for $-t_1 \le x \le 0$ and $\varepsilon(x)=\varepsilon_2$ for $0 \le x \le t_2$. The left (source) components of $\psi(x, y)$: u_{L1} of Eq. (A10.1) and u_{L2} of Eq. (A10.3) can be joined into

$$u_L(x,0)=\sum_{n=1}^{\infty}u_{Ln}(x) \tag{A10.15}$$

for $y=0$, where $u_{Ln}(x)$ is a piecewise eigenfunction defined as

$$u_{Ln}(x)=\begin{cases} b_{n1}\sin\left(\dfrac{\pi(x+t_1)}{\lambda_n}\right) & \text{for } -t_1 \le x \le 0 \qquad \text{(A10.16)} \\ b_{n2}\sin\left(\dfrac{\pi(x-t_2)}{\lambda_n}\right) & \text{for } 0 \le x \le t_2. \qquad \text{(A10.17)} \end{cases}$$

Note that b_{n1} and b_{n2} are related by Eq. (A10.5) and λ_n satisfies Eq. (A10.7). It can be mathematically shown that the orthogonality relationship in this case takes the form

$$\int_{-t_1}^{t_2}u_{Ln}(x)u_{Lm}(x)\varepsilon(x)dx=0 \tag{A10.18}$$

for $m \neq n$. Similar results apply to the three-region case as well.

Appendix 11 Drain Current Model of a Ballistic MOSFET

A ballistic MOSFET is a hypothetical device in which the mobile carriers suffer no collisions in the channel. This may happen, in principle, when the channel length is shorter than the mean free path, the average distance carriers travel between collisions. In an ordinary MOSFET, carriers moving from the source to the drain under the influence of the applied field (V_{ds}) collide with the silicon lattice, impurity (dopant) atoms, and surfaces. These collisions limit the velocity they can acquire from the field (Appendix 3), resulting in a reduced drain current. Under low field conditions, the effect of these collisions is lumped into a mobility factor proportional to the mean free time between successive collisions (Appendix 3). For long-channel MOSFETs, the drain current is proportional to the mobility (Section 3.1.2). For short-channel MOSFETs under high drain bias conditions, high-field scattering becomes important. This is usually modeled by velocity saturation (Section 3.2.2). In the absence of any scattering, carriers entering the channel from the source are accelerated by the applied field ballistically toward the drain. They can attain very high speeds especially in the high-field region near the drain. However, such high speeds (velocity overshoot) do not necessarily translate into large currents. Since current must be continuous from source to drain, it is bounded by the rate at which carriers are injected from the source. In a ballistic MOSFET then, the bottleneck is near the source where carriers move into the channel at relatively low velocities (before field acceleration). Current continuity is satisfied by a decreased carrier density near the drain such that the product of carrier density and velocity at the drain is the same as that at the source (see Fig. 3.31). This appendix describes the drain current model for a ballistic MOSFET published by Natori in 1994 (Natori, 1994).

A11.1 Source-Drain Current in a Ballistic MOSFET

The key to modeling the drain current of a ballistic MOSFET is to consider the average carrier velocity moving into the channel in the low field region near the source (Natori, 1994). In fact, as shown in the MOSFET band diagram in Fig. A11.1, there is typically a point of zero field near the source where the electron energy barrier is the highest. The conduction band energy at this point is designated as E_c'. Quasi-equilibrium condition holds at this point since there is no net force acting on the carriers. One expects the electron distribution function to take a Fermi–Dirac-like form here. Without any collisions in the channel, all the electrons moving from left to right ($v_y > 0$) at this point will eventually make it to the drain. This constitutes the positive component of the current. Conversely, the

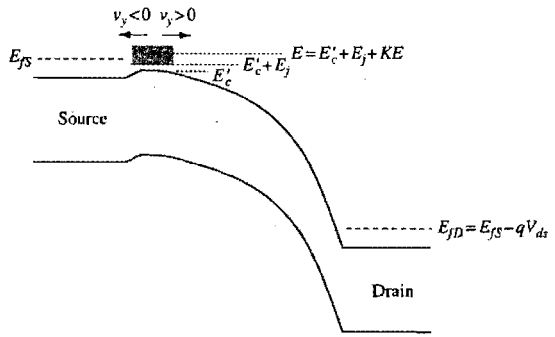


Figure A11.1. A schematic MOSFET band diagram under high-drain bias conditions. Dashed lines indicate the Fermi levels of the degenerately doped source and drain. At the point of highest energy barrier near the source, electrons populate states allowed in discrete subbands.

electrons moving from right to left ($v_y < 0$) at this point must have all originated from the drain, which make up the negative component of the current. Therefore, the left to right ($v_y > 0$) electron population is in quasi-equilibrium with the source and their distribution function is controlled by the source Fermi level E_{fS}, while the right to left ($v_y < 0$) electron distribution function is controlled by the drain Fermi level E_{fD} ($= E_{fS} - qV_{ds}$).

It is discussed in Appendix 12 that electrons confined in an inversion layer populate discrete subbands with a minimum energy E_j (jth subband) above the conduction band E_c' (Fig. A11.1). Consider electrons in the jth subband moving from left to right with a velocity between (v_y, v_z) and ($v_y + dv_y$, $v_z + dv_z$), where z is in the device width direction. From Eq. (A12.1), the electronic states per unit area is $(2g/h^2)m_y m_z dv_y dv_z$. The probability of each state being populated by an electron is given by the Fermi–Dirac distribution function, Eq. (2.4), with $E = E_c' + E_j + KE = E_c' + E_j + (1/2)m_y v_y^2 + (1/2)m_z v_z^2$ and $E_f = E_{fS}$. The charge density per unit area of these electrons is then

$$dQ_{j+} = \frac{(2g/h^2)q m_y m_z dv_y dv_z}{1 + e^{(E_c' + E_j + m_y v_y^2/2 + m_z v_z^2/2 - E_{fS})/kT}}. \tag{A11.1}$$

Since current per width equals the product of charge density and carrier velocity, the left-to-right current component is given by the summation over all subbands of the following double integral:

$$\begin{aligned} I_+ &= W \sum_j \int_{all\ v_z} \int_{v_y > 0} v_y dQ_{j+} \\ &= \sum_j \int_{-\infty}^{\infty} \int_0^{\infty} \frac{(2g/h^2) q W m_y m_z v_y}{1 + e^{(E_c' + E_j + m_y v_y^2/2 + m_z v_z^2/2 - E_{fS})/kT}} dv_y dv_z. \end{aligned} \tag{A11.2}$$

The integration with respect to v_y can be done analytically by a simple change of variable, $u = v_y^2$. Using integration by parts, the result of the second integration can be expressed in terms of the Fermi–Dirac integral,

$$F_{1/2}(u) \equiv \int_0^\infty \frac{\sqrt{y}dy}{1+e^{y-u}}, \tag{A11.3}$$

as

$$I_+ = \sum_j (4g/h^2) qW\sqrt{2m_z}(kT)^{3/2} F_{1/2}\left(\frac{E_{fS} - E_c' - E_j}{kT}\right). \tag{A11.4}$$

Likewise, the right-to-left current component is obtained by replacing E_{fS} with E_{fD},

$$I_- = \sum_j (4g/h^2) qW\sqrt{2m_z}(kT)^{3/2} F_{1/2}\left(\frac{E_{fD} - E_c' - E_j}{kT}\right). \tag{A11.5}$$

The full expression for the net drain to source current, $I_{ds} = I_+ - I_-$, is of the following form:

$$I_{ds} = \frac{4\sqrt{2}qW(kT)^{3/2}}{h^2}\left\{g\sqrt{m_z}\sum_j\left[F_{1/2}\left(\frac{E_{fS} - E_c' - E_j}{kT}\right) - F_{1/2}\left(\frac{E_{fS} - qV_{ds} - E_c' - E_j}{kT}\right)\right]\right.$$
$$\left. + g'\sqrt{m_z'}\sum_j\left[F_{1/2}\left(\frac{E_{fS} - E_c' - E_j'}{kT}\right) - F_{1/2}\left(\frac{E_{fS} - qV_{ds} - E_c' - E_j'}{kT}\right)\right]\right\}. \tag{A11.6}$$

Here, for silicon in the ⟨100⟩ direction, $g = 2$, $m_z = m_t$ for the unprimed valley, and $g' = 4$, $m_z' = (\sqrt{m_l} + \sqrt{m_t})^2/4$ for the primed valley. The unknowns in Eq. (A11.6) are $E_{fS} - E_c' - E_j$ and $E_{fS} - E_c' - E_j'$, the relative position of the various subbands at the point of highest electron barrier with respect to the source Fermi level. They are controlled by the gate voltage and, in general, must be solved numerically from the coupled Poisson's and Schrödinger's equations (Section 4.2.4).

With the same partition of charge into the positive and negative directions, the total integrated density of inversion charge per unit area can be expressed as

$$Q_i = Q_+ + Q_- = \sum_j \int_{\text{all } v_z} \int_{v_y>0} dQ_{j+} + \sum_j \int_{\text{all } v_z} \int_{v_y<0} dQ_{j-}. \tag{A11.7}$$

After carrying out the integrations,[1] an expression similar to Eq. (A12.7) is obtained:

$$Q_i = \frac{2\pi qkT}{h^2}\left\{gm_t\sum_j\left[\ln\left(1 + e^{(E_{fS} - E_c' - E_j)/kT}\right) + \ln\left(1 + e^{(E_{fS} - qV_{ds} - E_c' - E_j)/kT}\right)\right]\right.$$
$$\left. + g'\sqrt{m_l m_t}\sum_j\left[\ln\left(1 + e^{(E_{fS} - E_c' - E_j')/kT}\right) + \ln\left(1 + e^{(E_{fS} - qV_{ds} - E_c' - E_j')/kT}\right)\right]\right\}. \tag{A11.8}$$

[1] The double integration can be carried out by transforming v_y, v_z to elliptical-polar coordinates, as in Section A12.1.

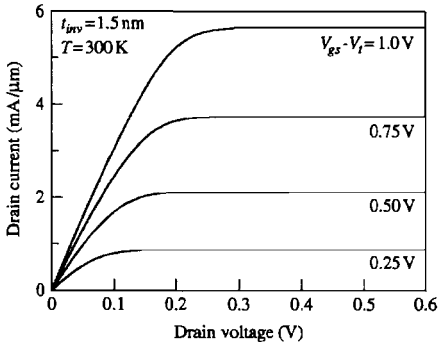


Figure A11.2. I_{ds}–V_{ds} curves of a ballistic MOSFET calculated from Eqs. (A11.9) and (A11.10). Here $C_{inv} = \varepsilon_{ox}/t_{inv}$.

On the other hand, inversion charge density at the source can be solved from electrostatics, e.g., Eq. (3.14) with $V = 0$, as a function of the gate voltage. Above the threshold voltage V_t, Q_i varies approximately linearly with V_{gs} and can be expressed as $C_{inv}(V_{gs} - V_t)$, where C_{inv} is the effective gate capacitance per unit area which lumps together oxide thickness, inversion layer depth (with quantum effect, Section 4.2.4.3), and polysilicon gate depletion effects.

A11.2 One Subband Approximation

In the case of strong quantum effects when either the temperature is low or the field is high, the spacing between E_j of successive subbands becomes larger than kT. It is a good approximation to keep only the lowest subband term, i.e., $j = 0$ of the unprimed valley (Section 4.2.4.1). Equating Eq. (A11.8) to $C_{inv}(V_{gs} - V_t)$ then allows $E_{fS} - E_c' - E_0$ to be solved analytically:

$$e^{(E_{fS} - E_c' - E_0)/kT} = \frac{1}{2}\left\{\sqrt{\left(e^{qV_{ds}/kT} - 1\right)^2 + 4\exp\left[\frac{h^2 C_{inv}(V_{gs} - V_t)}{4\pi qkTm_t} + \frac{qV_{ds}}{kT}\right]}\right.$$
$$\left. - 1 - e^{qV_{ds}/kT}\right\}. \tag{A11.9}$$

Under the same one subband approximation, Eq. (A11.6) is reduced to

$$I_{ds} = \frac{8\sqrt{2m_t}qW(kT)^{3/2}}{h^2}\left[F_{1/2}\left(\frac{E_{fS} - E_c' - E_0}{kT}\right)\right.$$
$$\left. - F_{1/2}\left(\frac{E_{fS} - E_c' - E_0}{kT} - \frac{qV_{ds}}{kT}\right)\right]. \tag{A11.10}$$

For given values of V_{gs} and V_{ds}, the current of a ballistic MOSFET can be calculated from Eqs. (A11.9) and (A11.10). An example is given in Fig. A11.2.[2] Obviously, $I_{ds}=0$ when $V_{ds}=0$ because $E_{fS}=E_{fD}$ and the positive moving ($v_y>0$ in Fig. A11.1) electron flux from the source exactly cancels the negative moving ($v_y<0$) electron flux from the drain. As V_{ds} increases, I_{ds} increases because the negative moving electron population decreases with decreasing $E_{fD}=E_{fS}-qV_{ds}$, while the positive moving electron population increases in order to conserve the total Q_i. The current saturates at a drain voltage V_{dsat} when the negative moving electron flux at the point of maximum barrier in Fig. A11.1 becomes negligibly small. While the drain current of a ballistic MOSFET exhibits saturation just like that of an ordinary MOSFET, the underlying physics is completely different. In an ordinary MOSFET, the drain current saturates due to either pinch-off or velocity saturation at the drain (Chapter 3). The current of a ballistic MOSFET saturates because for a given electron density, there is a maximum electron flux that can be extracted from the source, given by the positive half of the thermal distribution. Also note that the current of a ballistic MOSFET is independent of channel length, as it represents the highest current limit when $L \rightarrow 0$.

Further analytical expressions of V_{dsat} and I_{dsat} of a ballistic MOSFET can be derived under the *degenerate* condition in which the source Fermi level is at least a few kT above the minimum energy of the lowest subband, i.e., when $(E_{fS}-E_c'-E_0)/kT>1$. This happens at relatively high gate voltages when the dimensionless parameter in Eq. (A11.9), $h^2C_{inv}(V_{gs}-V_t)/(4\pi qkTm_t)$, is greater than one, i.e., when the electron sheet density, $C_{inv}(V_{gs}-V_t)/q$, is higher than half of the 2-D effective density of states, $4\pi kTm_t/h^2 \approx 2\times10^{12}\ \text{cm}^{-2}$. For example, for $t_{inv}=1.5$ nm and $V_{gs}-V_t=1$ V, $C_{inv}(V_{gs}-V_t)/q$ is approximately $1.3\times10^{13}\ \text{cm}^{-2}$. Fermi–Dirac integrals can then be approximated by $F_{1/2}(u)\approx(2/3)u^{3/2}$ for $u>>1$ (Blakemore, 1982).

In saturation, $qV_{ds}/kT>>1$, Eq. (A11.9) gives $(E_{fS}-E_c'-E_0)/kT\approx h^2C_{inv}(V_{gs}-V_t)/(4\pi qkTm_t)$. Substituting in Eq. (A11.10) yields the saturation current,

$$I_{dsat}=\frac{2\sqrt{2}hW}{3\pi\sqrt{\pi\hbar}m_t}\left[C_{inv}(V_{gs}-V_t)\right]^{3/2}. \tag{A11.11}$$

In the linear region, $qV_{ds}/kT<<1$, Eq. (A11.9) gives $(E_{fS}-E_c'-E_0)/kT\approx h^2C_{inv}(V_{gs}-V_t)/(8\pi qkTm_t)$. Substituting in Eq. (A11.10) yields the linear region conductance,

$$I_{dlin}/V_{ds}=\frac{4q^2W}{\sqrt{\pi\hbar}h}\sqrt{C_{inv}(V_{gs}-V_t)}. \tag{A11.12}$$

Note that the ballistic-limited conductance is independent of mobility and channel length. The saturation voltage can be estimated as the V_{ds} value for which $I_{dlin}=I_{dsat}$:

$$V_{dsat}=\frac{h^2C_{inv}(V_{gs}-V_t)}{3\sqrt{2}\pi q^2m_t}=\frac{h^2\varepsilon_{ox}/(3\sqrt{2}\pi q^2m_t)}{t_{inv}}(V_{gs}-V_t). \tag{A11.13}$$

[2] An explicit analytic function that approximates Fermi–Dirac integrals numerically is used (Blakemore, 1982).

Note that while $V_{dsat}\approx V_{gs}-V_t$ for a conventional long-channel MOSFET, V_{dsat} of a ballistic MOSFET is much smaller than $V_{gs}-V_t$ since in practice the effective gate oxide (t_{inv}) is much thicker than the constant, $h^2\varepsilon_{ox}/(3\sqrt{2}\pi q^2m_t)\approx 0.26$ nm, in Eq. (A11.13). It is also worth noting that both I_{dsat} and V_{dsat} are independent of temperature in the degenerate limit. In fact, they essentially equal the values at 0 K.

Note from above that $(E_{fS}-E_c'-E_0)/kT$ at high V_{ds} increases by a factor of two over the static value at $V_{ds}=0$. This means lowering of the energy barrier with respect to the source Fermi level in order to populate the higher positive-v_y states when the negative-v_y states are not occupied. While such a potential change in silicon consumes a part of the gate overdrive [see the gate bias equation (2.195)], the implicit assumption here is that it causes a negligible loss on the inversion charge density. In other words, the "density of states" capacitance, $C_{DoS}=8\pi m_t q^2/h^2$, is much larger than C_{inv}. Here C_{DoS} is defined as the charge of the filled electronic states per unit area per unit potential, which equals $q^2N(E)$ where $N(E)$ is the 2-D density of states given by Eq. (A12.3). It is a good approximation for silicon in which C_{DoS} is equivalent to that of a 1.3 Å thick oxide. For high-mobility semiconductors with light effective mass, C_{DoS} cannot be ignored and must be taken into account when evaluating Q_i under ballistic conditions (Rahman *et al.*, 2003).

A related assumption in treating Q_i to be the electrostatic value $C_{inv}(V_{gs}-V_t)$ is that the source region acts like an infinite reservoir of carriers maintained in quasi-thermal equilibrium. In practice, however, source doping is limited by technology factors and scattering processes may not ensure an equilibrium carrier distribution in ballistic conditions. With insufficient collisions, the states with momenta perpendicular to the channel direction are underpopulated. This means that additional potential drop in the semiconductor is needed to populate states with even higher v_y. It acts like a reduction in C_{DoS} which could result in a significantly lower inversion charge density than the electrostatic value in long-channel devices. Furthermore, if the source is not doped high enough to sufficiently supply carriers in the longitudinal momentum states, Q_i decreases even further owing to "source starvation" (Fischetti *et al.*, 2007).

Appendix 12 Quantum-Mechanical Solution in Weak Inversion

In this appendix, we derive the expressions for the 2DEG (two-dimensional electron gas) density of states and inversion charge density used in Section 4.2.4. A quantum-mechanical treatment of inversion layer is necessary because of the confinement of electron motion in the direction normal to the surface (Stern and Howard, 1967). Note that MKS units are used throughout this appendix (e.g., length must be in meters, not centimeters).

A12.1 2-D Density of States

Following an approach in parallel with that of Section 2.1.1.2 for the 3-D density of states, one can derive an expression for the density of states of a 2-D gas. Based on quantum mechanics, there is one allowed state in a phase space of volume $(\Delta y \Delta p_y) \times (\Delta z\, \Delta p_z) = h^2$, where p_y and p_z are the y- and z-components of the electron momentum and h is Planck's constant. If we let $N(E)dE$ be the number of electronic states per unit area with an energy between E and $E + dE$, then

$$N(E)dE = 2g\frac{dp_y dp_z}{h^2}, \tag{A12.1}$$

where $dp_y\, dp_z$ is the area in the momentum space within which the electron energy lies between E and $E + dE$, g is the degeneracy of the subband, and the factor of two arises from the two possible directions of electron spin.

If E_{min} is the ground-state energy of a particular subband, the energy–momentum relationship near the bottom of that subband is

$$E - E_{min} = \frac{p_y^2}{2m_y} + \frac{p_z^2}{2m_z}, \tag{A12.2}$$

where $E - E_{min}$ is the electron kinetic energy, and m_y, m_z are the effective masses. The area of the ellipse given by Eq. (A12.2) in momentum space is $2\pi(m_y m_z)^{1/2}(E - E_{min})$. Therefore, the area $dp_y\, dp_z$ within which the electron energy lies between E and $E + dE$ is $2\pi(m_y m_z)^{1/2}dE$ and Eq. (A12.1) becomes

$$N(E)dE = \frac{4\pi g\sqrt{m_y m_z}}{h^2}dE. \tag{A12.3}$$

The number of electrons per unit area in this subband is then given by

$$n = \int_{E_{min}}^{\infty} N(E) f_D(E) dE, \tag{A12.4}$$

where $f_D(E)$ is the Fermi–Dirac distribution function, Eq. (2.4). Since $N(E)$ is a constant and can be taken out of the integral, Eq. (A12.4) can be easily integrated to yield

$$n = \frac{4\pi g k T\sqrt{m_y m_z}}{h^2}\ln\left[1 + e^{(E_f - E_{min})/kT}\right]. \tag{A12.5}$$

A12.2 QM Inversion Charge Density

Figure A12.1 shows the energy-band diagram of a quantized inversion layer at the silicon surface, where the bottom of the conduction band at an energy E_c' is lower than the conduction band E_c in the bulk by an amount $q\psi_s$ due to the applied field from the gate. Here ψ_s is the surface potential or band bending described in Section 2.3.2. In the subthreshold region, the potential function can be approximated by a triangular well of slope $\mathscr{E}_s$ equal to the surface electric field the electrons experience. The quantized subband energies for the ⟨100⟩ direction are then represented by E_j ($g = 2$, $m_x = m_l = 0.92m_0$) and E_j' ($g' = 4$, $m_x' = m_t = 0.19m_0$) above the conduction band at the surface, where E_j and E_j' are given by Eqs. (4.46) and (4.48) in Section 4.2.4. For the jth sub-band, the minimum energy is

$$E_{min} = E_c' + E_j = E_c - q\psi_s + E_j. \tag{A12.6}$$

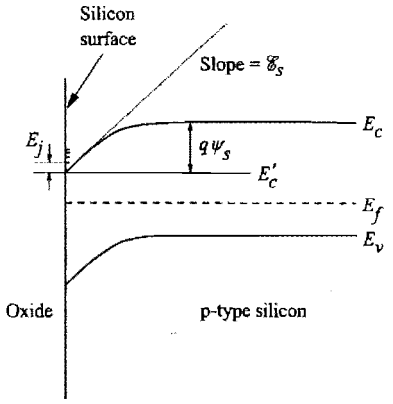


Figure A12.1. Schematic band diagram showing band bending in the subthreshold region and quantized electron energy levels in the inversion layer at the silicon surface.

Summing over all the subbands in both valleys using Eq. (A12.5), one obtains the total inversion charge density per unit area (Stern and Howard, 1967):

$$Q_i^{QM} = \frac{4\pi qkT}{h^2}\left\{gm_t\sum_j \ln\left[1+e^{(E_f-E_c-E_j+q\psi_s)/kT}\right] + g'\sqrt{m_t m_l}\sum_j \ln\left[1+e^{(E_f-E_c-E_j'+q\psi_s)/kT}\right]\right\}. \quad \text{(A12.7)}$$

Note that for the first group of subbands in Eq. (A12.7), the confinement (x) is in the longitudinal direction and the density-of-state effective mass $(m_y m_z)^{1/2}$ is $(m_t m_t)^{1/2}$ or m_t. For the second group of subbands in Eq. (A12.7), the confinement (x) is perpendicular to the longitudinal direction and the density-of-state effective mass $(m_y m_z)^{1/2}$ is $(m_t m_l)^{1/2}$. In the subthreshold region, the Fermi level is at least a few kT below the lowest subband of energy $E_c - q\psi_s + E_0$, and the factors $\ln(1 + e^x)$ can be approximated by e^x in both terms of Eq. (A12.7). Furthermore, from Eq. (2.9), $E_f - E_c$ in the bulk silicon is equal to $kT\ln(n/N_c)$ or $kT\ln[n_i^2/(N_a N_c)]$, where N_c is the effective density of states of the conduction band, and $n = n_i^2/N_a$ is the equilibrium electron concentration in the bulk silicon. For a nonuniformly doped substrate, N_a refers to the p-type concentration at the edge of the depletion layer. Equation (A12.7) is then simplified to

$$Q_i^{QM} = \frac{4\pi qkTn_i^2}{h^2 N_c N_a}\left(2m_t\sum_j e^{-E_j/kT} + 4\sqrt{m_t m_l}\sum_j e^{-E_j'/kT}\right)e^{q\psi_s/kT}, \quad \text{(A12.8)}$$

where $g = 2$ and $g' = 4$ have been substituted.

A12.3 Convergence of the QM Solution at Low Fields to the 3-D Continuum Case

The energy levels of the lower valley are given by Eq. (4.46) (Stern, 1972):

$$E_j = \left[\frac{3hq\mathscr{E}_s}{4\sqrt{2m_l}}\left(j+\frac{3}{4}\right)\right]^{2/3}, \quad j = 0, 1, 2, \ldots. \quad \text{(A12.9)}$$

When the surface field is low ($\mathscr{E}_s < 10^4$ V/cm at room temperature), the spacings between the quantized energy levels are small compared with kT and the 2-D quantum effect is weak. In this case, the serial summations in Eq. (A12.8) can be replaced by integrals using the identity

$$\sum_n e^{-(n\Delta y)^{2/3}}\Delta y = \int_0^\infty e^{-y^{2/3}}dy \quad \text{(A12.10)}$$

in the limit of $\Delta y \to 0$. By a simple change of variable ($u = y^{1/3}$), the integral on the RHS of Eq. (A12.10) can be easily converted to a gamma function, whose value is $3\pi^{1/2}/4$. Therefore,

$$\sum_j e^{-E_j/kT} \to \frac{3\sqrt{\pi}}{4}\left(\frac{4\sqrt{2m_l}(kT)^{3/2}}{3hq\mathscr{E}_s}\right) \quad \text{(A12.11)}$$

in the low-field limit. Now Eq. (A12.8) can be evaluated as

$$Q_i^{QM} = \frac{4\pi qkTn_i^2}{h^2 N_c N_a}\left[2m_t\frac{\sqrt{2\pi m_l}(kT)^{3/2}}{hq\mathscr{E}_s} + 4\sqrt{m_t m_l}\frac{\sqrt{2\pi m_t}(kT)^{3/2}}{hq\mathscr{E}_s}\right]e^{q\psi_s/kT}. \quad \text{(A12.12)}$$

Substituting N_c from Eq. (2.10) with the 3-D degeneracy factor $g = 6$ into Eq. (A12.12) yields

$$Q_i^{QM} = \frac{kTn_i^2}{\mathscr{E}_s N_a}e^{q\psi_s/kT}. \quad \text{(A12.13)}$$

This is the same as Eq. (3.36) for the 3-D inversion charge density per unit area in the subthreshold region. In other words, when the surface field is low and/or the temperature is high, the 2-D quantum solution converges to the 3-D continuum case. At high fields, however, Q_i^{QM} is lower than the 3-D inversion charge density for the same band bending, which results in a higher threshold voltage than predicted by the classical model, as discussed in Section 4.2.4.

Appendix 13 Power Gain of a Two-Port Network

In the small-signal linear analysis in the frequency domain, a two-port network (Fig. A13.1) is often represented by an admittance matrix:

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = [Y]\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}. \tag{A13.1}$$

To derive an expression for the maximum available gain, the direction of power flow needs to be specified, for example, from port 1 to port 2. Assume an output termination Y_L, then

$$i_2 + Y_L v_2 = 0, \tag{A13.2}$$

where $\mathrm{Re}(Y_L) > 0$, i.e., Y_L is passive. If $v_1 = |v_1|e^{j(\omega t+\alpha)}$, $i_1 = |i_1|e^{j(\omega t+\beta)}$, then the power input at port 1 is the time average of $|v_1|\cos(\omega t+\alpha) \times |i_1|\cos(\omega t+\beta)$, which equals $\frac{1}{2}|v_1||i_1|\cos(\alpha-\beta)$ or $\frac{1}{2}\mathrm{Re}(v_1 i_1^*)$. Likewise, the power output delivered at port 2 is $\frac{1}{2}\mathrm{Re}(-v_2 i_2^*)$. Combining Eqs. (A13.1) and (A13.2) to express v_1, v_2, i_2 in terms of i_1, one can evaluate the power gain:

$$G \equiv \frac{\mathrm{Re}(-v_2 i_2^*)}{\mathrm{Re}(v_1 i_1^*)} = \frac{|Y_{21}|^2 \mathrm{Re}(Y_L)}{|Y_{22}+Y_L|^2 \mathrm{Re}(Y_{11}) - \mathrm{Re}\left[Y_{12}^* Y_{21}^* (Y_{22}+Y_L)\right]}. \tag{A13.3}$$

Next, Y_L ($\equiv x+jy$, $x>0$) is varied to maximize G. Let $Y_{11} \equiv G_{11}+jB_{11}$ and $Y_{22} \equiv G_{22}+jB_{22}$, then

$$G = \frac{|Y_{21}|^2 x}{\left[(G_{22}+x)^2 + (B_{22}+y)^2\right]G_{11} - (G_{22}+x)\mathrm{Re}(Y_{12}Y_{21}) - (B_{22}+y)\mathrm{Im}(Y_{12}Y_{21})}. \tag{A13.4}$$

Here, the denominator is a quadratic function of x and y. Note that if either G_{11} or G_{22} is negative, the power gain G can be either negative (meaning oscillations)[1] or unbounded (infinite gain) for some choices of $x>0$ and y. If both G_{11} and G_{22} are positive, and if

$$2G_{11}G_{22} > \mathrm{Re}(Y_{12}Y_{21}) + |Y_{12}Y_{21}|, \tag{A13.5}$$

then the denominator in Eq. (A13.4) is positive definite (for $x>0$).[2] From basic algebra, an optimum set of $x>0$ and y can be chosen such that G reaches its maximum value, the maximum available gain,

[1] Since $\mathrm{Re}(-v_2 i_2^*) = \mathrm{Re}(Y_L^* |v_2|^2) > 0$, if $G<0$, then $\mathrm{Re}(v_1 i_1^*) < 0$, i.e., the two-port network is delivering power out of both port 2 and port 1.

[2] If Eq. (A13.5) is not satisfied, then there exist values of $x>0$, y such that the denominator in Eq. (A13.4) goes to zero, i.e., the power gain is unbounded. It can be shown that this is a stronger condition than $G_{\max} = 1$.

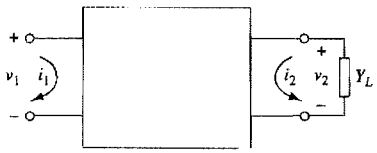


Figure A13.1. Current and voltage definitions at the input and output of a two-port network. The output is assumed to be terminated by an admittance Y_L.

$$G_{\max} = \frac{|Y_{21}|^2}{2G_{11}G_{22} - \mathrm{Re}(Y_{12}Y_{21}) + \sqrt{[2G_{11}G_{22} - \mathrm{Re}(Y_{12}Y_{21})]^2 - |Y_{12}Y_{21}|^2}}. \tag{A13.6}$$

The power gain criterion is $G_{\max} = 1$, which, after rearranging the terms and squaring to remove the square root, can be shown to yield

$$|Y_{21}|^2\left\{|Y_{21}|^2 + |Y_{12}|^2 - 2[2G_{11}G_{22} - \mathrm{Re}(Y_{12}Y_{21})]\right\} = 0. \tag{A13.7}$$

Obviously, $|Y_{21}| \neq 0$, for otherwise $G_{\max} = 0$. Therefore,

$$4G_{11}G_{22} = |Y_{12}|^2 + |Y_{21}|^2 + 2\mathrm{Re}(Y_{12}Y_{21}) = |Y_{12} + Y_{21}^*|^2. \tag{A13.8}$$

Equation (A13.8) is a necessary condition for $G_{\max} = 1$, but not sufficient. If Eq. (A13.8) is substituted into Eq. (A13.5), the square-root term in the denominator becomes $\frac{1}{2}|\,|Y_{12}|^2 - |Y_{21}|^2|$, which can take one or the other form depending on whether $|Y_{21}/Y_{12}| > 1$ or < 1. $G_{\max} = 1$ is obtained only if $|Y_{21}/Y_{12}| > 1$. Otherwise, $|Y_{21}/Y_{12}| < 1$, and $G_{\max} = |Y_{21}/Y_{12}|^2 < 1$. In other words, given Eq. (A13.8) which is symmetric with respect to indices 1 and 2, power gain is possible only if the circuit is used in the right direction. The criteria for power gain in a two-port network, i.e., $G_{\max} > 1$ is then $|Y_{21}/Y_{12}| > 1$ and

$$|Y_{12} + Y_{21}^*|^2 > 4G_{11}G_{22}. \tag{A13.9}$$

The above condition in terms of the matrix elements is used in Appendices 14 and 18 to derive the unity power gain frequency, $f_{\max}$, of transistors.

In 1954, Mason derived a unilateral power gain for linear two-port networks (Mason, 1954),

$$U \equiv \frac{|Y_{12} - Y_{21}|^2}{4[G_{11}G_{22} - \mathrm{Re}(Y_{12})\mathrm{Re}(Y_{21})]}. \tag{A13.10}$$

It is the maximum power gain obtained after the two-port network is "unilateralized" by means of lossless reciprocal embedding. "Unilateralized" means that after the transformation, $Y'_{12} = 0$ in the new admittance matrix so that there is no feedback coupling from port 2 to port 1. Since U is invariant under all lossless reciprocal embedding of

the two-port, it serves as a useful figure of merit for transistors. $U > 1$ means that the transistor is capable of delivering power gain at that frequency.

With the help of the identity, $|Y_{12}+Y_{21}^*|^2 = |Y_{12}-Y_{21}|^2+4\mathrm{Re}(Y_{12})\mathrm{Re}(Y_{21})$, it is readily shown that the condition $U=1$ is exactly the same as Eq. (A13.8), the condition for $G_{max}=1$. While U is the maximum power gain under Mason's lossless reciprocal embeddings, it is not the highest gain with lossy terminations. Often one finds $G_{max} > U$ under the power gain conditions, Eq. (A13.9) and $|Y_{21}/Y_{12}| > 1$ (Gupta, 1992).

Appendix 14 Unity-Gain Frequencies of a MOSFET Transistor

In this appendix, full expressions for the figure of merit of a MOSFET transistor in RF circuits, f_T (unity current gain frequency or current gain cutoff frequency) and f_{max} (unity power gain frequency or maximum oscillation frequency), are derived. Adding parasitic resistances R_g, R_s, R_d to the intrinsic MOSFET in Fig. 5.46, one obtains the equivalent circuit of an extrinsic MOSFET transistor in Fig. A14.1. The convention here is that the primed parameters refer to the intrinsic device, while the unprimed parameters are for the extrinsic device as measured externally. Also, v_1', i_1 represent v_{gs}', i_{gs}, and v_2', i_2 represent v_{ds}', i_{ds}, respectively.

The intrinsic admittance matrix of a MOSFET, Eq. (5.52), is expressed in the new notation as:

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = [\tilde{Y}'] \begin{bmatrix} v_1' \\ v_2' \end{bmatrix} = \begin{bmatrix} Y_{11}' & Y_{12}' \\ Y_{21}' & Y_{22}' \end{bmatrix} \begin{bmatrix} v_1' \\ v_2' \end{bmatrix} = \begin{bmatrix} j\omega(C_{gs}+C_{gd}) & -j\omega C_{gd} \\ g_m' - j\omega C_{gd} & g_{ds}' + j\omega C_{db} + j\omega C_{gd} \end{bmatrix} \begin{bmatrix} v_1' \\ v_2' \end{bmatrix}. \tag{A14.1}$$

The extrinsic small-signal voltages, v_1 ($=v_{gs}$) and v_2 ($=v_{ds}$), are related to the intrinsic voltages v_1', v_2' by the following matrix equation:

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1' \\ v_2' \end{bmatrix} + \begin{bmatrix} R_g+R_s & R_s \\ R_s & R_d+R_s \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \left\{ [\tilde{Y}']^{-1} + \begin{bmatrix} R_g+R_s & R_s \\ R_s & R_d+R_s \end{bmatrix} \right\} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}, \tag{A14.2}$$

where the inverse of Eq. (A14.1) has been used in the second step. This equation can be inverted to obtain the extrinsic admittance matrix, $[\tilde{Y}]$,

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = [\tilde{Y}] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = [\tilde{Y}'] \left\{ 1 + \begin{bmatrix} R_g+R_s & R_s \\ R_s & R_d+R_s \end{bmatrix} [\tilde{Y}'] \right\}^{-1} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}. \tag{A14.3}$$

If the parasitic resistances are not excessively large, $[\tilde{Y}]$ can be expressed to the first order of R_s, R_d, R_g as

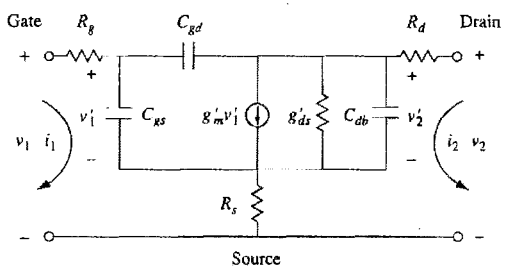


Figure A14.1. Small-signal equivalent circuit of an extrinsic MOSFET.

$$[\tilde{Y}] = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = [\tilde{Y}'] - [\tilde{Y}']\begin{bmatrix} R_g + R_s & R_s \\ R_s & R_d + R_s \end{bmatrix}[\tilde{Y}'] =$$
$$\begin{bmatrix} Y'_{11} - Y'^2_{11}R_g - Y'_{12}Y'_{21}R_d - (Y'_{11} + Y'_{12})(Y'_{11} + Y'_{21})R_s \\ Y'_{21} - Y'_{11}Y'_{21}R_g - Y'_{21}Y'_{22}R_d - (Y'_{21} + Y'_{11})(Y'_{21} + Y'_{22})R_s \\ Y'_{12} - Y'_{11}Y'_{12}R_g - Y'_{12}Y'_{22}R_d - (Y'_{12} + Y'_{11})(Y'_{12} + Y'_{22})R_s \\ Y'_{22} - Y'^2_{22}R_d - Y'_{12}Y'_{21}R_g - (Y'_{22} + Y'_{12})(Y'_{22} + Y'_{21})R_s \end{bmatrix} \quad \text{(A14.4)}$$

where Y'_{11}, Y'_{12}, Y'_{21}, Y'_{22} are given by Eq. (A14.1).

A14.1 Unity Current Gain Frequency, f_T

From Eq. (A14.4), the extrinsic transconductance can be expressed as

$$g_m = \frac{i_2}{v_1}\Big|_{v_2=0} = Y_{21} = g'_m - j\omega C_{gd} - j\omega(C_{gs} + C_{gd})(g'_m - j\omega C_{gd})R_g$$
$$- (g'_m - j\omega C_{gd})(g'_{ds} + j\omega C_{db} + j\omega C_{gd})R_d$$
$$- (g'_m + j\omega C_{gs})(g'_m + g'_{ds} + j\omega C_{db})R_s, \quad \text{(A14.5)}$$

which has a dc value,

$$g_m(\omega = 0) = g'_m\left[1 - g'_m R_s - g'_{ds}(R_s + R_d)\right]. \quad \text{(A14.6)}$$

Likewise, the extrinsic current gain β can also be obtained from Eq. (A14.4), with the assumption that the output port is short-circuited (again, keeping only terms to the first order of R_s, R_d, R_g):

$$\beta = \frac{i_2}{i_1}\Big|_{v_2=0} = \frac{Y_{21}}{Y_{11}} = \frac{Y'_{21}}{Y'_{11}}\left[1 - (Y'_{22} - Y'_{12}Y'_{21}/Y'_{11})R_d\right.$$
$$\left.-(Y'_{22} + Y'_{11}Y'_{22}/Y'_{21} - Y'_{12} - Y'_{12}Y'_{21}/Y'_{11})R_s\right]$$
$$= \frac{Y'_{21}}{Y'_{11}}\left[1 - (Y'_{22} - Y'_{12}Y'_{21}/Y'_{11})(R_s + R_d) - (Y'_{11}Y'_{22}/Y'_{21} - Y'_{12})R_s\right]. \quad \text{(A14.7)}$$

For the current gain magnitude, the imaginary part inside the square bracket in Eq. (A14.7) can be neglected because its contribution is of the second order of R_s, R_d. Therefore,

$$|\beta| = \frac{\sqrt{g'^2_m + (\omega C_{gd})^2}}{\omega(C_{gs} + C_{gd})}$$
$$\times\left[1 - g'_{ds}(R_s + R_d) - \frac{C_{gd}}{C_{gs} + C_{gd}}g'_m(R_s + R_d).\right.$$
$$\left.+\frac{\omega^2(C_{gs} + C_{gd})\left[C_{gd}g'_{ds} + (C_{db} + C_{gd})g'_m\right]}{g'^2_m + (\omega C_{gd})^2}R_s\right]. \quad \text{(A14.8)}$$

The extrinsic unity current gain frequency is then obtained by setting $|\beta| = 1$ in the above equation and solving for $f_T = \omega/2\pi$ to the zeroth and first order of R_s, R_d:

$$f_T = \frac{g'_m}{2\pi\sqrt{C^2_{gs} + 2C_{gs}C_{gd}}}\left\{1 - \frac{C_{gs} + C_{gd}}{C^2_{gs} + 2C_{gs}C_{gd}}\left[g'_{ds}(C_{gs}R_s + C_{gs}R_d + C_{gd}R_d)\right.\right.$$
$$\left.\left.+g'_m(C_{gd}R_d - C_{db}R_s)\right]\right\}. \quad \text{(A14.9)}$$

A14.2 Unity Power Gain Frequency, f_{max}

The unity power gain frequency or the maximum oscillation frequency, f_{max}, can be solved by applying the power gain condition, Eq. (A13.8) in Appendix 13,

$$4\mathrm{Re}(Y_{11})\mathrm{Re}(Y_{22}) = \left|Y_{12} + Y^*_{21}\right|^2 \quad \text{(A14.10)}$$

to the extrinsic matrix $[\tilde{Y}]$, Eq. (A14.4). Since the leading terms in $\mathrm{Re}(Y_{11})$ are of the first order of R_s, R_d, R_g, only the zeroth order term in $\mathrm{Re}(Y_{22})$ needs be kept. Therefore,

$$\mathrm{Re}(Y_{11})\mathrm{Re}(Y_{22}) = g'_{ds}\left[\omega^2(C_{gs} + C_{gd})^2 R_g + \omega^2 C^2_{gs}R_s + \omega^2 C^2_{gd}R_d\right]. \quad \text{(A14.11)}$$

Also,

$$\mathrm{Re}(Y_{12} + Y^*_{21}) = g'_m - 2\omega^2 C_{gd}(C_{gs} + C_{gd})R_g$$
$$- \left[g'_m g'_{ds} + 2\omega^2 C_{gd}(C_{db} + C_{gd})\right]R_d - \left[g'_m(g'_m + g'_{ds}) - 2\omega^2 C_{gs}C_{db})\right]R_s. \quad \text{(A14.12)}$$

The imaginary part of $(Y_{12} + Y^*_{21})$ is of the first order of R_s, R_d, R_g, therefore negligible when the absolute value is taken in Eq. (A14.10). The maximum oscillation frequency is then solved by substituting Eqs. (A14.11) and (A14.12) into Eq. (A14.10), and keeping only the first order terms of R_s, R_d, R_g,

$$\omega_{max}^2 = \frac{g_m'^2/4}{(C_{gs}+C_{gd})^2 g_{ds}' R_g + C_{gd}(C_{gs}+C_{gd}) g_m' R_g + C_{gd}\left[(C_{gd}+C_{db}) g_m' + C_{gd} g_{ds}'\right] R_d + C_{gs}(C_{gs} g_{ds}' - C_{db} g_m') R_s} \quad \text{(A14.13)}$$

Alternatively, it can be expressed in terms of $\omega_T = 2\pi f_T \approx g_m'/(C_{gs} + C_{gd})$ (zeroth-order approximation to Eq. (A14.9)),

$$\omega_{max} = \frac{\omega_T/2}{\sqrt{g_{ds}' R_g + \omega_T C_{gd} R_g + \frac{C_{gd}+C_{db}}{C_{gs}+C_{gd}} \omega_T C_{gd} R_d - \frac{C_{db}}{C_{gs}+C_{gd}} \omega_T C_{gs} R_s + \frac{C_{gd}^2 g_{ds}' R_d + C_{gs}^2 g_{ds}' R_s}{(C_{gs}+C_{gd})^2}}} \quad \text{(A14.14)}$$

Note that f_{max} is simply $\omega_{max}/2\pi$. An approximate expression for f_{max} commonly found in the literature is obtained by keeping only the second term in the square root:

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_g C_{gd}}}. \quad \text{(A14.15)}$$

Appendix 15 Determination of Emitter and Base Series Resistances

The ideal base current I_{B0} is proportional to $\exp(qV_{BE}/kT)$, where V_{BE} is the forward-bias voltage applied to the emitter and base terminals. In Section 6.3.1, it is shown that the effect of the parasitic emitter and base series resistances is to reduce the voltage appearing across the immediate emitter–base junction by an amount ΔV_{BE} given by

$$\Delta V_{BE} = I_C r_e + I_B(r_e + r_{bi} + r_{bx}), \quad \text{(A15.1)}$$

where r_e is the emitter series resistance, r_{bi} is the intrinsic-base series resistance, r_{bx} is the extrinsic-base series resistance, I_B is the measured base current, and I_C is the measured collector current. As a result of this voltage reduction, the measured base current is reduced from its ideal value by the ratio

$$\frac{I_B}{I_{B0}} = \exp\left(-\frac{q\Delta V_{BE}}{kT}\right). \quad \text{(A15.2)}$$

Using Eq. (A15.1), Eq. (A15.2) can be rearranged (Ning and Tang, 1984) to give

$$\frac{kT}{qI_C}\ln\left(\frac{I_{B0}}{I_B}\right) = \left(r_e + \frac{r_{bi}}{\beta_0}\right) + \frac{r_e + r_{bx}}{\beta_0}, \quad \text{(A15.3)}$$

where $\beta_0 = I_C/I_B$ is the measured static common-emitter current gain.

A15.1 The Case Where The Emitter Series Resistance is Constant, Independent of V_{BE}

For most bipolar device fabrication processes, the emitter series resistance is a constant, independent of V_{BE} or the device operating current. In this case, r_e and r_{bx} are constants. Also, it is shown in Section 7.2.1 that the collector current density is proportional to the intrinsic-base sheet resistivity. Therefore, r_{bi} is proportional to β_0, and the ratio r_{bi}/β_0 is constant, independent of current. Thus, if $(kT/qI_C)\ln(I_{B0}/I_B)$ is plotted as a function of $1/\beta_0$, the slope gives $r_e + r_{bx}$ and the intercept gives $r_e + r_{bi}/\beta_0$. This is illustrated in Fig. A15.1.

The ratio r_{bi}/β_0 can be obtained at low currents where the individual values of r_{bi} and β_0 are relatively independent of current. β_0 is directly measurable from I_C and I_B, and r_{bi} can be calculated as described in Appendix 16, or measured as described below.

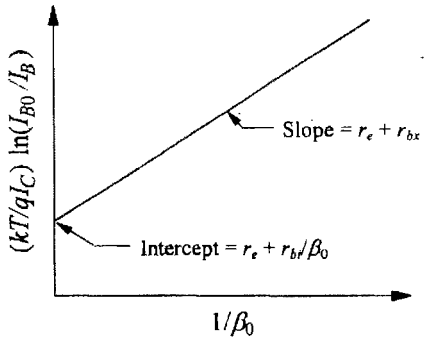


Figure A15.1. Schematic illustrating the determination of the emitter and base series resistances from Eq. (A15.3). This method works for devices where r_e is a constant, independent of V_{BE}.

Once the ratio r_{bi}/β_0 is determined, r_e and r_{bx} can be extracted from the intercept and the slope of the plot described above. Thus, all three resistance components can be obtained.

A15.2 The Case Where the Emitter Series Resistance is a Function of V_{BE}

Sometimes, a particular bipolar device fabrication process produces devices with an emitter series resistance that appears to depend on V_{BE} or on the device operating current. This is often the case with polysilicon-emitter processes, and certainly the case when the transport in and/or across the emitter is governed by tunneling (Yu *et al.*, 1984). In this case, a plot like Fig. A15.1 does not yield a straight line (Riccó *et al.*, 1984; Dubois *et al.*, 1994), and hence cannot be used to extract the emitter and base series resistances.

A15.3 Direct Measurement of Base Resistances

The total base resistance can be measured directly by using a test device structure with two separate base contacts, one on each side of the emitter stripe (Weng *et al.*, 1992). This device structure and the measurement technique are illustrated in Fig. A15.2. Here the mask dimension of the emitter stripe is W_{mask}, and the electrical emitter stripe width is W. Base contact B1 is used to operate the transistor as usual, while base contact B2 is used only to sense the potential difference that develops between B1 and B2 when a base current flows from B1 into the emitter terminal. As a terminal for voltage sensing only, B2 draws negligible base current. Thus, the device operates like a transistor with base contact on only side of the emitter stripe. The total base series resistance is simply the measured potential difference between B1 and B2, divided by the measured base current.

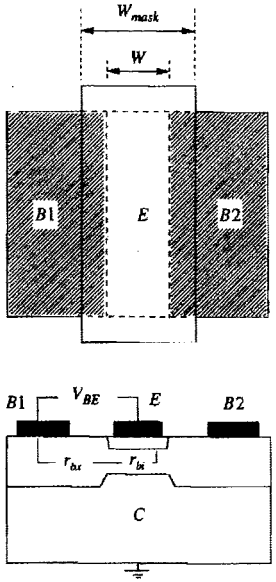


Figure A15.2. Experimental technique for measuring the total base series resistance. The special device structure has two separate base contacts, B1 and B2, one on each side of the emitter stripe. The transistor is operated as usual, using base contact B1. The voltage drop between B1 and B2, due to the lateral flow of base current, is sensed using base contact B2. (After Weng *et al.*, 1992.)

Let r_b be the total base resistance measured as described above. That is,

$$r_b = r_{bx} + r_{bi}. \tag{A15.4}$$

It is reasonable to assume that, for a given device fabrication process, W is related to its mask dimension W_{mask} by

$$W = W_{mask} - 2\Delta W, \tag{A15.5}$$

where ΔW represents the overlap per edge between the emitter mask and the extrinsic-base region. Also, it is shown in Appendix 16 that r_{bi} is proportional to the emitter stripe width. Therefore,

$$r_{bi} \propto W_{mask} - 2\Delta W. \tag{A15.6}$$

By using specially designed device structures of various W_{mask} dimensions and measuring r_b as a function of W_{mask}, r_{bx} can be determined (Weng *et al.*, 1992). This is illustrated in Fig. A15.3. Once r_{bx} is known, r_{bi} for a given W_{mask} can be determined.

A15.4 Dependence of Base Resistance on V_{BE}

The intrinsic-base sheet resistivity is given by Eq. (7.5), i.e.,

$$R_{Sbi} = \left(q \int_0^{W_B} p_p(x) \mu_p(x) dx \right)^{-1} . \tag{A15.7}$$

As the emitter–base diode is forward-biased, the emitter–base junction depletion-layer width decreases, and hence the intrinsic-base width W_B increases, as can be seen readily from Eq. (2.85). Also, as shown in Eq. (6.82), the electrons injected from the emitter and traversing the base–collector junction depletion layer causes the base width to increase. The amount of base-width increase increases with the injected electron concentration. As a result, at small emitter–base forward biases, R_{Sbi} decreases slowly as W_B increases with increasing V_{BE}. The total base resistance therefore decreases slowly with increasing V_{BE}. This is illustrated in Fig. A15.4.

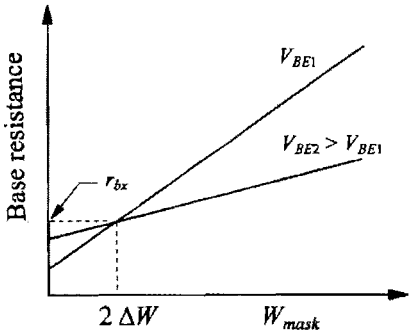


Figure A15.3. Schematic illustrating the determination of the total base resistance r_b and its extrinsic-base component r_{bx} using the device structure shown in Fig. A15.2.

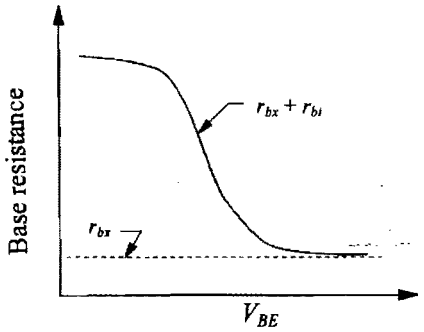


Figure A15.4. Schematic illustrating the dependence of the total base resistance on forward V_{BE}.

At large emitter–base forward biases, the base-widening effect becomes significant. When that happens, both $p_p(x)$ and W_B increase rapidly with further increase in V_{BE}. As a result, R_{Sbi}, and hence the total base resistance, decreases rapidly with increasing V_{BE}. The total base resistance approaches r_{bx} in the limit of very large forward V_{BE} (Weng *et al.*, 1992). This is also illustrated in Fig. A15.4.

Appendix 16 Intrinsic-Base Resistance

Consider the intrinsic part of a bipolar transistor, the cross section of which is shown schematically in Fig. A16.1. The base current I_B enters the intrinsic-base region at the base contact and then spreads out, turns upward, and enters the emitter. Thus, the effective intrinsic-base resistance r_{bi} as seen by the base current depends on how the base current spreads out inside the intrinsic-base layer. One commonly used method for evaluating r_{bi} is to consider the power dissipation P in the intrinsic base (Hauser, 1968), and define r_{bi} by

$$P = I_B^2 r_{bi}. \tag{A16.1}$$

A16.1 The Case of Negligible Current Crowding

The power dissipation can be evaluated readily for low-current situations where there is no current crowding and the base-current density J_B entering the emitter can be assumed to be uniform. That this is a good assumption for modern bipolar transistors will be shown later.

Let us assume the emitter stripe has a width W and a length L, and the base is contacted on one side of the emitter only, as shown in Fig. A16.1. Consider a slice of the intrinsic base between points y and $y+\Delta y$. The resistance of this slice as seen by the base current is

$$\Delta R = \frac{R_{Sbi}}{L}\Delta y, \tag{A16.2}$$

where R_{Sbi} is the intrinsic-base sheet resistivity given by Eq. (7.5). The base current passing through this slice is

$$i_B(y) = J_B L(W - y), \tag{A16.3}$$

and the power dissipated in this slice is

$$\Delta P = i_B^2(y)\Delta R = \frac{R_{Sbi}}{L} i_B^2(y)\Delta y. \tag{A16.4}$$

R_{Sbi} is a function of collector current density, which in turn is a function of the base–emitter diode voltage. The assumption of negligible current crowding implies that there is negligible lateral voltage drop along the intrinsic base, and R_{Sbi} is independent of y. Therefore, the total power dissipated by the base current in the intrinsic base is

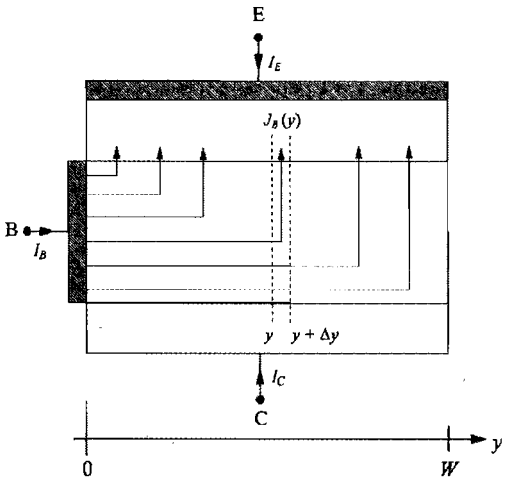


Figure A16.1. Schematic of the intrinsic part of a bipolar transistor illustrating the flow of base current. The transistor has an emitter stripe of width W and a base contact on only one side.

$$\begin{aligned} P &= \frac{R_{Sbi}}{L}\int_0^W i_B^2(y)dy \\ &= R_{Sbi}J_B^2 L\int_0^W (W-y)^2 dy \\ &= \frac{1}{3}R_{Sbi}J_B^2 L W^3 \\ &= \frac{1}{3}\left(\frac{W}{L}\right)R_{Sbi}I_B^2. \end{aligned} \tag{A16.5}$$

Comparison of Eq. (A16.1) with Eq. (A16.5) gives

$$r_{bi} = \frac{1}{3}\left(\frac{W}{L}\right)R_{Sbi}. \tag{A16.6}$$

It should be noted that, as illustrated in Fig. A15.4, ***R_{Sbi}, and hence r_{bi}, rolls off with collector current density.*** The rolloff can be very rapid once the collector current density is large enough to cause appreciable base widening.

A16.2 Other Emitter Geometries

The above calculations can be applied to other emitter geometries and/or base contact schemes. For the same emitter-stripe geometry, but with base contact on both sides of the emitter, the base resistance is reduced by a factor of 4, namely

$$r_{bi} = \frac{1}{12}\left(\frac{W}{L}\right)R_{Sbi} \qquad \text{for two-sided base contact.} \tag{A16.7}$$

The base resistance for a square emitter with base contact on all four sides is

$$r_{bi} = \frac{1}{32} R_{Sbi} \qquad \text{for four-sided base contact.} \tag{A16.8}$$

For a round emitter, it is

$$r_{bi} = \frac{1}{8\pi} R_{Sbi} \qquad \text{for round emitter.} \tag{A16.9}$$

A16.3 Estimation of Emitter Current-Crowding Effect

The assumption of uniform base current density used above is valid when the maximum lateral voltage drop in the intrinsic-base layer, caused by the lateral flow of the base current, is small compared to kT/q. If the lateral voltage drop is not negligible, then the base current density becomes a function of y, with

$$J_B(y) = J_B(0) \exp\left(\frac{qV_{BE}(y)}{kT} - \frac{qV_{BE}(0)}{kT}\right). \tag{A16.10}$$

Since $V_{BE}(y) \leq V_{BE}(0)$, we have $J_B(y) \leq J_B(0)$. That is, the emitter current density is larger at the emitter edge than towards the middle of the emitter. This is known as the emitter current-crowding effect. The general expressions for $J_B(y)$ and $V_{BE}(y)$ have been derived by Hauser (1964, 1968). The derivation is rather involved, and the interested reader is referred to the references for details. Here we simply give an upper-bound estimate of the emitter current-crowding effect.

When the base current density is a function of distance from the emitter edge, then instead of Eq. (A16.3), the base current passing through the intrinsic-base slice at point y is

$$i_B(y) = L \int_y^W J_B(y')dy'. \tag{A16.11}$$

The voltage drop across an intrinsic-base slice between y and $y + \Delta y$ is

$$\begin{aligned} \Delta V_{BE}(y) &= V_{BE}(y + \Delta y) - V_{BE}(y) \\ &= -i_B(y)\Delta R(y) = -\Delta y\, R_{Sbi}(J_C(y)) \int_y^W J_B(y')dy', \end{aligned} \tag{A16.12}$$

where Eq. (A16.2) has been used for ΔR. In Eq. (A16.12), R_{Sbi} is denoted explicitly as a function of J_C, which in turn is a function of y. The gradient of the base–emitter voltage along the intrinsic-base layer is therefore

$$\begin{aligned} \frac{dV_{BE}(y)}{dy} &= -R_{Sbi}(J_C(y)) \int_y^W J_B(y')dy' \\ &= -R_{Sbi}(J_C(y))J_B(0) \int_y^W \exp\left(\frac{qV_{BE}(y')}{kT} - \frac{qV_{BE}(0)}{kT}\right)dy', \end{aligned} \tag{A16.13}$$

where we have used Eq. (A16.10) for $J_B(y)$. Notice that this voltage gradient is negative. An upperbound of the magnitude of this voltage gradient can be obtained by replacing the exponential inside the integral, which is smaller than unity because $V_{BE}(y) \leq V_{BE}(0)$, by unity, and replacing R_{Sbi} by its low-current value. That is,

$$\begin{aligned} \left|\frac{dV_{BE}(y)}{dy}\right| &\leq R_{Sbi}(J_C \to 0)J_B(0) \int_y^W dy' \\ &= R_{Sbi}(J_C \to 0)J_B(0)(W - y). \end{aligned} \tag{A16.14}$$

Integrating Eq. (A16.14) gives an upper-bound estimate of the maximum voltage drop along the intrinsic base due to the lateral flow of base current. This estimate is

$$\begin{aligned} V_{BE}(0) - V_{BE}(W) &< R_{Sbi}(J_C \to 0)J_B(0) \int_0^W (W - y)dy \\ &= R_{Sbi}(J_C \to 0)J_B(0)\frac{W^2}{2}. \end{aligned} \tag{A16.15}$$

Consider a typical modern bipolar transistor. To avoid excessive base-widening effect, the collector current density is typically under 1 mA/μm^2. For a typical current gain of 100, J_B is therefore typically under 0.01 mA/μm^2. The low-current value of R_{Sbi} is typically 10 kΩ/□. If the transistor has an emitter-stripe width of 0.5 μm, with base contact on both sides of the stripe, then $W = 0.25$ μm. For this typical transistor, Eq. (A16.15) suggests that the maximum lateral voltage drop in the intrinsic base is less than 3 mV, much smaller than kT/q, which is about 26 mV at room temperature. As the intrinsic-base sheet resistivity rolls off with emitter–base forward bias, this lateral voltage drop becomes even smaller. Therefore, ***emitter current-crowding effect is negligible for modern bipolar transistors***. Of course, for transistors of earlier generations, where the emitter stripes were much wider, emitter current crowding could be significant. Our estimates are consistent with the results obtained from a more exact calculation of the internal base voltage distribution (Chiu *et al.*, 1992).

Appendix 17 Energy-Band Diagram of a Si–SiGe n–p Diode

When Ge is added to Si, the energy-band edges of the resulting SiGe alloy are shifted relative to those of Si such that the energy bandgap of the SiGe alloy is smaller than that of Si. Figure A17.1 is a schematic showing the shifts in the energy-band edges of SiGe relative to the energy-band edges of Si for two $Si_{1-x}Ge_x$ compositions (People, 1986). It shows that the bandgap narrowing in SiGe is caused mostly by shifts in the valence-band edge. Shifts in the conduction-band edge are relatively small.

To understand the operation of a SiGe-base bipolar transistor, we should first establish its energy-band diagram. Since the currents in a bipolar transistor in normal operation are controlled by the emitter–base diode, we need to consider only the energy-band diagram of the emitter–base diode. Therefore, we want to consider an n^+–p Si–SiGe diode. For simplicity, we assume that the bandgap narrowing in the SiGe is due entirely to shifts in the valence-band edge. This is a good approximation since shifts in the conduction-band edge are relatively small.

Let us first consider an n^+–p^+ Si–SiGe diode having a spatially uniform Ge distribution. (We assume the regions to be heavily doped to make it easier to understand the effect of adding Ge into Si because the Fermi level of an n^+-region is very close to its conduction-band edge while the Fermi level of a p^+-region is very close to its valance-band edge.) The energy-band diagrams of the Si and SiGe regions when they are physically apart are illustrated schematically in Fig. A17.2(a). The SiGe region has a smaller energy gap than the Si region. The conduction-band edges of the Si region and that of the SiGe region are at the same energy, since for simplicity we are ignoring the relatively small shift in the conduction-band edge of SiGe relative to that of Si. As discussed in Sections 2.1.1.3 and 2.1.4.5, the Fermi level is spatially constant at thermal equilibrium. As a result, when the n^+-Si and the p^+-SiGe regions are brought together to form a diode, the energy-band diagram of the resulting diode is that illustrated in Fig. A17.2(b), with the Fermi level being constant across the diode. Figure A.17.2(b) clearly shows that the reduction in energy barrier caused by the narrower SiGe energy bandgap occurs in the conduction band, not in the valence band. That is, compared to a Si–Si n^+–p diode, a Si–SiGe n^+–p diode has a smaller energy barrier for electron injection from the emitter into the base, but there is little change in the energy barrier for hole injection from the base into the emitter.

Let us next consider a p-type SiGe region having a triangular or linearly graded Ge distribution. Figure A17.3(a) shows the energy-band diagram of a p-type Si region. Figure A17.3(c) shows the energy-band diagram of a p-type SiGe region when a linearly graded Ge distribution, as illustrated in Fig. A17.3(b), is incorporated into the p-type Si

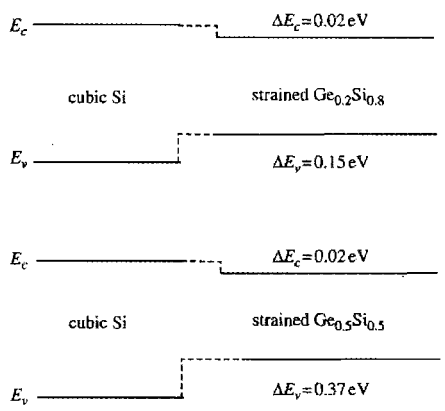


Figure A17.1. Schematic diagram showing shifts in energy-band edges of strained SiGe relative to regular cubic Si. (After People, 1986.)

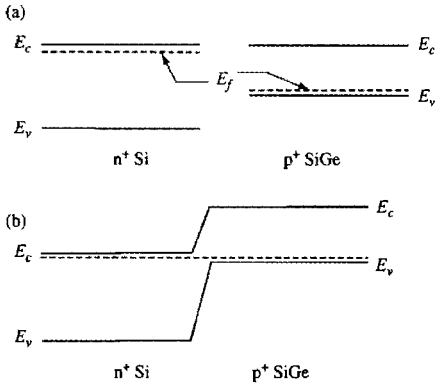


Figure A17.2. (a) Schematic energy-band diagram of an n^+-Si region and a p^+-SiGe region located physically apart. (b) Energy-band diagram when the two regions are brought together to form a diode.

region. The linearly graded Ge distribution causes the SiGe bandgap to narrow in proportion to the Ge concentration. This gradual narrowing of the energy bandgap and the spatially constant Fermi level, which is separated from the valence-band edge by a value given by Eq. (2.22), combine to cause the SiGe energy-band edges to become that illustrated in Fig. A17.3(c). The result is that the valence-band edge remains essentially spatially constant, while the conduction-band edge slopes downward at a rate proportional to the grading in the Ge profile. The amount of energy bandgap narrowing across the p-region is denoted by ΔE_g.

Next consider a Si–SiGe diode comprising an n^+-Si emitter and a p-type SiGe base having a linearly graded Ge profile. The energy-band diagrams of the physically

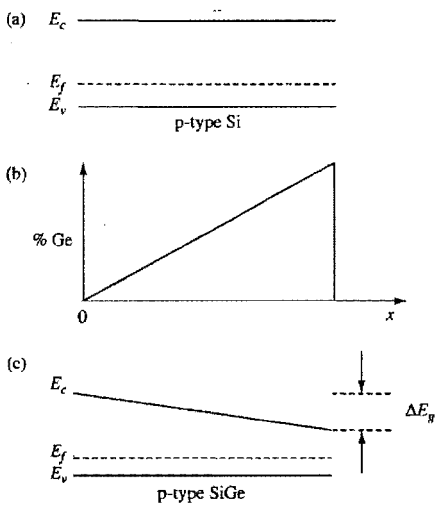


Figure A17.3. Schematics of (a) energy-band diagram of a p-type Si region, (b) a triangular or linearly graded Ge distribution, and (c) energy-band diagram of a p-type SiGe region having a linearly graded Ge distribution.

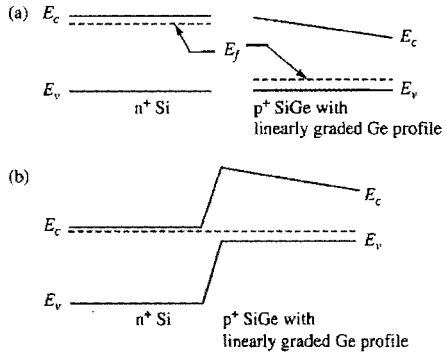


Figure A17.4. Schematics of (a) energy-band diagrams of an n^+-Si region and a p-type SiGe region located physically apart, and (b) energy-band diagram when the two regions are brought together to form a diode.

separated emitter and base regions are as illustrated in Fig. A17.4(a), and the energy-band diagram of the diode is as illustrated in Fig. A17.4(b). Since the Ge concentration in the base at the emitter-base junction is zero, the energy barrier for electron injection from the emitter into the base and that for hole injection from the base into the emitter are exactly the same as those for a regular Si–Si n^+–p diode. Any increase in electron injection current into the base is due to the bandgap grading in the base, not due to any change in energy barrier at the emitter–base junction.

Appendix 18 f_T and f_{max} of a Bipolar Transistor

In this appendix, we follow the admittance matrix approach developed in Appendices 13 and 14 for a MOSFET to derive the expressions for f_T and f_{max} of a bipolar transistor. The two-port network for an intrinsic bipolar transistor is shown in Fig. A18.1, derived from the small-signal equivalent circuit shown in Fig. 6.19. For simplicity, the collector–substrate capacitor has been dropped. This is justified because $C_{dCS,tot} \ll C_{dBC,tot}$ for a typical modern bipolar transistor. The corresponding admittance matrix can be written as

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = [\tilde{Y}_i] \begin{bmatrix} v_1' \\ v_2' \end{bmatrix} = \begin{bmatrix} Y_{11}' & Y_{12}' \\ Y_{21}' & Y_{22}' \end{bmatrix} \begin{bmatrix} v_1' \\ v_2' \end{bmatrix}, \tag{A18.1}$$

with the matrix elements given by:

$$Y_{11}' = \frac{1}{r_\pi'} + j\omega(C_\pi + C_\mu) = \frac{g_m'}{\beta_0} + j\omega(C_\pi + C_\mu), \tag{A18.2}$$

$$Y_{12}' = -j\omega C_\mu, \tag{A18.3}$$

$$Y_{21}' = g_m' - j\omega C_\mu, \tag{A18.4}$$

and

$$Y_{22}' = \frac{1}{r_0'} + j\omega C_\mu. \tag{A18.5}$$

In Eq. (A18.2), we have used the fact that $1/r_\pi' = g_m'/\beta_0$ where β_0 is the static current gain [see Eq. (6.100)]. Note that the real part of Y_{12}' is zero because the base current is not a function of collector voltage for a transistor biased in the forward active mode.

To derive the admittance matrix for an extrinsic transistor, we need to include the IR drops in the resistors r_b (= $r_{bi} + r_{bx}$), r_e and r_c shown in Figs 6.18 and 6.20. The small-signal forms of Eqs. (6.107) and (6.108) are

$$v_{be} = v_{be}' + i_b(r_b + r_e) + i_c r_e \tag{A18.6}$$

and

$$v_{ce} = v_{ce}' + i_b r_e + i_c(r_c + r_e), \tag{A18.7}$$

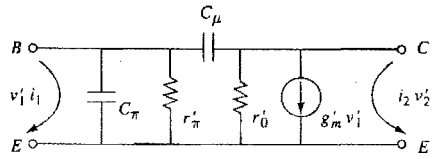


Figure A18.1. Two-port network for an intrinsic bipolar transistor.

which, in matrix notation, imply

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1' \\ v_2' \end{bmatrix} + \begin{bmatrix} r_b + r_e & r_e \\ r_e & r_c + r_e \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}. \tag{A18.8}$$

Comparing Eq. (A18.8) with Eq. (A14.2) of Appendix 14, we see the equivalencies between r_b and R_g, between r_e and R_s, and between r_c and R_d. Therefore, from Eq. (A14.4) of Appendix 14, we can write

$$Y_{11} = Y'_{11} - Y'^2_{11} r_b - Y'_{12} Y'_{21} r_c - (Y'_{11} + Y'_{12})(Y'_{11} + Y'_{21}) r_e, \tag{A18.9}$$

$$Y_{12} = Y'_{12} - Y'_{11} Y'_{12} r_b - Y'_{12} Y'_{22} r_c - (Y'_{11} + Y'_{12})(Y'_{12} + Y'_{22}) r_e, \tag{A18.10}$$

$$Y_{21} = Y'_{21} - Y'_{11} Y'_{21} r_b - Y'_{21} Y'_{22} r_c - (Y'_{11} + Y'_{21})(Y'_{21} + Y'_{22}) r_e, \tag{A18.11}$$

$$Y_{22} = Y'_{22} - Y'_{12} Y'_{21} r_b - Y'^2_{22} r_c - (Y'_{12} + Y'_{22})(Y'_{21} + Y'_{22}) r_e. \tag{A18.12}$$

The above equations can be used to relate the extrinsic device parameters to the intrinsic device parameters and parasitic resistors. The processes involved are quite tedious and the results are usually very complex, involving many terms. Both the process and the results can be greatly simplified if the appropriate approximations are made as early as possible. One of the approximations is $\mathrm{Re}(Y'_{21}) = \beta_0 \mathrm{Re}(Y'_{11}) >> \mathrm{Re}(Y'_{11})$. Another approximation is to assume a large Early voltage, which is equivalent to assuming r_0' being large compared to $1/g_m'$, r_e, r_b and r_c. The Early voltage is typically about 40 V for a Si-base transistor (see Section 6.3.2.1), and larger than 100 V for a SiGe-base transistor. Therefore, for a collector current of 1 mA, $r_0' = V_A/I_C > 4 \times 10^4 \Omega$ for a Si-base transistor and $> 10^5\ \Omega$ for a SiGe-base transistor. From Eq. (6.99), $1/g_m' = kT/qI_C$ is about 26 Ω. Also, $r_b \gg r_e$ and $r_b \gg r_c$ are good approximations. For a well designed bipolar transistor, r_b should be dominated by the intrinsic base resistance r_{bi}. For a typical bipolar transistor having a rectangular emitter geometry and two-sided base contact, Eq. (A16.7) gives $r_{bi} = (W/L)R_{Sbi}/12$, where W is the emitter stripe width and L is the emitter stripe length, and R_{Sbi} is the sheet resistivity of the intrinsic base layer. A typical value for R_{Sbi} is 10 kΩ/□. Therefore, for a transistor with $(W/L) = 0.2$, typical r_{bi} is about 400 Ω. Therefore, the assumption of r_0' being large compared to $1/g_m'$, r_e, r_b and r_c is valid for typical bipolar transistors. We will point out when and where these approximations are used.

A18.1 Cutoff (Unity Current Gain) Frequency, f_T

The short-circuit-load current gain for the intrinsic device is

$$\beta' = \left.\frac{i_2}{i_1}\right|_{v_2'=0} = \frac{Y'_{21}}{Y'_{11}}. \tag{A18.13}$$

Therefore, the cutoff frequency of the intrinsic device is given by

$$1 = \left|\frac{Y'_{21}}{Y'_{11}}\right| = \sqrt{\frac{g_m'^2 + \omega_T'^2 C_\mu^2}{\dfrac{g_m'^2}{\beta_0^2} + \omega_T'^2 (C_\pi + C_\mu)^2}}, \tag{A18.14}$$

which can be rearranged to give

$$\omega_T' = 2\pi f_T' = \sqrt{\frac{g_m'^2 - \dfrac{g_m'^2}{\beta_0^2}}{(C_\pi + C_\mu)^2 - C_\mu^2}} \approx \frac{g_m'}{\sqrt{(C_\pi + C_\mu)^2 - C_\mu^2}}, \tag{A18.15}$$

where we have use the approximation $1/\beta_0 \ll 1$. For a typical modern bipolar transistor which has reduced C_μ ($= C_{dBC,tot}$), we have $C_\pi (= C_{dBE,tot} + C_{DE}) \gg C_\mu$. Therefore, Eq. (A18.15) can be simplified to give the commonly used approximation of

$$\omega_T' = 2\pi f_T' \approx \frac{g_m'}{C_\pi + C_\mu}. \tag{A18.16}$$

The short-circuit-load extrinsic current gain, to the first order of r_b, r_c, and r_e, is

$$\begin{aligned} \beta = \left.\frac{i_2}{i_1}\right|_{v_2=0} &= \frac{Y_{21}}{Y_{11}} = \frac{Y'_{21}}{Y'_{11}} \left(\frac{1 - r_b Y'_{11} - r_c Y'_{22} - r_e (Y'_{11} + Y'_{21})(Y'_{21} + Y'_{22})/Y'_{21}}{1 - r_b Y'_{11} - r_c Y'_{12} Y'_{21}/Y'_{11} - r_e (Y'_{11} + Y'_{12})(Y'_{11} + Y'_{21})/Y'_{11}} \right) \\ &\approx \frac{Y'_{21}}{Y'_{11}} \left[1 - (r_c + r_e)(Y'_{22} - Y'_{12} Y'_{21}/Y'_{11}) + r_e (Y'_{12} - Y'_{11} Y'_{22}/Y'_{21}) \right]. \end{aligned} \tag{A18.17}$$

Notice that β is independent of r_b. Noting that $\mathrm{Re}(Y'_{12}) = 0$ and that $(r_c + r_e)\,\mathrm{Re}(Y'_{22}) \ll 1$, we have, to the first order of r_b, r_c, and r_e,

$$\begin{aligned} |\beta| &= \left|\frac{Y'_{21}}{Y'_{11}}\right| \left[1 + (r_c + r_e)\mathrm{Re}(Y'_{12} Y'_{21}/Y'_{11}) - r_e \mathrm{Re}(Y'_{11} Y'_{22}/Y'_{21}) \right] \\ &\approx \left|\frac{Y'_{21}}{Y'_{11}}\right| \left[1 + (r_c + r_e)\mathrm{Re}(Y'_{12} Y'_{21}/Y'_{11}) \right]. \end{aligned} \tag{A18.18}$$

where we have used the fact that $r_e \mathrm{Re}(Y'_{11} Y'_{22}/Y'_{21}) \ll 1$ for a typical transistor (see Exercise 8.9) in writing the last equation. The cutoff frequency is obtained by setting $|\beta| = 1$, i.e.,

$$1 = \sqrt{\frac{g_m'^2 + \omega_T^2 C_\mu^2}{(g_m'/\beta_0)^2 + \omega_T^2 (C_\pi + C_\mu)^2}} \left[1 - \frac{(r_c + r_e)\omega_T^2 g_m' (C_\mu^2/\beta_0 + C_\pi C_\mu + C_\mu^2)}{(g_m'/\beta_0)^2 + \omega_T^2 (C_\pi + C_\mu)^2} \right]. \tag{A18.19}$$

Noting that $1/\beta_0 \ll 1$ and keeping only the terms to the first order of r_c and r_e, we can rearrange Eq. (A18.19) to give

$$g_m'^2 + \omega_T^2 C_\mu^2 = \omega_T^2 (C_\pi + C_\mu)^2 + 2(r_c + r_e) g_m' \omega_T^2 (C_\pi + C_\mu) C_\mu, \tag{A18.20}$$

which in turn can be rearranged to give

$$\frac{1}{\omega_T} = \frac{\sqrt{(C_\pi + C_\mu)^2 - C_\mu^2}}{g_m'} \left[1 + (r_c + r_c) \frac{g_m' (C_\pi + C_\mu) C_\mu}{(C_\pi + C_\mu)^2 - C_\mu^2} \right]. \tag{A18.21}$$

Using Eqs. (A18.15) and (A18.16) and the approximations involved in them, we can simplify Eq. (A18.21) to the commonly used approximation of

$$\frac{1}{\omega_T} \approx \frac{1}{\omega_T'} + (r_e + r_c) C_\mu = \frac{1}{\omega_T'} + (r_e + r_c) C_{dBC,tot}. \tag{A18.22}$$

A18.2 Unity Power Gain (Maximum Oscillation) Frequency, f_{max}

From Eq. (A14.10), f_{max} can be obtained from solving the equation

$$4\mathrm{Re}(Y_{11})\mathrm{Re}(Y_{22}) = \left| Y_{12} + Y_{21}^* \right|^2. \tag{A18.23}$$

Using Eqs. (A18.9) and (A18.12) and keeping only the terms to the first order of r_b, r_c, and r_e, we have

$$\begin{aligned} \mathrm{Re}(Y_{11})\mathrm{Re}(Y_{22}) = {} & g_m'/\beta_0 r_0' - r_b \left[g_m'^2/\beta_0^2 r_0' - g_m' \omega^2 C_\mu^2/\beta_0 - \omega^2 (C_\pi + C_\mu)^2/r_0' \right] \\ & - r_c \left[g_m'/\beta_0 r_0'^2 - g_m' \omega^2 C_\mu^2/\beta_0 - \omega^2 C_\mu^2/r_0' \right] \\ & - r_e \left[2g_m'^2/\beta_0 r_0' - \omega^2 C_\pi^2/r_0' + g_m'/\beta_0 r_0'^2 \right]. \end{aligned} \tag{A18.24}$$

A similar expansion and approximation can be made for $\left| Y_{12} + Y_{21}^* \right|^2$, and Eq. (A18.23) can then be used to determine f_{max}. However, if we make the ***large Early voltage approximation*** early on, the procedure and the final expression for f_{max} can be greatly simplified. For a typical modern bipolar transistor operating at $I_C = 1$ mA, C_μ ($= C_{dBC,tot}$) > 1 fF, $\beta_0 > 100$, $r_0' > 10^5\ \Omega$, $g_m' \approx 4 \times 10^{-2}\ \Omega^{-1}$ and $f_{max} > 30$ GHz. Thus, for a typical modern bipolar transistor, the terms of the second order in $1/r_0'$ in Eq. (A18.24) can be ignored compared to those of the first order in $1/r_0'$. That is, Eq. (A18.24) can be approximated by

$$\begin{aligned} \mathrm{Re}(Y_{11})\mathrm{Re}(Y_{22}) \approx {} & g_m'/\beta_0 r_0' - r_b \left[g_m'^2/\beta_0^2 r_0' - g_m' \omega^2 C_\mu^2/\beta_0 - \omega^2 (C_\pi + C_\mu)^2/r_0' \right] \\ & + r_c \left[g_m' \omega^2 C_\mu^2/\beta_0 + \omega^2 C_\mu^2/r_0' \right] - r_e \left[2g_m'^2/\beta_0 r_0' - \omega^2 C_\pi^2/r_0' \right]. \end{aligned} \tag{A18.25}$$

Again, keeping only the terms to the first order of r_b, r_c, and r_e, we have

$$\begin{aligned} \left| Y_{12} + Y_{21}^* \right|^2 &= g_m'^2 - r_b \left[4g_m' \omega^2 (C_\pi + C_\mu) C_\mu + 2g_m'^3/\beta_0 \right] \\ &\quad - r_c \left[4g_m' \omega^2 C_\mu^2 + 2g_m'^2/r_0' \right] - r_e \left[2g_m'^3 + 2g_m'^2/\beta_0 r_0' + 2g_m'^2/r_0' \right] \\ &\approx g_m'^2 - r_b \left[4g_m' \omega^2 (C_\pi + C_\mu) C_\mu + 2g_m'^3/\beta_0 \right] - 4r_c g_m' \omega^2 C_\mu^2 - 2r_e g_m'^3, \end{aligned} \tag{A18.26}$$

where in writing the last equation, we have used the fact that $g_m' \gg 1/r_0'$, $r_e/r_0' \ll 1$ and $r_c/r_0' \ll 1$. Substituting Eqs. (A18.25) and (A18.26) into Eq. (A18.23), and again noting that $g_m' \gg 1/r_0'$, $r_c \ll r_0'$, $r_b \ll r_0'$, $r_e \ll r_0'$ and $\beta_0 \gg 1$, we have the equation for f_{max} in the form

$$\begin{aligned} & g_m'^2 (1 - 2g_m' r_b/\beta_0 - 2g_m' r_e) \\ & \quad = \omega_{max}^2 4g_m' \left\{ r_b \left[(C_\pi + C_\mu) C_\mu + (C_\pi + C_\mu)^2/g_m' r_0' \right] + r_c C_\mu^2 + r_e C_\pi^2/g_m' r_0' \right\}. \end{aligned} \tag{A18.27}$$

To the first order of r_b, r_c, and r_e, Eq. (A18.27) can be rearranged to give

$$\omega_{max} = \frac{g_m' (1 - g_m' r_b/\beta_0 - g_m' r_e)}{\sqrt{4g_m' \left\{ r_b \left[(C_\pi + C_\mu) C_\mu + (C_\pi + C_\mu)^2/g_m' r_0' \right] + r_c C_\mu^2 + r_e C_\pi^2/g_m' r_0' \right\}}}. \tag{A18.28}$$

For a typical modern bipolar transistor, $g_m' r_0' = qV_A/kT > 4 \times 10^3$ (see Section 6.4.3). While it is true that $C_\pi = (C_{dBE,tot} + C_{DE}) \gg C_\mu$, C_π is not larger than C_μ by a factor of $g_m' r_0'$ without severe base widening occurring. In other words, for a transistor operated at typical current densities of interest, the terms proportional to $1/g_m' r_0'$ in the denominator can be ignored. That is, we have

$$\omega_{max} \approx \frac{g_m' (1 - g_m' r_b/\beta_0 - g_m' r_e)}{\sqrt{4g_m' \left[r_b (C_\pi + C_\mu) C_\mu + r_c C_\mu^2 \right]}}. \tag{A18.29}$$

Equation (A18.29) is equivalent to that obtained by making the large Early voltage approximation ($1/r_0' \approx 0$) from the very beginning. The familiar approximation for ω_{max}

is obtained if we approximate the numerator by g'_m and keep only the term containing r_b in the denominator, i.e.,

$$\omega_{\max} \approx \frac{g'_m}{\sqrt{4g'_m(C_\pi + C_\mu)C_\mu r_b}} = \sqrt{\frac{f'_T}{8\pi C_\mu r_b}}, \tag{A18.30}$$

where we have used Eq. (A18.16) for f'_T.

References

Abbas, S. A. (1974). Substrate current – a device and process monitor, *IEEE IEDM Technical Digest*, 404–407.

Abbas, S. A. and Dockerty, R. C. (1975). Hot-carrier instability in IGFETs, *Appl. Phys. Lett.* **27**, 147–148.

Andrews, J. M. (1974). The role of the metal–semiconductor interface in silicon integrated technology, *J. Vac. Sci. Technol.* **11**, 972–984.

Andrews, J. M. and Phillips, J. C. (1975). Chemical bonding and structure of metal–semiconductor interfaces, *Phys. Rev. Lett.* **35**, 56–59.

Ansley, W. E., Cressler, J. D. and Richey, D. M. (1998). Base-profile optimization for minimum noise figure in advanced UHV/CVD SiGe HBTs, *IEEE Trans. Microwave Theory and Techniques* **46**, 653–660.

Arora, N. (1993). *MOSFET Models for VLSI Circuit Simulation*. Wien: Springer-Verlag.

Arora, N. D. and Gildenblat, G. S. (1987). A semi-empirical model of the MOSFET inversion layer mobility for low-temperature operation, *IEEE Trans. Electron Devices* **ED-34**, 89–93.

Asbeck, P. M. and Nakamura, T. (2001). Bipolar transistor technology: past and future directions, *IEEE Trans. Electron Devices* **48**, 2455–2456.

Ashburn, P. (1988). *Design and Realization of Bipolar Transistors*. Chichester: Wiley.

Auberton-Hervé, A. J. (1996). SOI: materials to systems, *IEEE IEDM Technical Digest*, 3–10.

Avenier, G., Schwartzmann, T., Chevalier, P., Vandelle, B., Rubaldo, L., Dutartre, D., Boissonnet, L., Saguin, F., Pantel, R., Frégonèse, S., Maneux, C., Zimmer, T. and Chantre, A. (2005). A self-aligned vertical HBT for thin SOI SiGeC BiCMOS, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 218–131.

Baccarani, G. and Sai-Halasz, G. A. (1983). Spreading resistance in submicron MOSFETs, *IEEE Electron Device Lett.* **EDL-4**, 27–29.

Baccarani, G. and Wordeman, M. R. (1983). Transconductance degradation in thin-oxide MOSFETs, *IEEE Trans. Electron Device*, **ED-30**, 1295–1304.

Baccarani, G., Wordeman, M. R. and Dennard, R. H. (1984). Generalized scaling theory and its application to a 1/4 micrometer MOSFET design, *IEEE Trans. Electron Devices* **ED-31**, 452–462.

Bakoglu, H. B. (1990). *Circuits, Interconnections, and Packaging for VLSI*. Addison-Wesley.

Balk, P., Burkhardt, P. G. and Gregor, L. V. (1965). Orientation dependence of built-in surface charge on thermally oxidized silicon, *IEEE Proc.*, **53**, 2133.

Balland, B. B. and Barbottin, G. (1989). The trapping and detrapping of carriers injected into SiO_2. In *Instabilities in Silicon Devices*, vol. 2, ed. G. Barbottin and A. Vapaille. Amsterdam: North-Holland, chapter 10.

Bardeen, J. (1947). Surface states and rectification at a metal semiconductor contact, *Phys. Rev.* **71**, 717–727.

Berger, H. H. (1972). Models for contacts to planar devices, *Solid-State Electron.* **15**, 145–158.

Berglund, C. N. and Powell, R. J. (1971). Photoinjection into SiO_2: electron scattering in the image force potential well, *J. Appl. Phys.* **42**, 573–579.

Bhavnagarwala, A. J., Tang, X. and Meindl, J. D. (2001). The impact of intrinsic device fluctuations on CMOS SRAM cell stability, *IEEE J. Solid-State Circuits* **36**, 658–665.

Blakemore, J. S. (1982). Approximations for Fermi-Dirac integrals, especially the function $F_{1/2}(\eta)$ used to describe electron density in a semiconductor, *Solid-State Electronics* **25**, 1067–1076.

Blat, C. E., Nicollian, E. H. and Poindexter, E. H. (1991). Mechanism of negative-bias-temperature instability, *J. Appl. Phys.* **69**, 1712–1720.

Bourgoin, J. C. (1989). Radiation-induced defects in the Si-SiO_2 structure. In *Instabilities in Silicon Devices*, vol. 2, ed. G. Barbottin and A. Vapaille. Amsterdam: North-Holland, chapter 17.

Brews, J. R. (1978). A charge sheet model of the MOSFET, *Solid-State Electron.* **21**, 345–355.

Brews, J. R. (1979). Threshold shifts due to nonuniform doping profiles in surface channel MOSFETs, *IEEE Trans. Electron Devices* **ED-26**, 1696–1710.

Brews, J. R. (1981). Physics of the MOS transistor, in *Applied Solid State Science*, suppl. 2A, ed. D. Kahng. New York: Academic.

Bulucea, C. D. (1968). Diffusion capacitance of p–n junctions and transistors, *Electronics Lett.* **4**, 559–561.

Burghartz, J. N., Bartek, M., Rajaei, B., Sarro, P. M., Polyakov, A., Pham, N. P., Boullard, E. and Ng, K. T. (2002). Substrate options and add-on process modules for monolithic RF silicon technology, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, 17–23.

Burns, J. R. (1964). Switching response of complementary-symmetry MOS transistor logic circuits, *RCA Review* **25**, 627.

Cai, J. and Ning, T. H. (2006). SiGe HBTs on CMOS-compatible SOI. In *Silicon Heterostructure Handbook*, ed. J. D. Cressler. New York: CRC Press, 233–247.

Cai, J., Ajmera, A., Ouyang, C., Oldiges, P., Steigerwalt, M., Stein, K., Jenkins, K., Shahidi, G. and Ning, T. (2002a). Fully-depleted-collector polysilicon-emitter SiGe-base vertical bipolar transistor on SOI, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, 172–173.

Cai, J., Taur, Y., Huang, S.-F., Frank, D. J., Kosonocky, S. and Dennard, R. H. (2002b). Supply voltage strategies for minimizing the power of CMOS processors, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, 102–103.

Cai, J., Kumar, M., Steigerwalt, M., Ho, H., Schonenberg, K., Stein, K., Chen, H., Jenkins, K., Ouyang, Q., Oldiges, P. and Ning, T. (2003). Vertical SiGe-base bipolar transistors on CMOS-compatible SOI substrate, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, 215–218.

Cai, Y., Sato, T., Orshansky, M., Sylvester, D. and Hu, C. (2000). New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation, *Proc. Custom Integrated Circuit Conference (CICC)*, 201–204.

Caughey, D. M. and Thomas, R. E. (1967). Carrier mobilities in silicon empirically related to doping and field, *Proc. IEEE* **55**, 2192.

Chan, T. Y., Chen, J., Ko, P. K. and Hu, C. (1987a). The impact of gate-induced drain leakage current on MOSFET scaling, *IEEE IEDM Technical Digest*, 718–721.

Chan, T. Y., Wu, A. T., Ko, P. K. and Hu, C. (1987b). Effects of the gate-to-drain/source overlap on MOSFET characteristics, *IEEE Electron Device Lett.* **EDL-8**, 326–328.

Chan, T. Y., Young, K. K. and Hu, C. (1987c). A true single-transistor oxide-nitride-oxide EEPROM device, *IEEE Electron Device Lett.* **EDL-8**, 93–95.

Chang, C. Y., Fang, Y. K. and Sze, S. M. (1971). Specific contact resistance of metal–semiconductor barriers, *Solid-State Electron.* **14**, 541–550.

Chang, L. L., Stiles, P. J. and Esaki, L. (1967). Electron tunneling between a metal and a semiconductor: characteristics of Al–Al_2O_3–SnTe and –GeTe junctions, *J. App. Phys.* **38**, 4440–4445.

Chang, W.-H., Davari, B., Wordeman, M. R., Taur, Y., Hsu, C. C.-H. and Rodriguez, M. D. (1992). A high-performance 0.25-μm CMOS technology: I – design and characterization, *IEEE Trans. Electron Devices* **39**, 959–966.

Chen, I. C., Holland, S. and Hu, C. (1986). Oxide breakdown dependence on thickness and hole current-enhanced reliability of ultra-thin oxides, *IEEE IEDM Technical Digest*, 660–663.

Chen, T. C., Cressler, J. D., Toh, K. Y., Warnock, J., Lu, P. F., Jenkins, K. A., Basavaiah, S., Manny, M. P., Ng, H. Y., Tang, D. D., Li, G. P., Chuang, C. T., Polcari, M. R., Ketchen, M. B. and Ning, T. H. (1989). A submicron high performance bipolar technology, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, 87–88.

Chen, W., Taur, Y., Sadana, D., Jenkins, K. A., Sun, J. Y.-C. and Cohen, S. (1996). Suppression of the SOI floating-body effects by linked-body structure, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 92–93.

Chen, T., Bellini, M., Zhao, E., Comeau, J. P., Sutton, A. K., Grens, C. M., Cressler, J. D., Cai, J. and Ning, T. H. (2005). Substrate bias effects in vertical SiGe HBTs fabricated on CMOS-compatible thin film SOI, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 256–259.

Chern, J. G. J., Chang, P., Motta, R. F. and Godinho, N. (1980). A new method to determine MOSFET channel length, *IEEE Electron Device Lett.* **EDL-1**, 170–173.

Chiu, T.-Y., Chin, G. M., Lau, M. Y., Hanson, R. C., Morris, M. D., Lee, K. F., Voshchenkov, A. M., Swartz, R. G., Archer, V. D. and Finegan, S. N. (1987). A high speed super self-aligned bipolar-CMOS technology, *IEEE IEDM Technical Digest*, 24–27.

Chiu, T.-Y., Tien, P. K., Sung, J. and Liu, T.-Y. M. (1992). A new analytical model and the impact of base charge storage on base potential distribution, emitter current crowding and base resistance, *IEEE IEDM Technical Digest*, 573–576.

Chor, E. F., Ashburn, P. and Brunnschweiler, A. (1985). Emitter resistance of arsenic- and phosphorus-doped polysilicon emitter transistors, *IEEE Electron Device Lett.* **EDL-6**, 516–518.

Chor, E. F., Brunnschweiler, A. and Ashburn, P. (1988). A propagation-delay expression and its application to the optimization of polysilicon emitter ECL processes, *IEEE J. Solid-State Circuits* **23**, 251–259.

Chuang, C. T., Chin, K., Stork, J. M. C., Patton, G. L., Crabbé, E. F. and Comfort, J. H. (1992). On the leverage of high-f_T transistors for advanced high-speed bipolar circuits, *IEEE J. Solid-State Circuits* **27**, 225–228.

Chung, J. E., Jeng, M.-C., Moon, J. E., Ko, P.-K. and Hu, C. (1990). Low-voltage hot-electron currents and degradation in deep-submicrometer MOSFETs, *IEEE Trans. Electron Devices* **37**, 1651–1657.

Chynoweth, A. G. (1957). Ionization rates for electrons and holes in silicon, *Phys. Rev.* **109**, 1537–1540.

Chynoweth, A. G., Feldmann, W. L., Lee, C. A., Logan, R. A. and Pearson, G. L. (1960). Internal field emission at narrow silicon and germanium p–n junctions, *Phys. Rev.* **118**, 425–434.

Coen, R. W. and Muller, R. S. (1980). Velocity of surface carriers in inversion layers on silicon, *Solid-State Electron.* **23**, 35–40.

Colinge, J.-P. (2004). Multiple-gate SOI MOSFETs, *Solid-State Electron.* **48**, 897–905.

Colinge, J.-P., Hashimoto, K., Kamins, T., Chiang, S.-Y., Liu, E.-D., Peng, S. and Rissman, P. (1986). High-speed, low-power, implanted-buried-oxide CMOS circuits, *IEEE Electron Device Lett.* **EDL-7**, 279–281.

Cowley, A. M. and Sze, S. M. (1965). Surface states and barrier height of metal–semiconductor systems, *J. Appl. Phys.* **36**, 3212–3220.

Crabbé, E. F., Comfort, J. H., Lee, W., Cressler, J. D., Meyerson, B. S., Megdanis, A. C., Sun, J. Y.-C. and Stork, J. M. C. (1992). 73-GHz self-aligned SiGe-base bipolar transistors with phosphorus-doped polysilicon emitters, *IEEE Electron Device Lett.* **13**, 259–261.

Crabbé, E. F., Meyerson, B. S., Stork, J. M. C. and Harame, D. L. (1993a). Vertical profile optimization of very high frequency epitaxial Si- and SiGe-base bipolar transistors, *IEEE IEDM Technical Digest*, 83–86.

Crabbé, E. F., Cressler, J. D., Patton, G. L., Stork, J. M. C., Comfort, J. H. and Sun, J. Y.-C. (1993b). Current gain rolloff in graded-base SiGe Heterojunction bipolar transistors, *IEEE Electron Device Lett.* **14**, 193–195.

Cressler, J. D., Comfort, J. H., Crabbé, E. F., Patton, G. L., Stork, J. M. C., Sun, J. Y.-C. and Meyerson, B. S. (1993a). On the profile design and optimization of epitaxial Si- and SiGe-base bipolar technology for 77K applications – part I: transistor dc design considerations, *IEEE Trans. Electron Devices* **40**, 525–541.

Cressler, J. D., Crabbé, E. F., Comfort, J. H., Stork, J. M. C. and Sun, J. Y.-C. (1993b). On the profile design and optimization of epitaxial Si- and SiGe-base bipolar technology for 77K applications – part II: circuit performance issues, *IEEE Trans. Electron Devices* **40**, 542–554.

Crowell, C. R. (1965). The Richardson constant for thermionic emission in Schottky barrier diodes, *Solid-State Electron.* **8**, 395–399.

Crowell, C. R. (1969). Richardson constant and tunneling effective mass for thermionic and thermionic-field emission in Schottky barrier diodes, *Solid-State Electron.* **12**, 55–59.

Crowell, C. R. and Rideout, V. L. (1969). Normalized thermionic-field (T-F) emission in metal–semiconductor (Schottky) barriers, *Solid-State Electron.* **12**, 89–105.

Crowell, C. R. and Sze, S. M. (1966a). Current transport in metal-semiconductor barriers, *Solid-State Electron.* **9**, 1035–1048.

Crowell, C. R. and Sze, S. M. (1966b). Temperature dependence of avalanche multiplication in semiconductors, *Appl. Phys. Lett.* **9**, 242–244.

Crupi, F., Ciofi, C., Germano, A., Iannaccone, G., Stathis, J. H. and Lombardo, S. (2002). On the role of interface states in low-voltage leakage currents of metal–oxide–semiconductor structures, *Appl. Phys. Lett.* **80**, 4597–4599.

Davari, B., Ting, C. Y., Ahn, K. Y., Basavaiah, S., Hu, C. K., Taur, Y., Wordeman, M. R., Aboelfotoh, O., Krusin-Elbaum, L., Joshi, R. V. and Polcari, M. R. (1987). Submicron tungsten-gate MOSFET with 10-nm gate oxide, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 61–62.

Davari, B., Koburger, C., Furukawa, T., Taur, Y., Noble, W., Megdanis, A., Warnock, J. and Mauer, J. (1988). A variable-size shallow trench isolation technology with diffused sidewall doping for submicron CMOS, *IEEE IEDM Technical Digest*, 92–95.

Davari, B., Koburger, C. W., Schulz, R., Warnock, J. D., Furukawa, T., Jost, M., Taur, Y., Schwittek, W. G., DeBrosse, J. K., Kerbaugh, M. L. and Mauer, J. L. (1989). A new planarization technique using a combination of RIE and chemical mechanical polish, *IEEE IEDM Technical Digest*, 61–64.

Deal, B. E. (1980). Standardized terminology for oxide charges associated with thermally oxidized silicon, *IEEE Trans. Electron Devices* **27**, 606–608.

Deal, B. E., Sklar, M., Grove, A. S. and Snow, E. H. (1967). Characteristics of the surface-state charge of thermally oxidized silicon, *J. Electrochem. Soc.* **114**, 266–274.

Deal, B. E., MacKenna, E. L. and Castro, P. L. (1969). Characteristics of fast surface states associated with SiO_2–Si and Si_3N_4–Si structures, *J. Electrochem. Soc.* **116**, 997–1005.

de Graaff, H. C., Slotboom, J. W. and Schmitz, A. (1977). The emitter efficiency of bipolar transistors, *Solid-State Electron.* **20**, 515–521.

Degraeve, R., Groeseneken, G., Bellens, R., Depas, M. and Maes, H. E. (1995). A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxide, *IEEE IEDM Technical Digest*, 863–866.

Degraeve, R., Groeseneken, G., Bellens, R., Ogier, J. L., Depas, M., Roussel, P. J. and Maes, H. E. (1998). New insights in the relation between electron trap generation and the statistical properties of oxide breakdown, *IEEE Trans. Electron Devices* **43**, 904–911.

Deixler, P., Huizing, H. G. A., Donkers, J. J. T. M., Klootwijk, J. H., Hartskeerl, D., de Boer, W. B., Havens, R. J., van der Toorn, R., Paasschens, J. C. J., Kloosterman, W. J., van Berkum, J. G. M., Terpstra, D. and Slotboom, J. W. (2001). Explorations for high performance SiGe-heterojunction bipolar transistor integration, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 30–33.

del Alamo, J., Swirhun, S. and Swanson, R. M. (1985a). Measuring and modeling minority carrier transport in heavily doped silicon, *Solid-State Electron.* **28**, 47–54.

del Alamo, J., Swirhun, S. and Swanson, R. M. (1985b). Simultaneous measurement of hole lifetime, hole mobility and band-gap narrowing in heavily doped n-type silicon, *IEEE IEDM Technical Digest*, 290–293.

Dennard, R. H. (1968). Field-effect transistor memory, U. S. Patent 3,387,286 issued June 4.

Dennard, R. H. (1984). Evolution of the MOSFET dynamic RAM – a personal view, *IEEE Trans. Electron Devices* **ED-31**, 1549–1555.

Dennard, R. H. (1986). Scaling limits of silicon VLSI technology, in *The Physics and Fabrication of Microstructures and Microdevices*, ed. M. J. Kelly and C. Weisbuch. Berlin: Springer-Verlag.

Dennard, R. H., Gaensslen, F. H., Yu, H. N., Rideout, V. L., Bassous, E. and LeBlanc, A. R. (1974). Design of ion-implanted MOSFETs with very small physical dimensions, *IEEE J. Solid-State Circuits* **SC-9**, 256–268.

Depas, M., Nigam, T. and Heyns, M. H. (1996). Soft breakdown of ultrathin gate oxide layers, *IEEE Trans. Electron Devices* **43**, 1499–1504.

DiMaria, D. J. (1987). Correlation of trap creation with electron heating in silicon dioxide, *Appl. Phys. Lett.* **51**, 655–657.

DiMaria, D. J. and Cartier, E. (1995). Mechanism for stress-induced leakage currents in thin silicon dioxide films, *J. Appl. Phys.* **78**, 3883–3894.

DiMaria, D. J. and Stathis, J. H. (1997). Explanation for the oxide thickness dependence of breakdown characteristics of metal-oxide semiconductor structures, *Appl. Phys. Lett.* **70**, 2708–2710.

DiMaria, D. J., Cartier, E. and Arnold, D. (1993). Impact ionization, trap creation, degradation, and breakdown in silicon dioxide films on silicon, *J. Appl. Phys.* **73**, 3367–3384.

DiStefano, T. H. and Shatzkes, M. (1974). Impact ionization model for dielectric instability and breakdown, *Appl. Phys. Lett.* **25**, 685–687.

Dubois, E., Bricout, P.-H. and Robilliart, E. (1994). Accuracy of series resistances extraction schemes for polysilicon bipolar transistors, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 148–151.

Duvernay, J., Brossard, F., Borot, G., Boissonnet, L., Vandelle, B., Rubaldo, L., Deléglise, F., Avenier, G., Chevalier, P., Rauber, B., Dutartre, D. and Chantre, A. (2007). Development of a self-aligned pnp HBT for a complementary thin-SOI SiGeC BiCMOS technology, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 34–37.

Dziewior, J. and Schmid, W. (1977). Auger coefficients for highly doped and highly excited silicon, *Appl. Phys. Lett.* **31**, 346–348.

Dziewior, J. and Silber, D. (1979). Minority-carrier diffusion coefficients in highly doped silicon, *Appl. Phys. Lett.* **35**, 170–172.

Early, J. M. (1952). Effects of space-charge layer widening in junction transistors, *Proc. IRE* **40**, 1401–1406.

Ebers, J. J. and Moll, J. L. (1954). Large-signal behavior of junction transistors, *Proc. IRE* **42**, 1761–1772.

Eichelberger, E. B. and Bello, S. E. (1991). Differential current switch – high performance at low power, *IBM J. Res. Develop.* **35**, 313–320.

Eitan, B., Pavan, P., Bloom, I., Aloni, E., Frommer, A. and Finzi, D. (2000). NROM: a novel localized trapping, 2-bit nonvolatile memory cell, *IEEE Electron Device Lett.* **21**, 543–545.

El-Mansy, Y. A. and Boothroyd, A. R. (1977). A simple two-dimensional model for IGFET, *IEEE Trans. Electron Devices* **ED-24**, 254–262.

EMIS Datareviews Series No. 4 (1988). *Properties of Silicon*. London: INSPEC, The Institute of Electrical Engineers.

Fair, R. B. and Wivell, H. W. (1976). Zener and avalanche breakdown in As-implanted low voltage Si n–p junctions, *IEEE Trans. Electron Devices* **ED-23**, 512–518.

Farber, A. S. and Schlig, E. S. (1972). A novel high-performance bipolar monolithic memory cell, *IEEE J. Solid-State Circuits* **SC-7**, 297–298.

Filensky, W. and Beneking, H. (1981). New technique for determination of static emitter and collector series resistances of bipolar transistors, *Electron. Lett.* **17**, 503–504.

Fischetti, M. V. and Laux, S. E. (1996). Band structure, deformation potentials, and carrier mobility in strained Si, Ge, and SiGe alloys, *J. Appl. Phys.* **80** (4), 2234–2252.

Fischetti, M. V., Laux, S. E. and Crabbé, E. (1995). Understanding hot-electron transport in silicon devices: Is there a shortcut? *J. Appl. Phys.* **78**, 1058–1087.

Fischetti, M. V., Wang, L., Yu, B., Sachs, C., Asbeck, P. M., Taur, Y. and Rodwell, M. (2007). Simulation of electron transport in high-mobility MOSFETs: density of states bottleneck and source starvation, *IEEE IEDM Technical Digest*, 109–112.

Frank, D. J., Laux, S. E. and Fischetti, M. V. (1992). Monte Carlo simulation of a 30 nm dual-gate MOSFET: how short can silicon go? *IEEE IEDM Technical Digest*, 553–556.

Frank, D. J., Taur, Y. and Wong, H.-S. (1998). Generalized scale length for two-dimensional effects in MOSFETs, *IEEE Electron Device Lett.* **19**, 385–387.

Frank, D. J., Taur, Y., Ieong, M. and Wong, H.-S. P. (1999). Monte Carlo modeling of threshold variation due to dopant fluctuations, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 171–172.

Frank, D. J., Dennard, R. H., Nowak, E., Solomon, P. M., Taur, Y. and Wong, H.-S. P. (2001). Device scaling limits of Si MOSFETs and their application dependencies, *IEEE Proc.* **89**, 259–288.

Frohman-Bentchkowsky, D. (1971). A fully decoded 2048-bit electrically programmable FAMOS read-only memory, *IEEE J. Solid-State Circuits* **SC-6**, 301–306.

Gaensslen, F. H., Rideout, V. L., Walker, E. J. and Walker, J. J. (1977). Very small MOSFETs for low temperature operation, *IEEE Trans. Electron Devices* **ED-24**, 218–229.

Gautier, J., Jenkins, K. A. and Sun, J. Y.-C. (1995). Body charge related transient effects in floating-body SOI nMOSFETs, *IEEE IEDM Technical Digest*, 623–626.

Ghandhi, S. K. (1968). *The Theory and Practice of Microelectronics*. New York: Wiley.

Gildenblat, G., Li, X., Wu, W., Wang, H., Jha, A., van Langevelde, R., Smit, G. D. J., Scholten, A. J. and Klaassen, D. B. M., (2006). PSP: An advanced surface-potential-based MOSFET model for circuit simulation, *IEEE Trans. Electron Devices* **ED-53**, 1979–1993.

Good, R. H. Jr., and Müller, E. W. (1956). In *Handbuck der Physik*, vol. XXI. Berlin: Springer-Verlag, pp. 176–231.

Goodman, A. M. (1966). Photoemission of holes from silicon into silicon dioxide, *Phys. Rev.* **152**, 780–784.

Grant, W. H. (1973). Electron and hole ionization rates in epitaxial silicon at high electric fields, *Solid-State Electron.* **16**, 1189–1203.

Gray, P. E., DeWitt, D., Boothroyd, A. R. and Gibbons, J. F. (1964). *Physical Electronics and Circuit Models of Transistors*, vol. 2, Semiconductor Electronics Education Committee. New York: Wiley.

Green, M. A. (1990). Intrinsic concentration, effective densities of states, and effective mass in silicon, *J. Appl. Phys.* **67**, 2944–2954.

Grove, A. S. (1967). *Physics and Technology of Semiconductor Devices*. New York: Wiley.

Grove, A. S. and Fitzgerald, D. J. (1966). Surface effects on p–n junctions – characteristics of surface space-charge regions under non-equilibrium conditions, *Solid-State Electron.* **9**, 783–806.

Gummel, H. K. (1961). Measurement of the number of impurities in the base layer of a transistor, *Proc. IRE* **49**, 834.

Gummel, H. K. (1967). Hole-electron product of p–n junctions, *Solid-State Electron.* **10**, 209–212.

Guo, J. C., Chung, S. S. and Hsu, C. C. (1994). A new approach to determine the effective channel length and the drain-and-source series resistance of miniaturized MOSFETs, *IEEE Trans. Electron Devices* **ED-41**, 1811–1818.

Gupta, M. S. (1992). Power gain in feedback amplifiers, a classic revisited, *IEEE Trans. Microwave Theory Techniques* **40**, 864–879.

Hafez, W., Lai, J.-W. and Feng, M. (2003). Submicron InP-InGaAs single heterojunction bipolar transistors with f_T of 377 GHz, *IEEE Electron Device Lett.* **24**, 292–294.

Hall, R. N. (1952). Electron–hole recombination in germanium, *Phys. Rev.* **87**, 387.

Harame, D. L., Comfort, J. H., Cressler, J. D., Crabbé, E. F., Sun, J. Y.-C., Meyerson, B. S. and Tice, T. (1995a). Si/SiGe epitaxial-base transistors – part I: materials, physics, and circuits, *IEEE Trans. Electron Devices* **42**, 455–468.

Harame, D. L., Comfort, J. H., Cressler, J. D., Crabbé, E. F., Sun, J. Y.-C., Meyerson, B. S. and Tice, T. (1995b). Si/SiGe epitaxial-base transistors – part II: process integration and analog applications, *IEEE Trans. Electron Devices* **42**, 469–482.

Harari, E. (1978). Dielectric breakdown in electrically stressed thin films of thermal SiO_2, *J. Appl. Phys.* **49**, 2478–2489.

Hashimoto, T., Kikuchi, T., Watanabe, K., Ohashi, N., Saito, T., Yamaguchi, H., Wada, S., Natsuaki, N., Konda, M., Konda, S., Homma, Y., Owada, N. and Ikeda, T. (1998). A 0.2μm bipolar-CMOS technology on bonded SOI with copper metallization for ultra high-speed processors, *IEEE IEDM Technical Digest*, 209–212.

Hauser, J. R. (1964). The effects of distributed base potential on emitter current injection density and effective base resistance for stripe transistor geometries, *IEEE Trans. Electron Devices* **ED-11**, 238–242.

Hauser, J. R. (1968). Bipolar transistors. In *Fundamentals of Silicon Integrated Device Technology*, vol. II, ed. R. M. Burger and R. P. Donovan. Englewood Cliffs: Prentice-Hall.

Hayes, J. R., Capasso, F., Gossard, A. C., Malik, R. J. and Wiegmann, W. (1983). Bipolar transistor with graded band-gap base, *Electronics Lett.***19**, 410–411.

Hedenstierna, N. and Jeppson, K. O. (1987). CMOS circuit speed and buffer optimization, *IEEE Trans. Computer-Aided Design*, **CAD-6**, 270.

Heine, V. (1965). Theory of surface states, *Phys. Rev.* **138**, A1689–A1696.

Henisch, H. K. (1984). *Semiconductor Contacts*. Oxford: Clarendon Press.

Hill, C. F. (1968). Noise margin and noise immunity in logic circuits, *Microelectronics*, 16–21.

Hillen, M. W. and Verwey, J. F. (1986). Mobile ions in SiO_2 layers on Si. In *Instabilities in Silicon Devices*, vol. 1, ed. G. Barbottin and A. Vapaille. Amsterdam: North-Holland, chapter 8.

Hiramoto, T., Tamba, N., Yoshida, M., Hashimoto, T., Fujiwara, T., Watanabe, K., Usami, M. and Ikeda, T. (1992). A 27-Ghz double polysilicon bipolar technology on bonded SOI with embedded 58μm^2 CMOS cells for ECL-CMOS SRAM applications, *IEEE IEDM Technical Digest*, 39–42.

Hisamoto, D., Lee, W.-C., Kedzierski, J., Takeuchi, H., Asano, K., Kuo, C., Anderson, E., King, T.-J., Bokor, J. and Hu, C. (2000). FinFET – a self-aligned double-gate MOSFET scalable to 20 nm, *IEEE Trans. Electron Devices* **47**, 2320–2325.

Hobart, K. D., Kub, F. J., Papanicoloau, N. A., Kruppa, W. and Thompson, P. E. (1995). $Si/Si_{1-x}Ge_x$ heterojunction bipolar transistors for microwave power applications, *J. Crystal Growth* **157**, 215–220.

Hodges, D. A., editor (1972). *Semiconductor Memories*. New York: IEEE Press.

Hsu, C. C.-H. and Ning, T. H. (1991). Voltage and temperature dependence of interface trap generation by hot electrons in p- and n-poly gated MOSFETs, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 17–18.

Hsu, C. C.-H., Acovic, A., Dori, L., Wu, B., Lii, T., Quinlan, D., DiMaria, D., Taur, Y., Wordeman, M. R. and Ning, T. H. (1992). A high-speed low-power p-channel flash EEPROM using silicon-rich oxide as tunneling dielectric, *Int. Conf. Solid-State Devices and Materials, Extended Abstract*, 140–142.

Hu, C., editor (1991). *Nonvolatile Semiconductor Memories Technologies, Design, and Applications*. New York: IEEE Press.

Hu, C., Tam, S. C., Hsu, F.-C., Ko, P.-K., Chan, T.-Y. and Terrill, K. W. (1985). Hot-electron-induced MOSFET degradation – model, monitor, and improvement, *IEEE Trans. Electron Devices* **32**, 375–385.

Hu, S. M. and Schmidt, S. (1968). Interactions in sequential diffusion processes in semiconductors, *J. Appl. Phys.* **39**, 4272–4283.

Hui, J., Wong, S. and Moll, J. (1985). Specific contact resistivity of $TiSi_2$ to p^+ and n^+ junctions, *IEEE Electron Device Lett.* **EDL-6**, 479–481.

Huizing, H. G. A., Klootwijk, J. H., Aksen, E. and Slotboom, J. W. (2001). Base current tuning in SiGe HBT's by SiGe in the emitter, *IEEE IEDM Technical Digest*, 899–902.

Hurkx, G. A. M. (1994). Bipolar transistor physics. In *Bipolar and Bipolar-MOS Integration*, ed. P. A. H. Hart. Amsterdam: Elsevier.

Hurkx, G. A. M. (1996). The relevance of f_T and f_{max} for the speed of a bipolar CE amplifier stage, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 53–56.

Iinuma, T., Itoh, N., Nakajima, H., Inou, K., Matsuda, S., Yoshino, C., Tsuboi, Y., Katsumata, Y. and Iwai, H. (1995). Sub-20 ps high-speed ECL bipolar transistor with low parasitic architecture, *IEEE Trans. Electron Devices* **42**, 399–405.

Im, H.-J., Ding, Y. and Pelz, J. P. (2001). Nanometer-scale test of the Tung model of Schottky-barrier height inhomogeneity, *Phys. Rev. B* **64**, 0753101–0753109.

Irvin, J. C. and Vanderwal, N. C. (1969). Schottky barrier devices. In *Microwave Semiconductor Devices and Their Circuit Applications*, ed. H. A. Watson. New York: McGraw-Hill, pp. 349–369.

Itoh, K. (2001). *VLSI Memory Chip Design*. Berlin: Springer.

Itoh, Y., Momodomi, M., Shirota, R., Iwata, Y., Nakayama, R., Kirisawa, R., Tanaka, T., Toita, K., Inoue, S. and Masuoka, F. (1989). An experimental 4Mb CMOS EEPROM with a NAND structured cell, *IEEE ISSCC Digest of Technical Papers*, 134–135.

ITRS (1999). *The International Technology Roadmap for Semiconductors*, published by the Semiconductor Industry Association, San Jose, California.

ITRS (2007). *The International Technology Roadmap for Semiconductors*, published by the Semiconductor Industry Association, San Jose, California.

Iyer, S. S., Patton, G. L., Delage, S. S., Tiwari, S. and Stork, J. M. C. (1987). Silicon-Germanium base heterojunction bipolar transistors by molecular beam epitaxy," *IEEE IEDM Technical Digest*, 874–876.

Jagannathan, B., Meghelli, M., Chan, K., Rieh, J.-S., Schonenberg, K., Ahlgren, D., Subbanna, S. and Freeman, G. (2003). 3.9 ps SiGe HBT ECL ring oscillator and transistor design for minimum gate delay, *IEEE Electron Device Lett.* **24**, 324–326.

Jeppson, K. O. and Svensson, C. M. (1977). Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices, *J. Appl. Phys.* **48**, 2004–2014.

Jiménez, D., Iñiguez, B., Suñé, J., Marsal, L. F., Pallarès, J., Roig, J. and Flores, D. (2004). Continuous analytic I–V model for surrounding-gate MOSFETs, *IEEE Electron Device Lett.* **25**, 571–573.

Jund, C. and Poirier, R. (1966). Carrier concentration and minority carrier lifetime measurement in semiconductor epitaxial layers by the MOS capacitance method, *Solid-State Electron.* **9**, 315–318.

Kaczer, B., Degraeve, R., Groeseneken, G., Rasras, M., Kubicek, S., Vandamme, E. and Badenes, B. (2000). Impact of MOSFET oxide breakdown on digital circuit operation and reliability, *IEEE IEDM Technical Digest*, 553–556.

Kahng, D. and Atalla, M. M. (1960). Silicon dioxide field surface devices, presented at IEEE Device Research Conference, Pittsburgh.

Kane, E. O. (1961). Theory of tunneling, *J. Appl. Phys.* **32**, 83–91.

Kannan, K. A. (2005). The effect of very thick and high-k gate insulators on short-channel effects of MOS transistors, Master thesis, Dept. Electrical and Computer Engneering, University of California, San Diego.

Kapoor, A. K. and Roulston, D. J., editors (1989). *Polysilicon Emitter Bipolar Transistors*. New York: IEEE Press.

Kay, L. E. and Tang, T.-W. (1991). Monte Carlo calculation of strained and unstrained electron mobilities in $Si_{1-x}Ge_x$ using an improved ionized-impurity model, *J. Appl. Phys.* **70**, 1483–1488.

Kerwin, R. E., Klein, D. L. and Sarace, J. C. (1969). Method for making MIS structures, U. S. Patent 3,475,234 issued Oct. 28.

Kesan, V. P., Subbanna, S., Restle, P. J., Tejwani, M. J., Aitken, J. M., Iyer, S. S. and Ott, J. A. (1991). High performance 0.25 μm p-MOSFETs with silicon–germanium channels for 300 K and 77 K operation, *IEEE IEDM Technical Digest*, 25–28.

Kianian, S., Levi, A., Lee, D. and Hu, Y.-W. (1994). A novel 3 volts-only, small sector erase, high density flash E^2PROM, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 71–72.

Kim, I.-G., Yanagidarira, K. and Hiramoto, T. (2003). Integration of fluorinated nano-crystal memory cells with $4.6F^2$ size by landing plug polysilicon contact and direct-tungsten bitline, *IEEE IEDM Technical Digest*, 605–608.

Kim, K. (2008). Future memory technology: challenges and opportunities, *Proceedings of Symp. on VLSI Technology, Systems, and Applications*, 5–9.

Kircher, C. J. (1975). Comparison of leakage currents in ion-implanted and diffused p–n junctions, *J. Appl. Phys.* **46**, 2167–2173.

Kirk Jr., C. T. (1962). A theory of transistor cutoff frequency (f_T) falloff at high current densities, *IEEE Trans. Electron Devices* **ED-9**, 164–174.

Klaassen, D. B. M. (1990). A unified mobility model for device simulation, *IEEE IEDM Technical Digest*, 357–360.

Klaassen, D. B. M., Slotboom, J. W. and de Graaff, H. C. (1992). Unified apparent bandgap narrowing in n- and p-type silicon, *Solid-State Electron.* **35**, 125–129.

Kolhatkar, J. S. and Dutta, A. K. (2000). A new substrate current model for submicron MOSFET's, *IEEE Trans. Electron Devices* **47**, 861–863.

Kondo, M., Kobayashi, T. and Tamaki, Y. (1995). Hetero-emitter-like characteristics of phosphorus doped polysilicon emitter transistors – part I: band structure in the polysilicon emitter obtained from electrical measurements, *IEEE Trans. Electron Devices* **42**, 419–426.

Kondo, M., Shimamoto, H. and Washio, K. (2001). Variation in emitter diffusion depth by $TiSi_2$ formation on polysilicon emitters of Si bipolar transistors, *IEEE Trans. Electron Devices* **48**, 2108–2117.

Koyanagi, M., Sunami, H., Hashimoto, N. and Ashikawa, M. (1978). Novel high density stacked capacitor MOS RAM, *IEEE IEDM Technical Digest*, 348–351.

Kroemer, H. (1957). Theory of a wide-gap emitter for transistors, *Proc. IRE* **45**, 1535–1537.

Kroemer, H. (1985). Two integral relations pertaining to the electron transport through a bipolar transistor with a non-uniform energy gap in the base region, *Solid-State Electron.* **28**, 1101–1103.

Kuno, H. J. (1964). Analysis and characterization of p–n junction diode switching, *IEEE Trans. Electron Devices* **ED-11**, 8–14.

Kunz, V. D., de Groot, C. H., Hall, S., Anteney, I. M., Abdul-Rahim, A. I. and Ashburn, P. (2002). Application of polycrystalline SiGe for gain control in SiGe HBTs, *Proc. European Solid-State Device Research Conf. (ESSDERC)*, 171–174.

Kunz, V. D., de Groot, C. H., Hall, S. and Ashburn, P. (2003). Polycrystalline silicon–germanium emitters for gain control, with application to SiGe HBTs, *IEEE Trans. Electron Devices* **50**, 1480–1486.

Kyung, M. (2005). Charge sheet models for MOSFET current: Examining their deviations from the Pao–Sah integral, Master thesis, Dept. Electrical and Computer Engineering, University of California, San Diego.

Lanzerotti, L. D., Sturm, J. C., Stach, E., Hull, R., Buyuklimanli, T. and Magee, C. (1996). Suppression of boron out diffusion in SiGe HBTs by carbon incorporation, *IEEE IEDM Technical Digest*, 249–252.

La Rosa, G., Guarin, F., Rauch, S., Acovic, A., Lukaitis, J. and Crabbé, E. (1997). NBTI-channel hot carrier effects in pMOSFETs in advanced CMOS technologies, *Proc. IEEE Int. Reliability Phys. Symp.*, 282–286.

Laux, S. E. (1984). Accuracy of an effective channel length/external resistance extraction algorithm for MOSFETs, *IEEE Trans. Electron Devices* **ED-31**, 1245–1251.

Laux, S. E. and Fischetti, M. V. (1988). Monte-Carlo simulation of submicrometer Si n-MOSFET's at 77 and 300 K, *IEEE Electron Device Lett.* **EDL-9**, 467–469.

Lax, M. (1960). Cascade capture of electrons in solids, *Phys. Rev.* **119**, 1502–1523.

Lee, J.-H., Kang, W.-G., Lyu, J.-S. and Lee, J. D. (1996). Modeling of the critical current density of bipolar transistor with retrograde collector doping profile, *IEEE Electron Device Lett.* **17**, 109–111.

Lenzlinger, M. and Snow, E. H. (1969). Fowler–Nordheim tunneling into thermally grown SiO_2, *J. Appl. Phys.* **40**, 278–283.

Li, G. P., Chuang, C. T., Chen, T. C. and Ning, T. H. (1987). On the narrow-emitter effect of advanced shallow-profile bipolar transistors, *IEEE Trans. Electron Devices* **35**, 1942–1950.

Li, G. P., Hackbarth, E. and Chen, T.-C. (1988). Identification and implication of a perimeter tunneling current component in advanced self-aligned bipolar transistors, *IEEE Trans. Electron Devices* **ED-35**, 89–95.

Li, M. F., He, Y. D., Ma, S. G., Cho, B.-J., Lo, K. F. and Xu, M. Z. (1999). Role of hole fluence in gate oxide breakdown, *IEEE Electron Devices Lett.* **20**, 586–588.

Liang, X. and Taur, Y. (2004). A 2-D analytical solution for SCEs in DG MOSFETs, *IEEE Trans. Electron Devices* **ED-51**, 1385–1391.

Lim, H.-K., and Fossum, J. G. (1984). Current-voltage characteristics of thin-film SOI MOSFETs in strong inversion, *IEEE Trans. Electron Devices* **ED-31**, 401–408.

Linder, B. P., Frank, D. J., Stathis, J. H. and Cohen, S. A. (2001). Transistor-limited constant voltage stress of gate dielectrics, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 93–94.

Linder, B. P., Lombardo, S., Stathis, J. H., Vayshenker, A. and Frank, D. J. (2002). Voltage dependence of hard breakdown growth and the reliability implication in thin dielectrics, *IEEE Electron Device Lett.* **23**, 661–663.

Lindmayer, J. and Wrigley, C. Y. (1965). *Fundamentals of Semiconductor Devices*. New York: Van Norstrand Reinhold.

Liu, M., Cai, M. and Taur, Y. (2006). Scaling limit of CMOS supply voltage from noise margin considerations, *Proceedings of 2006 International Conference on Simulation of Semiconductor Process and Devices (SISPAD)*, 287–289.

Lo, S.-H., Buchanan, D. A., Taur, Y. and Wang, W. (1997). Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's, *IEEE Electron Device Lett.* **18**, 209–211.

Lochtefeld, A. and Antoniadis, D. A. (2001). On experimental determination of carrier velocity in deeply scaled NMOS: how close to the thermal limit?, *IEEE Electron Device Lett.* **22**, 95–97.

Lohstroh, J., Seevinck, E. and de Groot, J. (1983). Worst-case static noise margin criteria for logic circuits and their mathematical equivalence, *IEEE J. Solid-State Circuits* **SC-18**, 803–807.

Lombardo, S., Stathis, J. H. and Linder, B. P. (2003). Breakdown transient in ultrathin gate oxides: transition in the degradation rate, *Phys. Rev. Lett.* **90**, 1676011–1676014.

Lu, H., Lu, W.-Y. and Taur, Y. (2008). Effect of body doping on double-gate MOSFET characteristics, *Semiconductor Science and Technology* **22**, 252835 (6 pp).

Lu, N., Cottrell, P., Craig, W., Dash, S., Critchlow, D., Mohler, R., Machesney, B., Ning, T., Noble, W., Parent, R., Scheuerlein, R., Sprogis, E. and Terman, L. (1985). The SPT cell – a new substrate-plate trench cell for DRAMs, *IEEE IEDM Technical Digest*, 771–772.

Lu, N. C. C., Cottrell, P. E., Craig, W. J., Dash, S., Critchlow, D. L., Mohler, R. L., Machesney, B. J., Ning, T. H., Noble, W. P., Parent, R. M., Scheuerlein, R. E., Sprogis, E. J. and Terman, L. M. (1986). A substrate-plate trench-capacitor (SPT) memory cell for dynamic RAM's, *IEEE J. Solid-State Circuits* **SC-21**, 627–734.

Lu, P.-F., and Chen, T.-C. (1989). Collector-base junction avalanche effects in advanced double-poly self-aligned bipolar transistors, *IEEE Trans. Electron Devices* **ED-36**, 1182–1188.

Lu, P.-F., Li, G. P. and Tang, D. D. (1987). Lateral encroachment of extrinsic base dopant in submicron bipolar transistors, *IEEE Electron Device Lett.* **8**, 496–498.

Lu, P.-F., Comfort, J. H., Tang, D. D., Meyerson, B. S. and Sun, J. Y.-C. (1990). The implementation of a reduced-field profile design for high-performance bipolar transistors, *IEEE Electron Device Lett.* **11**, 336–338.

Lu, W.-Y., and Taur, Y. (2006). On the scaling limit of ultrathin SOI MOSFETs, *IEEE Trans. Electron Devices* **ED-53**, 1137–1141.

Lundstrom, M. (1997). Elementary scattering theory of the Si MOSFET, *IEEE Electron Device Lett.* **18**, 361–363.

Lynes, D. J. and Hodges, D. A. (1970). Memory using diode-coupled bipolar transistor cells, *IEEE J. Solid-State Circuits* **SC-5**, 186–191.

Manku, T. and Nathan, A. (1992). Electron drift mobility model for devices based on unstrained and coherently strained $Si_{1-x}Ge_x$ grown on ⟨001⟩ silicon substrate, *IEEE Trans. Electron Devices* **39**, 2082–2089.

Many, A., Goldstein, Y. and Grover, N. B. (1965). *Semiconductor Surfaces*. New York: John Wiley & Sons.

Martinet, B., Romagna, F., Kermarrec, O., Campidelli, Y., Saguin, F., Baudry, H., Marty, M., Dutartre, D. and Chantre, A. (2002). An investigation of the static and dynamic characteristics of high speed SiGe:C HBTs using a poly-SiGe emitter, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 147–150.

Mason, S. J. (1954). Power gain in feedback amplifiers, *Trans. IRE Professional Group on Circuit Theory*, **CT-1**, 20–25.

Masuoka, F., Asano, M., Iwahashi, H., Komuro, T. and Tanaka, S. (1984). A new flash E^2 PROM cell using triple polysilicon technology, *IEEE IEDM Technical Digest*, 464–467.

Masuoka, F., Momodomi, M., Iwata, Y. and Shirota, R. (1987). New ultra high density EPROM and flash EEPROM with NAND structure cell, *IEEE IEDM Technical Digest*, 552–555.

Mayumi, H., Nokubo, J., Okada, K. and Shiba, H. (1974). A 25-ns read access bipolar 1-kbit TTL RAM, *IEEE J. Solid-State Circuits* **SC-9**, 283–284.

Mead, C. A. and Spitzer, W. G. (1964). Fermi level position at metal–semiconductor interfaces, *Phys. Rev.* **134**, A713–A716.

Meyer, C. S., Lynn, D. K. and Hamilton, D. J., editors (1968). *Analysis and Design of Integrated Circuits*. New York: McGraw-Hill.

Meyer, R. G. and Muller, R. S. (1987). Charge-control analysis of the collector–base space-charge-region contribution to bipolar-transistor time constant τ_f, *IEEE Trans. Electron Devices* **ED-34**, 450–452.

Mii, Y., Wind, S., Taur, Y., Lii, Y., Klaus, D. and Bucchignano, J. (1994). An ultra-low power 0.1 μm CMOS, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 9–10.

Miller, S. L. (1955). Avalanche breakdown in Germanium, *Phys. Rev.* **99**, 1234–1241.

Mizuno, T., Okamura, J. and Toriumi, A. (1994). Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's, *IEEE Trans. Electron Devices* **41**, 2216–2221.

Moll, J. L. (1964). *Physics of Semiconductors*. New York: McGraw-Hill.

Moll, J. L. and Ross, I. M. (1956). The dependence of transistor parameters on base resistivity, *Proc. IRE* **44**, 72–78.

Muller, R. S. and Kamins, T. I. (1977). *Device Electronics for Integrated Circuits*. New York: Wiley.

Na, M. H., Nowak, E. J., Haensch, W. and Cai, J. (2002). The effective drive current in CMOS inverters, *IEEE IEDM Technical Digest*, 121–124.

Nakamura, T. and Nishizawa, H. (1995). Recent progress in bipolar transistor technology, *IEEE Trans. Electron Devices* **ED-42**, 390–398.

Nanba, M., Uchino, T., Kondo, M., Nakamura, T., Kobayashi, T., Tamaki, Y. and Tanabe, M. (1993). A 64-GHz f_T and 3.6-V BV_{CEO} Si bipolar transistor using *in situ* phosphorus-doped and large-grained polysilicon emitter contacts, *IEEE Trans. Electron Devices* **ED-40**, 1563–1565.

Naruke, K., Yamada, S., Obi, E., Taguchi, S. and Wada, M. (1989). A new flash-erase EEPROM cell with a sidewall select-gate on its source side, *IEEE IEDM Technical Digest*, 603–606.

Natori, K. (1994). Ballistic metal-oxide-semiconductor field effect transistor, *J. Appl. Phys.* **76**, 4879–4890.

Ng, K. K. and Lynch, W. T. (1986). Analysis of the gate-voltage dependent series resistance of MOSFETs, *IEEE Trans. Electron Devices* **ED-33**, 965–972.

Nguyen, T. N. (1984). Small-geometry MOS transistors: physics and modeling of surface- and buried-channel MOSFETs. Unpublished Ph. D. thesis, Stanford University.

Nguyen, T. N. and Plummer, J. D. (1981). Physical mechanisms responsible for short-channel effects in MOS devices, *IEEE IEDM Technical Digest*, 596–599.

Nicollian, E. H., Berglund, C. N., Schmidt, P. F. and Andrews, J. M. (1971). Electrochemical charging of thermal SiO_2 films by injected electron currents, *J. Appl. Phys.* **42**, 5654–5664.

Nicollian, E. H. and Brews, J. R. (1982). *MOS Physics and Technology*. New York: Wiley.

Ning, T. H. (1976a). Capture cross section and trap concentration of holes in silicon dioxide, *J. Appl. Phys.* **47**, 1079–1081.

Ning, T. H. (1976b). High-field capture of electrons by Coulomb-attractive centers in silicon dioxide, *J. Appl. Phys.* **47**, 3203–3208.

Ning, T. H. (1978). Electron trapping in SiO_2 due to electron-beam deposition of aluminum, *J. Appl. Phys.* **49**, 4077–4082.

Ning, T. H. (1995). Trade-offs between SiGe and GaAs bipolar ICs, *Proc. Fourth International Conference on Solid State and Integrated-Circuit Technology*, 434–438.

Ning, T. H. (2001). History and future perspective of the modern silicon bipolar transistor, *IEEE Trans. Electron Devices* **48**, 2485–2491.

Ning, T. H. (2003a). Polysilicon-emitter SiGe-base bipolar transistors – what happens when Ge gets into the emitter? *IEEE Trans. Electron Devices* **50**, 1346–1352.

Ning, T. H. (2003b). Silicon technology – emerging trends from a system application perspective, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 5–8.

Ning, T. H. and Isaac, R. D. (1980). Effect of emitter contact on current gain of silicon bipolar devices, *IEEE Trans. Electron Devices* **ED-27**, 2051–2055.

Ning, T. H. and Sah, C. T. (1971). Multivalley effective-mass approximation for donor states in silicon, I. Shallow-level group-V impurities, *Phys. Rev.* **B4**, 3468–3481.

Ning, T. H. and Tang, D. D. (1984). Method for determining the emitter and base series resistances of bipolar transistors, *IEEE Trans. Electron Devices* **ED-31**, 409–412.

Ning, T. H. and Yu, H. N. (1974). Optically induced injection of hot electrons into SiO_2, *J. Appl. Phys.* **45**, 5373–5378.

Ning, T. H., Osburn, C. M. and Yu, H. N. (1975). Electron trapping at positively charged centers in SiO_2, *Appl. Phys. Lett.* **26**, 248–250.

Ning, T. H., Osburn, C. M. and Yu, H. N. (1976). Threshold instability in IGFET's due to emission of leakage electrons from silicon substrate into silicon dioxide, *Appl. Phys. Lett.* **29**, 198–200.

Ning, T. H., Osburn, C. M. and Yu, H. N. (1977a). Emission probability of hot electrons from silicon into silicon dioxide, *J. Appl. Phys.* **48**, 286–293.

Ning, T. H., Osburn, C. M. and Yu, H. N. (1977b). Effect of electron trapping on IGFET characteristics, *J. Electronic Materials* **6**, 65–76.

Ning, T. H., Cook, P. W., Dennard, R. H., Osburn, C. M., Schuster, S. E. and Yu, H. N. (1979). 1-μm MOSFET VLSI technology: part IV. hot-electron design constraints, *IEEE Trans. Electron Devices* **ED-26**, 346–353.

Ning, T. H., Isaac, R. D., Solomon, P. M., Tang, D. D., Yu, H. N., Feth, G. C. and Wiedmann, S. K. (1981). Self-aligned bipolar transistors for high-performance and low-power-delay VLSI, *IEEE Trans. Electron Devices* **ED-28**, 1010–1013.

Nissan-Cohen, Y. (1986). A novel floating-gate method for measurement of ultra-low hole and electron gate currents in MOS transistors, *IEEE Electron Devices Lett.* **EDL-7**, 561–563.

Niu, G., Zhang, S., Cressler, J. D., Joseph, A. J., Fairbanks, J. S., Larson, L. E., Webster, C. S., Ansley, W. E. and Harame, D. L. (2003). Noise modeling and SiGe profile design tradeoffs for RF applications, *IEEE Trans. Electron Devices* **7**, 2037–2044.

Nokubo, J., Tamura, T., Nakamae, M., Shiraki, H., Ikushima, T., Akashi, T., Mayumi, H., Kubota, T. and Nakamura T. (1983). A 4.5-ns access time 1k × 4 bit ECL RAM, *IEEE J. Solid-State Circuits* **SC-18**, 515–520.

Noble, W. P., Voldman, S. H. and Bryant, A. (1989). The effects of gate field on the leakage characteristics of heavily doped junctions, *IEEE Trans. Electron Devices* **ED-36**, 720–726.

Oda, K., Ohue, E., Tanabe, M., Shimamoto, H., Onai, T. and Washio, K. (1997). 130-GHz f_T SiGe HBT Technology, *IEEE IEDM Technical Digest*, 791–794.

Ogura, S., Codella, C. F., Rovedo, N., Shepard, J. F. and Riseman, J. (1982). A half-micron MOSFET using double-implanted LDD, *IEEE IEDM Technical Digest*, 718–721.

Oh, C. S., Chang, W. H., Davari, B. and Taur, Y. (1990). Voltage dependence of the MOSFET gate-to-source/drain overlap, *Solid-State Electron.* **33**, 1650–1652.

Oh, S.-H., Monroe, D. and Hergenrother, J. M. (2000). Analytic description of short-channel effects in fully-depleted double-gate and cylindrical, surrounding-gate MOSFETs, *IEEE Electron Device Lett.* **9**, 445–447.

Ohkura, Y. (1990). Quantum effects in Si n-MOS inversion layer at high substrate concentration, *Solid-State Electron.* **33**, 1581–1585.

Oka, T., Hirata, K., Ouchi, K., Uchiyama, H., Mochizuki, K. and Nakamura, T. (1997). InGaP/GaAs HBTs with high-speed and low-current operation fabricated using WSi/Ti as the base electrode and buried SiO2 in the extrinsic collector, *IEEE IEDM Technical Digest*, 739–742.

Okada, K. (1997). Extended time dependent dielectric breakdown model based on anomalous gate area dependence of lifetime in ultra-thin dioxides, *Jpn. J. Appl. Phys.* **36**, 1143–1147.

Ouyang, Q. C., Cai, J., Ning, T., Oldiges P. and Johnson, J. B. (2002). A simulation study of thin SOI bipolar transistors with fully or partially depleted collector, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 28–31.

Paasschens, J. C. J., Kloosterman, W. J. and Havens, R. J. (2001). Modeling two SiGe HBT specific features for circuit simulation, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 38–41.

Pao, H. C. and Sah, C. T. (1966). Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors, *Solid-State Electron.* **9**, 927–937.

Padovani, F. A. and Stratton, R. (1966). Field and thermionic–field emission in Schottky barriers, *Solid-State Electron.* **9**, 695–707.

Patton, G. L., Stork, J. M. C., Comfort, J. H., Crabbé, E. F., Meyerson, B. S., Harame, D. L. and Sun, J. Y.-C. (1990). SiGe-base heterojunction bipolar transistors: physics and design issues, *IEEE IEDM Technical Digest*, 13–16.

Pavan, P., Bez, R., Olivo, P. and Zanoni, E. (1997). Flash memory cells – an overview, *IEEE Proc.* **85**, 1248–1271.

People, R. (1985). Indirect band gap of coherently strained Ge_xSi_{1-x} bulk alloys on ⟨001⟩ silicon substrate, *Phys. Rev. B* **32**, 1405–1408.

People, R. (1986). Physics and applications of Ge_xSi_{1-x}/Si strained-layer heterostructures, *IEEE J. Quantum Electron.* **QE-22**, 1696–1710.

Poon, H. C., Gummel, H. K. and Scharfetter, D. L. (1969). High injection in epitaxial transistors, *IEEE Trans. Electron Devices* **ED-16**, 455–457.

Prinz, E. J. and Sturm, J. C. (1990). Base transport in near-ideal graded-base $Si/Si_{1-x}Ge_x/Si$ heterojunction bipolar transistors from 150 K to 370 K, *IEEE IEDM Technical Digest*, 975–978.

Prinz, E. J. and Sturm, J. C. (1991). Current gain–Early voltage products in heterojunction bipolar transistors with non-uniform base bandgaps, *IEEE Electron Device Lett.* **12**, 661–663.

Prinz, E. J., Garone, P. M., Schwartz, P. V., Xiao, X. and Sturm, J. C. (1989). The effect of base-emitter spacers and strain-dependent densities of states in $Si/Si_{1-x}Ge_x/Si$ heterojunction bipolar transistors, *IEEE IEDM Technical Digest*, 639–642.

Pritchard, R. L. (1955). High-frequency power gain of junction transistors, *Proc. IRE* **43**, 1075–1085.

Pritchard, R. L. (1967). *Electrical Characteristics of Transistors*. New York: McGraw-Hill.

Pugh, E. W., Critchlow, D. L., Henle, R. A. and Russell, L. A. (1981). Solid state memory development in IBM, *IBM J. Res. Develop.* **25**, 585–602.

Rahman, A., Guo, J., Datta, S. and Lundstrom, M. S. (2003). Theory of ballistic nanotransistors, *IEEE Trans. Electron Devices* **50**, 1853–1864.

Ranade, P., Ghani, T., Kuhn, K., Mistry, K., Pae, S., Shifren, L., Stettler, M., Tone, K., Tyagi, S. and Bohr, M. (2005). High performance 35nm L_{gate} CMOS transistors featuring NiSi metal gate (FUSI), uniaxial strained silicon channels and 1.2nm gate oxide, *IEEE IEDM Technical Digest*, 217–220.

Ranfft, R. and Rein, H.-M. (1982). A simple optimization procedure for bipolar subnanosecond ICs with low power dissipation, *Microelectron. J.* **13**, 23–28.

Rao, G. S., Gregg, T. A., Price, C. A., Rao, C. L. and Repka, S. J. (1997). IBM S/390 Parallel Enterprise Servers G3 and G4, *IBM J. Res. Develop.* **41**, 397–403.

Rauch III, S. E. (2002). The statistics of NBTI-induced V_T and β mismatch shifts in pMOSFETs, *IEEE Trans. Device and Materials Reliability* **2**, 89–93.

Razavi, B., Yan, R.-H. and Lee, K. F. (1994). Impact of distributed gate resistance on the performance of MOS devices, *IEEE Trans. Circuits & Systems* **41**, 750.

Razouk, R. R. and Deal, B. E. (1979). Dependence of interface state density on silicon thermal oxidation process variables, *J. Electrochem. Soc.* **126**, 1573–1581.

Riccó, B., Stork, J. M. C. and Arienzo, M. (1984). Characterization of non-ohmic behavior of emitter contacts of bipolar transistors, *IEEE Electron Device Lett.* **EDL-5**, 221–223.

Rideout, V. L., Gaensslen, F. H. and LeBlanc, A. (1975). Device design consideration for ion-implanted n-channel MOSFETs, *IBM J. Res. and Devel.* **19**, 50–59.

Rieh, J.-S., Cai, J., Ning, T., Stricker, A. and Freeman, G. (2005). Reverse active mode current characteristics of SiGe HBTs, *IEEE Trans. Electron Devices* **52**, 1291–1222.

Rios, R. and Arora, N. D. (1994). Determination of ultra-thin gate oxide thickness for CMOS structures using quantum effects, *IEEE IEDM Technical Digest*, 613–616.

Roulston, D. J. (1990). *Bipolar Semiconductor Devices*. New York: McGraw-Hill.

Rücker, H., Heinemann, B., Bolze, D., Knoll, D., Krüger, D., Kurps, R., Osten, H. H., Schley, P., Tillack, B. and Zaumseil, P. (1999). Dopant diffusion in C-doped Si and SiGe: physical model and experimental verification, *IEEE IEDM Technical Digest*, 345–348.

Sabnis, A. G. and Clemens, J. T. (1979). Characterization of the electron mobility in the inverted ⟨100⟩ Si surface, *IEEE IEDM Technical Digest*, 18–21.

Sah, C. T. (1966). The spatial variation of the quasi-Fermi Potential in p–n junctions, *IEEE Trans. Electron Devices* **ED-13**, 839–846.

Sah, C. T. (1988). Evolution of the MOS transistor – from concept to VLSI, *Proc. IEEE* **76**, 1280–1326.

Sah, C. T. (1991). *Fundamentals of Solid-State Electronics*. Singapore: World Scientific.

Sah, C. T., Noyce, R. N. and Shockley, W. (1957). Carrier generation and recombination in p–n junction and p–n junction characteristics, *Proc. IRE* **45**, 1228–1243.

Sah, C. T., Ning, T. H. and Tschopp, L. L. (1972). The scattering of electrons by surface oxide charges and by lattice vibrations at the silicon–silicon dioxide interface, *Surface Sci.* **32**, 561–575.

Sai-Halasz, G. A. (1995). Performance trends in high-end processors, *IEEE Proc.* **83**, 20–36.

Sai-Halasz, G. A., Wordeman, M. R., Kern, D. P., Rishton, S. and Ganin, E. (1988). High transconductance and velocity overshoot in NMOS devices at the 0.1 μm gate-length level, *IEEE Electron Device Lett.* **EDL-9**, 464–466.

Sai-Halasz, G. A., Wordeman, M. R., Kern, D. P., Rishton, S. A., Ganin, E., Chang, T. H. P. and Dennard, R. H. (1990). Experimental technology and performance of 0.1 μm gate length FETs operated at liquid-nitrogen temperature, *IBM J. Res. and Devel.* **34**, 452–465.

Sakurai, T. (1983). Approximation of wiring delay in MOSFET LSI, *IEEE J. Solid-State Circuits* **SC-18**, 418.

Salmon, S. L., Cressler, J. D., Jaeger, R. C. and Harame, D. L. (1997). The impact of Ge profile shape on the operation of SiGe HBT precision voltage references, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 100–103.

Samachisa, G., Su, C.-S., Kao, Y.-S., Smarandoiu, G., Wang, C.-Y. M., Wong, T. and Hu, C. (1987). A 128K flash EEPROM using double-polysilicon technology, *IEEE J. Solid-State Circuits* **SC-22**, 676–683.

Schaper, L. W. and Amey, D. I. (1983). Improved electrical performance required for future MOS packaging, *IEEE Trans. Components, Hybrids, and Manufacturing Tech.* **CHMT-6**, 283.

Schroder, D. K. (1990). *Semiconductor Material and Device Characterization*. New York: Wiley.

Schroder, D. K. and Babcock, J. A. (2003). Negative bias temperature instability: road to cross in deep submicron silicon semiconductor manufacturing, *J. Appl. Phys.* **96**, 1–18.

Schuegraf, K. F. and Hu, C. (1994). Metal–oxide–semiconductor field-effect-transistor substrate current during Fowler–Nordheim tunneling stress and silicon dioxide reliability, *J. Appl. Phys.* **76**, 3695–3700.

Schuegraf, K. F., King, C. C. and Hu, C. (1992). Ultra-thin dioxide leakage current and scaling limit, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 18–19.

Schüppen, A. and Dietrich, H. (1995). High speed SiGe heterobipolar transistors, *J. Crystal Growth* **157**, 207–214.

Schüppen, A., Dietrich, H., Gerlach, S., Höhnemann, H., Arndt, J., Seiler, U., Götzfried, R., Erben, U. and Schumacher, H. (1996). SiGe – technology and components for mobile communication systems, *Proc. Bipolar/BiCMOS Circuit & Technology Meeting*, IEEE, pp. 147–150.

Seevinck, E., List, F. J. and Lohstroh, J. (1987). Static-noise margin analysis of MOS SRAM cells, *IEEE J. Solid-State Circuits* **SC-22**, 748–754.

Selmi, L., Sangiorgi, E., Bez, R. and Riccò, B. (1993). Measurement of the hot hole injection probability from Si into SiO_2 in p-MOSFET, *IEEE IEDM Technical Digest*, 333–336.

Shahidi, G. G., Antoniadis, D. A. and Smith, H. I. (1989). Indium channel implants for improved MOSFET behavior at the 100 nm channel length regime, DRC Abstract, *IEEE Trans. Electron Devices* **ED-36**, 2605.

Shatzkes, M., Av-Ron, M. and Anderson, R. M. (1974). On the nature of conduction and switching in SiO_2, *J. Appl. Phys.* **45**, 2065–2077.

Sherony, M. J., Wei, A. and Antoniadis, D. A. (1996). Effect of body charge on fully- and partially-depleted SOI MOSFET design, *IEEE IEDM Technical Digest*, 125–128.

Sheu, B. J. and Ko, P. K. (1984). A capacitive method to determine channel lengths for conventional and LDD MOSFETs, *IEEE Electron Device Lett.* **EDL-5**, 491–493.

Shiba, T., Tamaki, Y., Onai, T., Saitoh, M., Kure, T., Murai, F. and Nakamura, T. (1991). SPOYEC – a sub-10-μm^2 bipolar transistor structure using fully self-aligned sidewall polycide base technology, *IEEE IEDM Technical Digest*, 455–458.

Shiba, T., Uchino, T., Ohnishi, K. and Tamaki, Y. (1996). In-situ phosphorus-doped polysilicon emitter technology for very high-speed, small emitter bipolar transistors, *IEEE Trans. Electron Devices* **43**, 889–897.

Shibib, M. A., Lindholm, F. A. and Therez, R. (1979). Heavily doped transparent-emitter regions in junction solar cells, diodes, and transistors, *IEEE Trans. Electron Devices* **ED-26**, 959–965.

Shockley, W. (1949). The theory of p–n junctions in semiconductors and p–n junction transistors, *Bell Syst. Tech. J.* **28**, 435–489.

Shockley, W. (1950). *Electrons and Holes in Semiconductors*. Princeton: D. Van Nostrand.

Shockley, W. (1961). Problems related to p-n junctions in silicon, *Solid-State Electron.* **2**, 35–67.

Shockley, W. and Read, W. T. (1952). Statistics of the recombination of holes and electrons, *Phys. Rev.* **87**, 835–842.

Shrivastava, R. and Fitzpatrick, K. (1982). A simple model for the overlap capacitance of a VLSI MOS device, *IEEE Trans. Electron Devices* **ED-29**, 1870–1875.

Sleva, S. and Taur, Y. (2005). The influence of source and drain junction depth on the short-channel effect in MOSFETs, *IEEE Trans. Electron Devices* **ED-52**, 2814–2816.

Slotboom, J. W. and de Graaff, H. D. (1976). Measurements of bandgap narrowing in Si bipolar transistors, *Solid-State Electron.* **19**, 857–862.

Sodini, C. G., Ekstedt, T. W. and Moll, J. L. (1982). Charge accumulation and mobility in thin dielectric MOS transistors, *Solid-State Electron.* **25**, 833–841.

Sodini, C. G., Ko, P. K. and Moll, J. L. (1984). The effect of high fields on MOS device and circuit performance, *IEEE Trans. Electron Devices* **ED-31**, 1386–1393.

Solomon, P. M. (1982). A comparison of semiconductor devices for high speed logic, *IEEE Proc.* **70**, 489.

Solomon, P. M. and Tang, D. D. (1979). Bipolar circuit scaling, *IEEE ISSCC Digest of Technical Papers*, 86–87.

Solomon, P. M., Jopling, J., Frank, D. J., D'Emic, C., Dokumaci, O., Ronsheim, P. and Haensch, W. E. (2004). Universal tunneling behavior in technologically relevant p–n junction diodes, *J. Appl. Phys.* **95**, 5800–5812.

Stathis, J. H. (1999). Percolation models for gate oxide breakdown, *J. Appl. Phys.* **86**, 5757–5766.

Stathis, J. H. and DiMaria, D. J. (1998). Reliability projection for ultrathin oxides at low voltage, *IEEE IEDM Technical Digest*, 167–170.

Stern, F. (1972). Self-consistent results for n-type Si inversion layers, *Phys. Rev. B* **5**, 4891–4899.

Stern, F. (1974). Quantum properties of surface space-charge layers, *CRC Crit. Rev. Solid-State Sci.* **4**, 499.

Stern, F. and Howard, W. E. (1967). Properties of semiconductor surface inversion layers in the electric quantum limit, *Phys. Rev.* **163**, 816–835.

Stolk, P. A., Eaglesham, D. J., Gossmann, H.-J. and Poate, J. M. (1995). Carbon incorporation in silicon for suppressing interstitial-enhanced boron diffusion, *Appl. Phys. Lett.* **66**, 1370–1372.

Stork, J. M. C. and Isaac, R. D. (1983). Tunneling in base–emitter junctions, *IEEE Trans. Electron Devices* **ED-30**, 1527–1534.

Su, L. T., Jacobs, J. B., Chung, J. and Antoniadis, D. A. (1994). Deep-submicrometer channel design in silicon-on-insulator (SOI) MOSFETs, *IEEE Electron Device Lett.* **15**, 183–185.

Suehle, J. S. (2002). Ultrathin gate oxide reliability: physical models, statistics, and characterization, *IEEE Trans. Electron Devices* **49**, 958–971.

Sun, J. Y.-C., Wordeman, M. R. and Laux, S. E. (1986). On the accuracy of channel length characterization of LDD MOSFETs, *IEEE Trans. Electron Devices* **ED-33**, 1556–1562.

Sun, J. Y.-C., Taur, Y., Dennard, R. H. and Klepner, S. P. (1987). Submicrometer-channel CMOS for low-temperature operation, *IEEE Trans. Electron Devices* **ED-34**, 19–27.

Sun, J. Y.-C., Wong, C. Y., Taur, Y. and Hsu, C. (1989). Study of boron penetration through thin oxide with p^+ polysilicon gate, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 17–18.

Sun, S. C. and Plummer, J. D. (1980). Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces, *IEEE Trans. Electron Devices* **ED-27**, 1497–1508.

Suñé, J. and Wu, E. Y. (2002). Statistics of successful breakdown events for ultra-thin gate oxides, *IEEE IEDM Technical Digest*, 147–150.

Suñé, J., Wu, E. Y. and Lai, W. L. (2004). Successive oxide breakdown statistics; correlation effects, reliability methodologies, and their limits, *IEEE Trans. Electron Devices* **51**, 1584–1592.

Suzuki, K. (1991). Optimum base doping profile for minimum base transit time, IEEE Trans. *Electron Devices* **38**, 2128–2133.

Suzuki, E., Miura, K., Hayasi, Y., Tsay, R.-P. and Schroder, D. K. (1989). Hole and electron current transport in metal–oxide–nitride–oxide–silicon memory structures, *IEEE Trans. Electron Devices* **36**, 1145–1149.

Swanson, R. M. and Meindl, J. D. (1972). Ion-implanted complementary MOS transistors in low-voltage circuits, *IEEE J. Solid-State Circuits* **SC-7**, 146–153.

Swirhun, S. E., Kwark, Y.-H. and Swanson, R. M. (1986). Measurement of electron lifetime, electron mobility and band-gap narrowing in heavily doped p-type silicon, *IEEE IEDM Technical Digest*, 24–27.

Sze, S. M. (1981). *Physics of Semiconductor Devices*. New York: Wiley.

Sze, S. M., Crowell, C. R. and Kahng, D. (1964). Photoelectric determination of the image force dielectric constant for hot electrons in Schottky barriers, *J. Appl. Phys.* **35**, 2534–2536.

Takagi, S., Iwase, M. and Toriumi, A. (1988). On universality of inversion-layer mobility in n- and p-channel MOSFETs, *IEEE IEDM Technical Digest*, 398–401.

Takeda, E., Suzuki, N. and Hagiwara, T. (1983). Device performance degradation due to hot-carrier injection at energies below the Si-SiO_2 energy barrier, *IEEE IEDM Technical Digest*, 396–399.

Takeda, E., Yang, C. Y. and Miura-Hamada, A. (1995). *Hot-Carrier Effects in MOS Devices*. San Diego: Academic Press.

Tang, D. D. (1980). Heavy doping effects in pnp bipolar transistors, *IEEE Trans. Electron Devices* **ED-27**, 563–570.

Tang, D. D. and Lu, P.-F. (1989). A reduced-field design concept for high-performance bipolar transistors, *IEEE Electron Device Lett.* **10**, 67–69.

Tang, D. D. and Solomon, P. M. (1979). Bipolar transistor design for optimized power-delay logic circuits, *IEEE J. Solid-State Circuits* **SC-14**, 679–684.

Tang, D. D., MacWilliams, K. P. and Solomon, P. M. (1983). Effects of collector epitaxial layer on the switching speed of high-performance bipolar transistors, *IEEE Electron Device Lett.* **EDL-4**, 17–19.

Taur, Y. (2000). An analytical solution to a double-gate MOSFET with undoped body, *IEEE Electron Device Lett.* **21**, 245–247.

Taur, Y., Chang, W. H. and Dennard, R. H. (1984). Characterization and modeling of a latchup-free 1-μm CMOS technology, *IEEE IEDM Technical Digest*, 398–401.

Taur, Y., Hu, G. J., Dennard, R. H., Terman, L. M., Ting, C. Y. and Petrillo, K. E. (1985). A self-aligned 1 μm channel CMOS technology with retrograde n-well and thin epitaxy, *IEEE Trans. Electron Devices* **ED-32**, 203–209.

Taur, Y., Sun, J. Y.-C., Moy, D., Wang, L. K., Davari, B., Klepner, S. P. and Ting, C. Y. (1987). Source-drain contact resistance in CMOS with self-aligned $TiSi_2$, *IEEE Trans. Electron Devices* **ED-34**, 575–580.

Taur, Y., Zicherman, D. S., Lombardi, D. R., Restle, P. J., Hsu, C. H., Hanafi, H. I., Wordeman, M. R., Davari, B. and Shahidi, G. G. (1992). A new shift and ratio method for MOSFET channel length extraction, *IEEE Electron Device Lett.* **EDL-13**, 267–269.

Taur, Y., Hsu, C. H., Wu, B., Kiehl, R., Davari, B. and Shahidi, G. (1993a). Saturation transconductance of deep-submicron-channel MOSFETs, *Solid-State Electron.* **36**, 1085–1087.

Taur, Y., Cohen, S., Wind, S., Lii, T., Hsu, C., Quinlan, D., Chang, C., Buchanan, D., Agnello, P., Mii, Y., Reeves, C., Acovic, A. and Kesan, V. (1993b). Experimental 0.1-μm p-channel MOSFET with p^+ polysilicon gate on 35-Å gate oxide, *IEEE Electron Device Lett.* **EDL-14**, 304–306.

Taur, Y., Wind, S., Mii, Y., Lii, Y., Moy, D., Jenkins, K., Chen, C. L., Coane, P. J., Klaus, D., Bucchignano, J., Rosenfield, M., Thomson, M. and Polcari, M. (1993c). High performance 0.1 μm CMOS devices with 1.5 V power supply, *IEEE IEDM Technical Digest*, 127–130.

Taur, Y., Mii, Y.-J., Frank, D. J., Wong, H.-S., Buchanan, D. A., Wind, S. J., Rishton, S. A., Sai-Halasz, G. A. and Nowak, E. J. (1995a). CMOS scaling into the 21th century: 0.1 μm and beyond, *IBM J. Res. and Develop.* **39**, 245–260.

Taur, Y., Mii, Y.-J., Logan, R. and Wong, H. S. (1995b). On effective channel length in 0.1 μm MOSFETs, *IEEE Electron Device Lett.* **EDL-16**, 136–138.

Taur, Y., Buchanan, D. A., Chen, W., Frank, D. J., Ismail, K. E., Lo, S.-H., Sai-Halasz, G. A., Viswanathan, R. G., Wann, H.-J. C., Wind, S. J. and Wong, H.-S. (1997). CMOS scaling into the nanometer regime, *IEEE Proc.* **85**, 486–504.

Taur, Y., Wann, C. H. and Frank, D. J. (1998). 25 nm CMOS design considerations, *IEEE IEDM Technical Digest*, 789–792.

Taur, Y., Liang, X., Wang, W. and Lu, H. (2004). A continuous, analytic drain-current model for double-gate MOSFETs, *IEEE Electron Device Lett.* **25**, 107–109.

Taylor, G. W. (1984). Velocity saturated characteristics of short-channel MOSFETs, *AT&T Bell Labs. Tech. J.* **63**, 1325.

Taylor, G. W. and Simmons, J. G. (1986). Figure of merit for integrated bipolar transistors, *Solid-State Electron.* **29**, 941–946.

Terman, L. M. (1971). MOSFET memory circuits, *Proc. IEEE* **59**, 1044–1058.

Thornton, R. D., DeWitt, D., Gray, P. E. and Chenette, E. R. (1966). *Characteristics and Limitations of Transistors*, vol. 4, Semiconductor Electronics Education Committee. New York: Wiley.

Ting, C. Y., Iyer, S. S., Osburn, C. M., Hu, G. J. and Schweighart, A. M. (1982). The use of $TiSi_2$ in a self-aligned silicide technology, presented at ECS Symposium on VLSI Science and Technology.

Tiwari, S., Rana, F., Hanafi, H., Hartstein, A., Crabbé, E.F. and Chan, K. (1996). A silicon nanocrystals based memory, *Appl. Phys. Lett.* **68**, 1377–1379.

Troutman, R.R. (1979). VLSI limitations from drain-induced barrier lowering, *IEEE Trans. Electron Devices* **ED-26**, 461–469.

Trumbore, F.A. (1960). Solid solubilities of impurity elements in germanium and silicon, *Bell System Tech. J.* **39**, 205.

Tuinhout, H., Pelgrom, M., de Vries, R.P. and Vertregt, M. (1996). Effects of metal coverage on MOSFET matching, *IEEE IEDM Technical Digest*, 735–738.

Tuinhout, H.P., Montree, A.H., Schmitz, J. and Stolk, P.A. (1997). Effects of gate depletion and boron penetration on matching of deep submicron CMOS transistors, *IEEE IEDM Technical Digest*, 631–634.

Tung, R.T. (1992). Electron transport at metal–semiconductor interfaces: general theory, *Phys. Rev. B* **45**, 13 509–13 523.

Uchida, K., Watanabe, H., Kinoshita, A., Koga, J., Numata, T. and Takagi, S.-I. (2002). Experimental study on carrier transport mechanism in ultrathin body SOI n and p MOSFETs with SOI thickness less than 5 nm, *IEEE IEDM Technical Digest*, 47–50.

Uchino, T., Shiba, T., Kikuchi, T., Tamaki, Y., Watanabe, A. and Kiyota, Y. (1995). Very-high-speed silicon bipolar transistors with *in-situ* doped polysilicon emitter and rapid vapor-phase doping base, *IEEE Trans. Electron Devices* **42**, 406–412.

Van de Walle C.G. and Martin, R.M. (1986). Theoretical calculations of heterojunction discontinuities in the Si/Ge system, *Phys. Rev. B* **34**, 5621–5634.

van Dort, M.J., Woerlee, P.H. and Walker, A.J. (1994). A simple model for quantization effects in heavily-doped silicon MOSFETs at inversion conditions, *Solid-State Electron.* **37**, 411–414.

van Overstraeten, R. and de Man, H. (1970). Measurement of the ionization rates in diffused silicon p–n junctions, *Solid-State Electron.* **13**, 583–608.

van Overstraeten, R.J., de Man, H.J. and Mertens, R.P. (1973). Transport equations in heavy doped silicon, *IEEE Trans. Electron Devices* **ED-20**, 290–298.

Verwey, J.F. (1972). Hole currents in thermally grown SiO_2, *J. Appl. Phys.* **43**, 2273–2277.

Wang, S.T. (1979). On the I–V characteristics of floating-gate MOS transistors, *IEEE Trans. Electron Devices* **26**, 1292–1294.

Wanlass, F. and Sah, C.T. (1963). Nanowatt logic using field-effect metal-oxide-semiconductor triodes, *IEEE ISSCC Digest of Technical Papers*, 32–33.

Warnock, J.D. (1995). Silicon bipolar device structures for digital applications: technology trends and future directions, *IEEE Trans. Electron Devices* **42**, 377–389.

Washio, K., Ohue, E., Shimamoto, H., Oda, K., Hayami, R., Kiyota, Y., Tanabe, M., Kondo, K., Hashimoto, T. and Harada, T. (2000). A 0.2-μm 180GHz f_{max} 6.7-ps-ECL SOI/HRS self-aligned SEG SiGe HBT/CMOS technology for microwave and high-speed digital applications, *IEEE IEDM Technical Digest*, 741–744.

Washio, K., Ohue, E., Shimamoto, H., Oda, K., Hayami, R., Kiyota, Y., Tanabe, M., Kondo, M., Hashimoto, T. and Harada, T. (2002). A 0.2-μm 180-GHz-f_{max} 6.7-ps-ECL SOI-HRS self-aligned SEG SiGe HBT/CMOS technology for microwave and high-speed digital applications, *IEEE Trans. Electron Devices* **49**, 271–278.

Webster, W.M. (1954). On the variation of junction-transistor current amplification factor with emitter current, *Proc. IRE* **42**, 914–920.

Weng, J., Holz, J. and Meister, T.F. (1992). New method to determine the base resistance of bipolar transistors, *IEEE Electron Device Lett.* **13**, 158–160.

Werner, W.M. (1976). The influence of fixed interface charges on the current-gain falloff of planar npn transistors, *J. Electrochem. Soc.* **123**, 540–543.

Wiedmann, S.K. and Berger, H.H. (1971). Small-size low-power bipolar memory cell, *IEEE J. Solid-State Circuits* **SC-6**, 283–288.

Wong, C.Y., Sun, J.Y.-C., Taur, Y., Oh, C.S., Angelucci, R. and Davari, B. (1988). Doping of n^+ and p^+ polysilicon in a dual-gate CMOS process, *IEEE IEDM Technical Digest*, 238–241.

Wong, C.Y., Piccirillo, J., Bhattacharyya, A., Taur, Y. and Hanafi, H.I. (1989). Sidewall oxidation of polycrystalline-silicon gate, *IEEE Electron Device Lett.* **EDL-10**, 420–422.

Wong, H.-S., and Taur, Y. (1993). Three-dimensional atomistic simulation of discrete random dopant distribution effects in sub-0.1 μm MOSFETs, *IEEE IEDM Technical Digest*, 705–708.

Wong, H.-S., Chan, K.K., Lee, Y., Roper, P. and Taur, Y. (1997). Fabrication of ultrathin, highly uniform thin-film SOI MOSFETs with low series resistance using pattern-constrained epitaxy, *IEEE Trans. Electron Devices* **44**, 1131–1135.

Wong, H.-S., Frank, D.J., Solomon, P.M., Wann, H.J. and Welser, J.J. (1999). Nanoscale CMOS, *IEEE Proc.* **87**, 537–570.

Wordeman, M.R. (1986). Design and modeling of miniaturized MOSFETs. Unpublished Ph. D. thesis, Columbia University.

Wordeman, M.R. (1989). Private communications.

Wordeman, M.R., Sun, Y.C. and Laux, S.E. (1985). Geometry effects in MOSFET channel length extraction algorithms, *IEEE Electron Device Lett.* **EDL-6**, 186–188.

Wu, A.T., Chan, T.Y., Ko, P.K. and Hu, C. (1986). A novel higb-speed 5-volt programming EPROM structure with source-side injection, *IEEE IEDM Technical Digest*, 584–587.

Yamada, S., Yamane, T., Amemiya, K. and Naruke, K. (1996). A self-convergence erase for NOR flash EEPROM using avalanche hot carrier injection, *IEEE Trans. Electron Devices* **43**, 1937–1941.

Yan, R.H., Ourmazd, A., Lee, K.F., Jeon, D.Y., Rafferty, C.S. and Pinto, M.R. (1991). Scaling the Si metal-oxide-semiconductor field-effect transistor into the 0.1-μm regime using vertical doping engineering, *Appl. Phys. Lett.* **59**, 3315–3317.

Yoshimi, M., Hazama, H., Takahashi, M., Kambayashi, S., Wada, T., Kato, K. and Tango, H. (1989). Two-dimensional simulation and measurement of high-performance MOSFETs made on a very thin SOI film, *IEEE Trans. Electron Devices* **ED-36**, 493–503.

Yoshino, C., Inou, K., Matsuda, S., Nakajima, H., Tsuboi, Y., Naruse, H., Sugaya, H., Katsumata, Y. and Iwai, H. (1995). A 62.8-GHz f_{max} LP-CVD epitaxially grown silicon-base bipolar transistor with extremely high Early voltage of 85.7 V, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 131–132.

Yu, A.Y.C. (1970). Electron tunneling and contact resistance of metal-silicon contact barriers, *Solid-State Electron.* **13**, 239–247.

Yu, B., Song, J., Yuan, Y., Lu, W.-Y. and Taur, Y. (2008). A unified analytic drain current model for multiple-gate MOSFETs, *IEEE Trans. Electron Devices* **ED-55**, 2157–2163.

Yu, D., Lee, K., Kim, B., Ontiveros, D., Vargason, K., Kuo, J.M. and Kao, Y.C. (2003). Ultra high-speed InP-InGaAs SHBTs with f_{max} of 478 GHz, *IEEE Electron Device Lett.* **24**, 384–386.

Yu, H.N. (1971). Transistor with limited-area base-collector junction, U.S. Patent Re. 27,045, reissued February 2.

Yu, Z., Riccó, B. and Dutton, R.W. (1984). A comprehensive analytical and numerical model of polysilicon emitter contacts in bipolar transistors, *IEEE Trans. Electron Devices* **ED-31**, 773–785.

Zafar, S., Lee, B.H., Stathis, J. and Ning, T. (2004). A model for negative bias temperature instability (NBTI) in oxide and high κ pFETs, *Symp. VLSI Technology Digest of Tech. Papers*, IEEE, pp. 208–209.

Index

abrupt junction, 38
acceptor, 17
acceptor level, 18–9
access transistor, 477–8, 496
accumulation, 76–7
accumulation layer, 250
  charge density, 250
  resistance, 274–5
  sheet resistivity, 251
ac equivalent circuit, 356
ac model, 355
active power, 220–1, 264–5, 299–301
activity factor of circuits, 265, 301
admittance matrix, 598
  bipolar, 617
  extrinsic, 601, 618
  intrinsic, 309, 601
  MOSFET, 309
alignment tolerance, 271
analog bipolar devices, 463
analytic potential drain-current model:
  DG MOSFET, 529, 532
  surrounding-gate MOSFET, 536
anode, 130
antifuse, 500
apparent bandgap narrowing, 325
applied voltage, 32, 35, 40
arsenic, 18
Auger recombination, 34
avalanche breakdown, 122
avalanche hot electron injection, 503
avalanche hot hole injection, 503
avalanche multiplication, 123

ballistic MOSFET, 192
  current model, 588
  saturation current, 592
  saturation voltage, 592
ballistic transport, 192–4, 588–93
band bending:
  in MOS device, 77, 79
  in polysilicon gate, 92
band diagram:
  bipolar transistor, 319
  buried-channel MOSFET, 222
  MOS capacitor, 73–5, 77, 79, 95–7
  MOSFET, 179, 193, 589, 595
  n-type silicon, 18
  p–n junction, 35
  p-type silicon, 18
  SiGe-base bipolar transistor, 391
  Si–SiGe n–p diode, 614
  Si-SiO_2, 73
bandgap, 12
  temperature coefficient, 12
band-to-band tunneling, 108, 125
barrier height, 109
barrier lowering, 114, 569
base:
  bipolar transistor, 318, 377
  extrinsic, 350, 377
  intrinsic, 350, 377
  p–n diode, 55
base charge, 341
base–collector depletion layer delay time (*see* base–collector transit time)
base-collector diode, 322, 341, 385
base–collector junction avalanche, 343, 367
base-collector transit time, 361, 452
base-conductivity modulation, 344
base current, 320, 330
base delay time, 361
base design, 377
base-emitter depletion layer delay time (*see* base–emitter transit time)
base-emitter transit time, 361, 451
base Gummel number, 330, 379
base junction depth, 320, 375
base resistance:
  extrinsic, 338, 605
  intrinsic, 338, 605, 610
base series resistance (*see* base resistance)
base sheet resistivity, 379
base transit time, 65, 384, 452
base transport factor, 368
base widening:
  at collector end, 345
  at emitter end, 419
base width, 320, 375
base width modulation:
  by V_{BE}, 419
  by V_{BC}, 340
basic equations for device operation, 27–35
BESOI, 517
bias point trajectory in switching, 292
BiCMOS, 523
bipolar digital circuit:
  current-switch emitter-follower circuit, 442
  delay components, 442
  device structure, 429, 445
  differential-current-switch circuit (differential ECL), 442
  emitter-coupled logic (ECL), 441
  layout, 445
bipolar inverter, 489
bipolar latch, 488
bipolar SRAM cell, 487
bipolar transistor:
  as an amplifier, 362, 463
  as an on-off switch, 489
  breakdown voltages, 366
  device model, 352
  device optimization, 447, 463
  figures of merit, 437
  forward-mode *I–V*, 336
  in saturation, 336, 491
  multi-emitter, 488
  n–p–n, 318, 429
  optimization for digital circuits, 447
  optimization for RF and analog applications, 463
  p–n–p, 318
  reverse-mode *I–V*, 423
  saturation currents, 369
  scaling, 457, 463
  scaling limits, 460, 468, 471
  SiGe base, 389, 431
bird's beak, 279
bistable latch, 478, 486
bitline, 476
  complementary, 478, 497
  folded, 499
body effect, 157–8, 166–7, 307
  coefficient, 157–8, 581
Boltzmann's constant, 14
Boltzmann's relations, 30
Boltzmann transport equation, 192
bonding (of silicon wafer), 517
boosted wordline, 497
boron, 18, 222
BOX (buried oxide of SOI), 520
boxlike doping profile, 380
breakdown voltage:
  bipolar transistor, 366
  MOSFET, 200
  p–n junction, 122
BSIM compact model, 290
buffer stage, 297, 316
built-in electric field, 324
built-in potential, 37
bulk mobility, 23–4
buried-channel MOSFET, 222
burn in, 214
butterfly plot, 478

capacitance:
  depletion-layer, 41, 87
  diffusion, 70, 359, 562
  interconnect (wire), 283–5
  interface-trap, 104
  intrinsic MOSFET, 172–3
  inversion layer, 87, 89, 173, 175, 593
  junction, 41, 278
  Miller, 302–3
  MOS, 85–90
  overlap, 279
capacitance-voltage characteristics:
  high-frequency, 88–9
  low-frequency (quasi-static), 88
  MOS, 86–90
  p–n junction, 42
capture and emission at a trap center, 553
capture cross section, 102
  field dependence, 102
  for electrons, 101
  for holes, 102
  temperature dependence, 102
capture rate:
  electron, 553
  hole, 554
carrier concentration:
  extrinsic, 20–1
  intrinsic, 16
carrier confinement, 82, 234, 594
carrier transport, 23–7, 192–4
carrier velocity
  at source, 193–4
  at drain, 192
cathode, 130
channel doping:
  counter-doping, 222, 231
  extreme retrograde, 230
  graphical interpretation, 232
  halo (pocket), 233
  high-low, 225
  laterally nonuniform, 233
  low-high, 229
  nonuniform, 224–34
  nonuniform, band diagram, 233
  profile design, 224
  pulse-shaped (delta), 230
  retrograde, 229
  superhalo, 234
channel length, 149
  definition, 242
  effective, 243
  extraction, 244–7
  extraction by *C–V*, 252

channel length (cont.)
metallurgical, 243
minimum, 183, 219
channel length bias, 244
channel length modulation, 159, 195
channel profile design, 217–34
channel resistance, 196, 244
channel sheet resistivity, 196
short-channel device, 248
channel width, 150
charge-control analysis, 67, 361
charge erasure (in NVRAM):
by Fowler-Nordheim tunneling, 507
by high-energy photons, 507
charge injection (in NVRAM):
by avalanche hot electrons, 503
by channel hot electrons, 502
by tunneling, 505
source side, 512
charge in silicon dioxide:
fixed charge, 99
interface trapped charge, 99
mobile ionic charge, 99
oxide trapped charge, 99
charge neutrality, 20, 79
charge-sheet model, 153
charge storage (in NVRAM):
in floating gate, 505
in nanocrystals, 515
in silicon nitride, 505
charge to breakdown, 139
chemical-mechanical polish (CMP), 517, 538
circuit:
bipolar, 352, 441
digital, 441
static CMOS, 256–73
clock frequency, 220, 265
CMOS:
advanced, 307–15
circuit, 256
inverter, 256–66
low power, 221, 300
NAND, 266–70, 304
NOR, 266
performance-power tradeoff, 221, 299–301
process flow, 538–41
scaling, 204–12
scaling limit, 219
technology generations, 208, 220
technology trends, 220, 223
two-way NAND, 266, 304
0.1 μm parameters for circuit model, 291
CMOS delay sensitivity:
to body effect, 307
to channel length, 298
to device width, 296–7
to gate oxide thickness, 298
to junction capacitance, 303–4
to load capacitance, 297
to mobility, 311–2
to n- and p-device width ratio, 296
to overlap capacitance, 302
to power supply voltage, 299–301
to saturation velocity, 311–12
to source-drain series resistance, 301, 311–12
to temperature, 314
to threshold voltage, 299–301
CMOS inverter, 256
bias point trajectory, 292
cascade, 290
cross coupled, 478
delay equation, 294
folded layout, 272
input capacitance, 294, 304
intrinsic delay, 294
layout, 272
maximum voltage gain, 260
noise margin, 259
noise margin, graphical interpretation, 262
noise margin, regenerative, 260
output capacitance, 294, 304
power-delay tradeoff, 300
pull down delay, 264, 293
pull up delay, 264, 293
switching resistance, 294–5
transfer curve, 257–9
waveform, 291
CMOS NAND, 266
bottom switching, 268, 305
delay sensitivity to body effect, 307
layout, 273
noise margin, 270
propagation delay, 306
switching resistance, 307
top switching, 268, 305
two-input, 268, 304
waveform, 305
collector, 318, 385
collector current, 320, 329, 352
collector current falloff, 343
collector design, 385
collector resistance, 338
common-base current gain, 334, 367
common-emitter current gain, 334
compensated semiconductor, 35
complementary metal-oxide-semiconductor (*see* CMOS)
complementary MOSFET (*see* CMOS)
concentration gradient, 27, 544
conduction band, 11
constant-field scaling, 204–7
constant-source diffusion, 281
constant-voltage scaling, 209, 457
contact hole, 270
contact resistance, 276
contact resistivity:
between silicide and metal, 277
between silicon and silicide, 121, 276
continuity equation, 33
control gate, 506
Coulomb scattering, 23, 170
screening, 171
covalent bond, 11, 17
critical field for velocity saturation, 187
cross-coupled latch, 478
cross-over current, 265, 292, 300
current at MOSFET threshold, 213
current crowding, 276
current density equation, 31
for electrons in p-type base, 328
for holes in n-type emitter, 329
current gain, 322, 334
common-base, 334
common-emitter, 334
fall off at high currents, 338, 343
fall off at low currents, 338, 347
MOSFET, 310
current injection from surface into bulk, 250, 275
current-voltage characteristics (*see* I–V characteristics)
cutoff frequency (f_T), 437, 601, 617

dc equivalent circuit, 352
dc model, 352
Debye length, 30
deep emitter, 333
deep impurity level, 34
deep-trench isolation, 429
defect generation, 129
degeneracy, 13, 237, 594
degenerately doped silicon, 22
delay components:
bipolar circuits, 442
CMOS circuits, 263, 293
delay equation:
bipolar, 442
CMOS, 293
delay sensitivity (*see* CMOS delay sensitivity)
delay time:
emitter, 361, 451
base–collector depletion-layer, 361, 452
base–emitter depletion-layer, 361, 452
base, 361, 452
density of states, 12–4
effective, 16
2-D, 594
depletion:
in MOS, 78
in p–n junction, 38
depletion approximation, 38, 81
depletion charge, 82, 84
depletion layer, 38
depletion-layer capacitance:
MOS device, 87
p–n junction, 41
depletion-layer width:
drain junction, 205, 576
gate-controlled, 82
maximum, 82
p–n junction, 40
source junction, 576
depletion of polysilicon gate, 92–4
design points (for bipolar transistor), 447
device design:
bipolar, 374
CMOS, 204, 217
device reliability, 137, 198
device scaling:
bipolar, 457, 463
CMOS, 204–12
DIBL, 177
dielectric boundary condition, 28
dielectric breakdown: 137
breakdown event, 137
breakdown field, 138
catastrophic breakdown, 137
charge to breakdown, 139
degradation rate, 140
progressive breakdown, 140
soft breakdown, 138
successive breakdown, 140
time to breakdown, 139
dielectric constant, 13
dielectric relaxation time, 34
diffused emitter, 320, 375
diffusion capacitance, 70, 359, 562
diffusion coefficient, 27
diffusion current, 27
diffusion equation, 281
diffusion length, 34, 53
digital circuit:
bipolar, 441
CMOS, 256, 266
diode (*see* p–n junction)
diode equation, 46
diode leakage current, 57
direct tunneling, 128
discharge time:
for narrow-base diode, 69
for wide-base diode, 68
discrete dopant effects (*see* dopant number fluctuations)
donor, 17
donor level, 18–9
dopant, 17–8
dopant number fluctuations, 239–42
in SRAM cell, 485
double-gate (DG) MOSFET, 529
planar, 535

double-gate (DG) MOSFET (cont.)
vertical, 535
FinFET, 536
drain-current model, 149–59
drain-current overshoot in SOI, 519
drain-current saturation, 158–62, 186–92
drain-induced barrier lowering (*see* DIBL)
DRAM, 3, 495–500
DRAM cell, 496
planar capacitor, 497
read, 497
retention time, 500
scaling, 499
stacked capacitor, 497
trench capacitor, 497
write, 497
drift current, 23
drift-diffusion model, 192
drift transistor, 382
drift velocity, 23
dual n^+/p^+ polysilicon gates, 223, 538

Early voltage, 340
Ebers-Moll model, 352
ECL, 440
ECL scaling, 457
EEPROM, 501, 507
effective channel length:
definition, 242–3
extraction, 244–7
physical interpretation, 248–51
effective density of states, 16
effective electric field, 169, 325
effective generation and recombination centers, 556
effective intrinsic-carrier concentration, 325
effective mass of electrons, 13
density of state, 237, 596
longitudinal, 13
transverse, 13
effective mobility, 169–71
effective vertical field, 169
eigenfunction, orthogonality, 586
Einstein relations, 27, 543
electric field, 23, 28
at silicon surface, 76, 236
in intrinsic base, 381
in quasi-neutral region, 323, 381
normal, 29, 584
tangential, 29, 584
electromagnetic wave, 288
electromigration, 286
electron affinity, 73
electron diffusion coefficient, 27
electron-hole pair generation, 34, 122
electron mobility, 23, 170
electron trap, 101
electrostatic potential, 28, 177
emission rate:
electron, 554
hole, 554
emitter, 55, 318, 331, 374
emitter-base diode, 320, 350
emitter-coupled logic (*see* ECL)
emitter current, 334, 352
emitter current crowding, 328, 612
emitter delay time, 361, 451
emitter depth variation, 410
emitter design, 374
emitter diffusion capacitance, 359
emitter Gummel number, 332
emitter injection efficiency, 368
emitter junction depth, 320, 375
emitter series resistance, 338, 375, 605
emitter stripe width, 459, 607, 610
energy band (*see* band diagram)
energy gap (*see* bandgap)
energy-momentum relationship, 13, 594
epitaxially grown base, 380, 467
EPROM, 501
equipartition of energy, 545
equivalent circuit:
bipolar, 352
MOS capacitor, 87
MOSFET, 167, 245, 309
equivalent oxide charge per unit area, 104
erase, 507
excess minority carriers:
in the base, 64, 359
in the emitter, 359
extrinsic base, 339, 377
extrinsic-base resistance, 338, 605
extrinsic Debye length, 30–1
extrinsic silicon, 17

f_{max}:
bipolar, 620
MOSFET, 310, 603
f_T:
bipolar, 619
MOSFET, 310, 602
fall time, 263, 291
FAMOS, 511
fan in, 266, 307
fan out, 293
feedforward, 303
Fermi-Dirac distribution function, 14
Fermi integral, 590
Fermi level, 14–16, 20–2, 32
local, 32
Fermi potential, 29, 31
field crowding, 136
field-dependent mobility, 169
field-effect transistor, 148
field emission, 119
field oxide, 149, 270–1
FinFET, 536
fixed oxide charge, 101
Flash, 4, 507
Flash memory arrays, 507
bitline, 508
erase, 508
NAND, 509
NOR, 509
over erasure, 509
read, 508
wordline, 508
write, 508
flat-band condition, 74, 92
flat-band voltage, 74, 91
floating-body effect, 518–19
floating gate NVRAM cell, 511
coupling factor, 506
select gate, 512
sidewall floating gate, 512
split gate, 513
stacked gate, 513
folded bitline, 499
folded layout, 272
forward-active mode, 336, 355, 442
forward bias, 40
forward current gain, 352
forward transit time, 360, 450
Fowler-Nordheim tunneling, 127, 505
free electron level, 72–4
freeze-out, 21
fully-depleted SOI, 518, 520
fully velocity saturated current, 188
fuse, 500

GaAs HBT, 469
gamma function, 16, 597
gate-all-around MOSFET, 536
gate bias (voltage) equation, 76, 84
gate capacitance:
to channel, 173
to source-drain, 278–9
to substrate, 172
gate-controlled depletion charge, 180
gated diode, 94, 135
gate depletion width, 82
maximum, 82, 183, 218
gate-induced drain leakage (*see* GIDL)
gate length, 242–3
gate overdrive, 189
gate oxide, 73
breakdown, 137
limit, 219
gate resistance, 280–2
gate-to-source/drain overlap region, 278–9
gate tunneling current, 127–9
gate workfunction, 74, 91
effect on threshold voltage, 221
effect on channel profile design, 221–3
Gaussian doping profile, 227, 380
Gauss's law, 28, 80
Ge in base:
constant distribution, 406
linearly graded distribution, 390
optimal distribution, 414
trapezoidal distribution, 401
Ge in emitter, 396
generalized scale length, 184, 582
generalized scaling, 207–9
generation and recombination centers, 107, 553
generation rate (*see* emission rate)
generation-recombination, 34, 553
GIDL, 136
graded base bandgap, 390
graded doping profile, 382
gradual channel approximation, 150, 160, 189, 531
graphical representation of nonuniform channel doping, 232
ground-plane MOSFET, 230, 232
Gummel number:
base, 330
emitter, 332
Gummel plot, 61, 337

halo doping, 233–4
HBT, 390, 426, 469
heat dissipation, 209
heavily-doped silicon, 22–3
heavy doping effects, 325
heterojunction bipolar transistor (*see* HBT)
high-field effect:
in diode, 122
in gated diode, 135
in oxide, 127
high-field region, 192, 197
high field transport, 186, 192
high-frequency capacitance, 88–9
high-κ (permittivity) dielectric, 184
high-level injection, 47, 51, 325, 347
high-low doping profile, 225
history effect in SOI MOSFET, 519
hole diffusion coefficient, 27
hole mobility, 24, 171
hole trap, 102
hot-carrier effect:
channel hot electron, 198
channel hot hole, 199
substrate hot electron, 199
hot carriers, 133, 192, 197
hot-electron emission probability, 134
hot-electron injection, 133, 502
hot-electron reliability, 192, 217
hot hole injection, 199
hybrid-π model, 357

ideal current-voltage characteristics, 327
ideal diode, 61
ideality factor:
  collector current, 337, 422
  diode current, 61
i-layer, 43
image-force-induced barrier lowering, 133, 569
impact ionization, 122
impact ionization rate, 123
implanted emitter, 318, 374
implanted polysilicon gate, 91
impurity ionization energy, 19
impurity energy level, 18
impurity scattering, 23, 170
input capacitance, 294
input resistance, 357
input voltage, 257, 308, 441
input waveform, 263, 291
integrated base dose, 379
interconnect capacitance:
  fringing field, 283–4
  parallel-plate, 283–4
  wire-to-wire, 284–5
interconnect RC delay:
  global wires, 287–8
  local wires, 286–7
interconnect resistance, 286
interconnect scaling, 284–6
interdigitated layout, 282
interface-state generation, 129
interface states, 99
interface-trap capacitance, 104
interface trapped charge, 99
interface traps, 99
inter-poly oxide, 506
interstitial, 18
intrinsic base, 350, 378
intrinsic-base dopant distribution, 380
intrinsic-base resistance, 338, 605, 610
intrinsic capacitance of MOSFET, 172–3
intrinsic carrier concentration, 16
intrinsic Fermi level, 16
intrinsic potential, 28
intrinsic inverter delay, 294
intrinsic silicon, 16
inversion, 78
  strong, 78
  weak, 78
inversion layer:
  capacitance, 89, 173
  charge, 83
  quantum effect, 234–9
  thickness, 83, 239
inverter:
  propagation delay, 289–92
  switching waveform, 291
  transfer curve, 257–9
ion implantation:
  dose, 227
  Gaussian profile, 227–8
  straggle, 227
ionization energy, 18–19
isolation, 148–9, 429, 538
ITRS, 4
I–V characteristics:
  ballistic MOSFET, 591
  MOSFET, 159, 162, 186
  nMOSFET, 258, 292
  n–p–n, 327, 336, 337
  pMOSFET, 258, 292
  p–n junction, 51
  Schottky barrier diode, 115

junction breakdown, 122, 366
junction capacitance, 42, 278, 303–4
junction depth, 149, 183–4
junction isolation, 429

kinetic energy of electrons, 13, 545, 594
kink effect in SOI, 518
Kirk effect, 345

Laplace equation, 575
large-signal analysis, 352
latch up, 538
lateral field, 189
lateral source-drain doping gradient, 248–50
lateral transistor, 318
layout:
  bipolar transistor, 445
  CMOS inverter, 272
  folded, 272
  groundrules, 271
  MOSFET, 271
  two-way NAND, 273
leakage current, 57
lifetime, 52, 557
lightly-doped drain (LDD), 199, 201
linearly-extrapolated threshold voltage, 175, 291
linearly graded bandgap, 390
linear region, 156–7
lithography, 3
load capacitance, 290, 294, 297, 441
load resistor, 441
logic gate, 266, 290, 441
logic swing, 441
long-channel MOSFET (*see* MOSFET)
low-frequency capacitance, 88
low-high doping profile, 229–31
low-high-low profile, 231
low-level injection, 47
low-power CMOS, 221
low-temperature CMOS, 312–5
lucky electrons, 134

majority carrier, 21, 34, 37
majority-carrier response time (*see* dielectric relaxation time)
manufacturing tolerance (*see* process tolerances)
Matthiessen's rule, 23
maximum available gain, 598
maximum electric field, 40
maximum gate depletion width, 82, 183, 218
maximum oscillation frequency (f_{max}), 440, 603, 617
maximum oxide field, 218–19
Maxwell-Boltzmann statistics, 15–16
Maxwell equations, 27
mean free path, 134, 543
metallurgical channel length, 242–3
metal-oxide-semiconductor (*see* MOS)
metal-silicon contact, 108
metal work function, 74
microprocessor, 3
midgap work-function gate, 223
Miller effect, 302–3
minimum channel length, 183, 219
  bulk MOSFET, 183
  DG MOSFET, 534
  FD-SOI MOSFET, 521
  with high-κ gate dielectric, 185
minimum feature size, 2–3, 271
minimum overlap, 280
minimum threshold voltage, 214, 220
minority carriers, 21, 34, 37
minority-carrier current, 54
minority-carrier diffusion length, 34, 62
minority-carrier lifetime, 34, 62
  electrons, 52, 558
  holes, 557
minority-carrier mobility, 62
MNOS:
  reversed source-drain mode, 514
  two-bits per cell, 515
mobile ions, 101
mobility:
  bulk, 24
  effective, 169
  electron, 24
  hole, 24
  minority carrier, 62
  MOSFET channel, 169–72
  temperature dependence, 23, 172
  universal, 170–1
  vertical field dependence, 170–1
MOS:
  band diagram, 73–5, 77, 84, 92
  capacitor, 72
  C–V characteristics, 86–90
  equivalent circuit, 87
  under nonequilibrium, 94–8
MOSFET:
  admittance matrix, extrinsic, 602
  admittance matrix, intrinsic, 309
  body effect, 166
  breakdown, 200
  buried channel, 222
  channel mobility, 169
  common-source RF circuit, 308
  current gain, extrinsic, 603
  current gain, intrinsic, 310
  drain current model, 149
  effective channel length, 242
  equipotential contours, 178
  intrinsic capacitance, 172
  I–V characteristics, 155
  linear region characteristics, 156
  output conductance, 255, 308
  p-channel, 162
  saturation region characteristics, 158
  scale length, 183–4, 582
  scaling, 204
  short-channel, 175
  small-signal equivalent circuit, 309, 602
  subthreshold characteristics, 163
  threshold voltage, 156, 212
  transconductance, extrinsic, 602
  transconductance, intrinsic, 192, 308
  unity-current-gain frequency, extrinsic, 603
  unity-current-gain frequency, intrinsic, 310
  unity-power-gain frequency, 604
  voltage gain, 310
multiple-gate (MG) MOSFET:
  omega-gate, 536
  pi-gate, 536
  quadruple-gate, 536
  surrounding-gate, 536
  triple-gate, 536
multiplication factor, 123, 367, 573

NAND
  EEPROM, 510
NAND gates, 266
  two-way, 268, 304
nanowire MOSFET, 536
narrow-base diode, 57
n-channel MOSFET (*see* MOSFET)
negative bias temperature instability (NBTI), 199
net recombination rate, 556
noise margin, 259, 269, 482
  CMOS inverter, 259
  CMOS NAND, 269
  SRAM cell, 482
nonequilibrium transport, 192
non-ideal base current, 347
non-scaling factors:
  primary, 210
  secondary, 211
non-transparent emitter (*see* deep emitter)
non-transparent polysilicon emitter, 400
nonuniform channel doping, 224–34

nonvolatile memory:
  cell, 511, 514
  charge injection, 502
  charge storage, 505
  data retention time, 501, 510
  endurance, 510
  erasure, 507
  programming, 508
  read, 508
  write, 508
nonvolatile random access memory (NVRAM), 500
NOR:
  EEPROM, 509
  logic gate, 266
n–p–n transistor (*see* bipolar transistor)
n-type silicon, 17
n-well, 257, 538

off current, 169, 210, 213
  requirement, 214
ohmic contact, 120, 276
Ohm's law, 24
on current, 214, 263, 295
one-sided junction, 42
optical phonon, 26
output capacitance, 294
output conductance, 195, 255, 308
output resistance, 357
output voltage, 257, 260, 310, 441
over erasure, 509
overlap capacitance, 279, 302
oxide (*see* silicon dioxide)
oxide charge, 98, 103
oxide field, 76, 128
  maximum, 219
oxide trapped charge, 101

packing density, 1
Pao-Sah's double integral, 153
parabolic band, 13
parabolic region, 157
parallel-plate capacitance, 283
parasitic capacitance, 277, 356
parasitic resistance, 196, 245, 274, 301, 338, 356
partially-depleted SOI, 518
p-channel MOSFET, 162
pedestal collector, 431
performance factor of bipolar circuits:
  analog, 463
  digital, 441
performance factor of CMOS circuits, 256
  advanced, 307
  low temperature, 312
  SiGe, 311
  SOI, 519
permittivity:
  Si, 28
  SiO_2, 75
  vacuum, 28
phonon scattering, 23, 169
phosphorus, 18
physical properties of Si and SiO_2, 13
pinch-off, 159
  current, 158
  voltage, 158
p–i–n diode, 43
Planck's constant, 12
p–n diode:
  breakdown, 122
  capacitance, 41
  depletion approximation, 38
  forward-biased, 40
  I–V characteristics, 51
  leakage, 57
  reverse-biased, 40
  saturation current, 57
p–n junction, 35
pn product, 17
p–n–p transistor (*see* bipolar transistor)
pocket doping, 233
Poisson's equation, 27
polySiGe emitter, 400
polysilicon base contact, 430
polysilicon emitter, 321, 333, 430
  transparent, 333
  non-transparent, 400
polysilicon gate, 91
  depletion, 91–4
  depletion capacitance, 93
  dual n^+/p^+, 5, 538
potential barrier, 177–9
power-delay product, 206, 209, 300
power density, 206, 209, 462
power dissipation:
  active, 220, 265, 299
  cross-over current, 265
  standby, 213
power gain, 598
power supply voltage, 208, 219–21, 300
power vs. delay trade-off:
  bipolar circuits, 453
  CMOS circuits, 299
principle of detailed balance, 554
process flow:
  bipolar, 542
  CMOS, 538
process tolerances, 212
PROM, 500
propagation delay, 289
  CMOS inverter chain, 289–91
  ECL, 440
  two-way NAND, 304–6
p-type silicon, 18
pull-down delay, 264, 293
pull-up delay, 264, 293
punch-through in short-channel MOSFET, 179

quantum confinement, 234, 594
quantum effect on threshold voltage, 234, 237
quasi-Fermi level, 32
quasi-Fermi potential, 33
quasineutrality, 47
quasineutral region, 37, 320
quasistatic assumption, 265
quasistatic C–V curve, 86

radiative process, 34
random access memory (RAM), 476
RC delay:
  interconnect, 287
  MOSFET gate, 282
reach-through, 318, 385
read, 480, 497, 508
reciprocity, 353
recombination, 34, 553
rectifier, 41
refresh, 496
resistivity of aluminum, 287
resistivity of bulk silicon, 25
retention time, 89
  DRAM data, 500
  NVRAM, 501
retrograde doping profile, 229, 388
  extreme, 230
reverse bias, 40
reverse current, 352
reverse current gain, 352
reverse Early effect, 419
RF (radio-frequency) circuits, 308
Richardson's constant, 118
ring oscillator, 290, 440
rise time, 289

Sah–Noyce–Shockley diode equation, 60, 561
saturated base current density, 332
saturated collector current density, 330
saturation current of MOSFET, 158, 188, 191–2, 592
saturation currents in bipolar, 369
saturation point, 160–1, 189
saturation region, 336
saturation velocity, 26, 186
saturation voltage of MOSFET, 158, 188, 191, 592
scale length:
  bulk MOSFET, 183–4
  double-gate MOSFET, 533
  surrounding-gate MOSFET, 537
scale length model:
  generalized, 582
  high-κ gate dielectric, 184
  one region, 183
  two region, 582
  three region, 584
scaling:
  bipolar, 457
  constant-field, 204
  constant-voltage, 209
  generalized, 207
  ideal, 204
  interconnect, 284
  MOSFET, 204–12
scaling limit:
  bipolar, 460
  bulk MOSFET, 219
  DG MOSFET, 534
  FD-SOI MOSFET, 522
  with high-κ gate dielectric, 219
scaling rules:
  bipolar, 458
  interconnect, 286
  MOSFET–constant field, 206
  MOSFET–generalized, 209
scattering:
  interface, 169
  ionized impurity, 23
  phonon, 23
  surface roughness, 170
scattering theory of MOSFET, 192
Schottky barrier:
  barrier height, 109
  effect of electric field, 114
  for electrons, 109
  for holes, 114
  measured barrier heights, 113
Schottky barrier diode (Schottky diode), 108
Schottky barrier emission:
  field emission, 119
  thermionic emission, 116
  thermionic-field emission, 119
Schrödinger equation, 235
self-aligned base contact, 430
self-aligned silicide, 276
sense amplifier, 477, 498
series resistance, 196, 245, 301
shallow emitter, 331
shallow impurities, 18
shallow trench isolation, 5, 538
sheet resistance, 25
sheet resistivity, 25
  aluminum, 274
  base, 379
  channel, 196
  gate, 280
  silicide, 276
  source drain, 276
  thin film, 25
shift and ratio (S&R) method, 247
Shockley diode current equation, 54

Shockley diode equation, 50
Shockley-Read recombination, 34, 553
short-channel effect, 176, 575
  DIBL, 177
  extreme retrograde doped MOSFET, 581
  ground-plane MOSFET, 581
  substrate sensitivity, 581
  subthreshold current slope, 580
  threshold voltage (V_t) rolloff, 176, 182
short-channel MOSFET, 175
short-circuit current, 265
SiGe base, 389, 431
SiGe-base bipolar transistor:
  comparison with GaAs HBT, 469
  heterojunction nature, 426
SiGe MOSFET, 311
silicide, 276, 538
silicon (Si):
  bandgap, 12
  covalent bonds, 11
  degenerate, 22
  dielectric constant, 13
  energy bands, 12, 18
  extrinsic, 17
  intrinsic, 16
  lattice constant, 13
  n-type, 17
  permittivity, 28
  physical properties, 13
  polycrystalline, 91
  p-type, 17
  resistivity, 25
  solid solubility, 20
  thermal conductivity, 13
silicon dioxide:
  band diagram, 73
  breakdown event, 137
  breakdown field, 138
  charge, 98
  charge to breakdown, 139
  damage, 105
  defect generation, 129
  defects, 129
  dielectric constant, 13
  permittivity, 75
  physical properties, 13
  successive breakdown, 137
  time to breakdown, 139
  trapped charge, 99, 101
  tunneling, 127
silicon-germanium base (*see* SiGe base)
silicon-on-insulator (SOI), 517
silicon-rich oxide, 512
SIMOX, 517
Si–SiGe n–p diode, 614
Si-SiO_2 interface, 99
Si-SiO_2 system, 73, 98
slow states, 105
small-signal analysis, 308, 356, 438
Smart-Cut, 517
sodium ion contamination, 101
soft error, 518
SOI bipolar, 523
SOI CMOS, 517
solid solubility, 20
source and drain junction depth, 183, 274
source-drain series resistance, 196, 245, 274
source-drain sheet resistivity, 276
source starvation, 593
source (thermal) injection velocity, 194
source-to-body potential, 268, 305
source-to-drain current, 149
source-to-drain current at threshold, 213
space-charge region (space-charge layer), 38
space-charge region current, 553
specific contact resistivity, 121, 276
SPICE, 290
spin, 13
split C–V measurement, 89–90
spreading resistance, 275
SRAM (static random access memory), 477
SRAM cell:
  bipolar, 487
  depletion load, 487
  device sizing, 482
  full CMOS, 478
  read, 480
  resistor load, 487
  scaling, 485
  static noise margin, 482
  TFT load, 487
  write, 481
standardized signal, 290
standby power, 213, 221, 301
static CMOS circuit, 256
static power dissipation, 213
step input, 263
step profile, 225, 229
storage capacitor, 496
stored minority-carrier charge (*see* excess minority carriers)
straggle, 227
strained-silicon MOSFET, 311
strong inversion, 78, 80
sub-band, 235
  energy level, 236–7
subcollector, 318, 339, 385
substitutional, 18
substrate bias, 166
  forward, 207, 231
  reverse, 168, 207
substrate current, 197–8
substrate sensitivity, 166
subthreshold characteristics, 164–6
subthreshold current, 165, 210
subthreshold non-scaling, 210
subthreshold slope, 165, 580
surface-channel MOSFET, 223
surface electric field, 76
  at threshold, 232
surface generation-recombination
    center, 107
surface potential, 76
  effect of oxide charge, 103
surface potential based compact model, 155
surface recombination velocity, 331
surface scattering, 169
surface state density, 100
surface states, 99
surrounding-gate MOSFET, 536
switch current, 441
switching delay, 264, 293, 442
switching energy, 300
switching power, 299
switching resistance, 294
  inverter, 294
  two-way NAND, 307
switching trajectory, 292
switching waveform:
  abrupt input transition, 263
  linear chain of inverters, 291
  two-way NAND, 305
symbols:
  bipolar transistor, 319
  nMOSFET, 257
  pMOSFET, 257

temperature dependence:
  CMOS performance, 314
  Fermi level, 22
  mobility, 23
  off-current, 213
  threshold voltage, 168
thermal energy, 14
thermal velocity, 543
thermal voltage, 210
thermionic emission, 116
thermionic-field emission, 119
threshold voltage:
  design, 213, 219
  discrete dopant effect, 239
  linearly-extrapolated, 175
  long-channel MOSFET, 156
  minimum, 214
  multiple, 221
  nonuniformly-doped MOSFET, 224
  performance sensitivity, 215
  quantum correction, 238
  requirement, 213
  roll-off, 176
  short-channel MOSFET, 182
  substrate-bias dependence, 166
  temperature dependence, 167
  tolerances, 218
  work-function effect, 221
  $2\psi_B$ defined, 156
time-dependent analysis, 361
time to breakdown, 139
t-inversion (t_{inv}), 175, 239
transconductance:
  bipolar, 357
  large signal, 295, 299
  linear, 175
  MOSFET, 308
  saturation, 192
transfer characteristics:
  inverter, 259
  two-way NAND, 269
transfer device, 496, 500
transfer length, 276
transfer ratio (DRAM read), 498
transistor equation, 365
transit time:
  base, 65, 384, 452
  base-collector, 361, 452
  base-emitter, 361, 451
  emitter, 361, 451
  forward, 360, 450
  MOSFET, 265
transmission coefficient (in tunneling), 121
transmission line delay, 288
transmission line model, 276
transparent emitter (*see* shallow emitter)
trap-assisted tunneling:
  bulk-trap-assisted, 130
  interface-trap-assisted, 131
  into an electron trap, 129
traps:
  Coulomb-attractive, 102
  Coulomb-repulsive, 102
  electron, 102
  hole, 102
  neutral, 102
trench isolation, 429
triangular potential well, 236, 595
triode region (*see* linear region)
tunneling:
  band-to-band, 108, 125
  direct, 128
  Fowler-Nordheim, 127
  into silicon dioxide, 127
  through silicon dioxide, 127, 219
tunneling current, 126, 128
tunnel oxide, 506
  silicon rich, 512
two-dimensional effect, 177
two-dimensional electron gas
    (2DEG), 235
two-port network, 598

two-way NAND:
  propagation delay, 306
  switching waveform, 305
  top and bottom switching, 268, 304
  transfer curves, 269

uniformly-doped channel, 223
unilateral power gain, 599
unity current gain frequency (f_T):
  bipolar, 619
  MOSFET, 309, 602
unity power-gain frequency:
  bipolar, 620
  MOSFET, 603
universal mobility behavior, 169

vacuum level (*see* free electron level)
valence band, 11
V_{BE} reference, 422
velocity-field relationship, 26, 187, 190
velocity of light, 288–9
velocity overshoot, 192
velocity saturation, 26, 186
  current limit, 188
velocity saturation model:
  $n = 1$, 187
  $n = \infty$, 190
  piecewise, 191
vertical transistor, 318
very-large-scale-integration (VLSI), 1–3
voltage gain, 310, 464
volume inversion, 530, 533

weak inversion, 78
Webster effect, 325
wide-base diode, 57
wide-gap emitter, 426, 470
wide-gap-emitter HBT, 426, 469
wire capacitance, 283
wire resistance, 286
wordline, 477
  boosted, 497
work function, 74
  metal, 74
  midgap, 223
  n^+ polysilicon, 91
  n-type silicon, 78
  p^+ polysilicon, 91
  p-type silicon, 74
write, 481, 497, 508
write disturb, 504