

Operation and Modeling of the MOS Transistor

Operation and Modeling of the MOS Transistor

THIRD EDITION

Yannis Tsividis
Columbia University

Colin McAndrew
Freescale Semiconductor

New York Oxford
OXFORD UNIVERSITY PRESS
2011

Oxford University Press, Inc., publishes works that further Oxford University's objective of excellence in research, scholarship, and education.

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Copyright © 2011, 1999 by Oxford University Press, Inc.; 1987 by McGraw-Hill, Inc.

For titles covered by Section 112 of the US Higher Education Opportunity Act, please visit www.oup.com/us/he for the latest information about pricing and alternate formats.

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016
<http://www.oup.com>

Oxford is a registered trademark of Oxford University Press

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

Library of Congress Cataloging-in-Publication Data

Tsividis, Yannis.

Operation and modeling of the MOS transistor / Yannis Tsividis, Colin McAndrew.—3rd ed.
p. cm.

Includes index.

ISBN 978-0-19-517015-3

1. Metal oxide semiconductors—Mathematical models. 2. Metal oxide semiconductor field-effect transistors—Mathematical models. I. McAndrew, Colin. II. Title.

III. Title: MOS transistor.

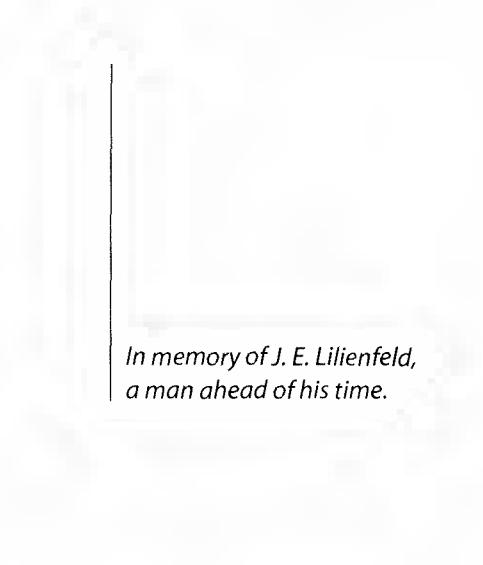
TK7871.99.M44T77 2010

621.3815'284—dc22

1052387

Printing number: 9 8 7 6 5 4 3 2 1

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



*In memory of J. E. Lilienfeld,
a man ahead of his time.*

About the Authors

Yannis Tsividis received the BS degree from the University of Minnesota, Minneapolis, in 1972, and the MS and PhD degrees from the University of California, Berkeley, in 1973 and 1976, respectively.

His involvement with MOS transistors began in the early seventies as part of his industrial and doctoral work. In 1975, to prove the feasibility of MOS technology for analog and mixed-signal integrated circuits, he designed and built a fully integrated MOS operational amplifier and demonstrated its use in a PCM codec. Since that time, much of his work has focused on device and circuit issues in the merging of analog and digital circuits on the same chip. In addition to this book, he has written *Mixed Analog-Digital Devices and Technology* (World Scientific, 2002) and has edited several other books. He is the inventor/coinventor in over 20 patents.

Dr. Tsividis is Charles Batchelor Professor of Electrical Engineering at Columbia University in New York. He has also taught, as part of regular or visiting appointments, at the University of California, Berkeley, the Massachusetts Institute of Technology, and the National Technical University of Athens. He has worked for Motorola Semiconductor and for AT&T Bell Laboratories.

Dr. Tsividis received the 1984 IEEE W. R. G. Baker Award for the best IEEE publication, and is recipient or corecipient of best paper awards from the European Solid-State Circuits Conference in 1986, the IEEE International Solid-State Circuits Conference in 2003, and the IEEE Circuits and Systems Society (Darlington Award, 1987; Guillemin-Cauer Award, 1998 and 2008). He has received Columbia's Presidential Award for Outstanding Teaching in 2003, and the IEEE Undergraduate Teaching Award in 2005. He is a Fellow of the IEEE and received the IEEE Gustav Robert Kirchhoff Award in 2007.

Colin McAndrew received the BE (Hons) degree in Electrical Engineering from Monash University, Victoria, Australia, in 1978 and the MSc and PhD degrees in Systems Design Engineering from the University of Waterloo, Ontario, Canada, in 1982 and 1984, respectively.

He became involved with modeling semiconductor devices for circuit simulation in 1987 and has contributed to the development of models for MOS, bipolar, and passive devices. He developed the backward-propagation-of-variation technique for statistical modeling and has been a primary advocate of the use of Verilog-A and compilers for device modeling.

He is a Fellow of the IEEE, was the recipient of the Ian Langlands Medal of the Institute of Engineers of Australia in 1978, and was recognized with the IEEE BCTM (Bipolar/BiCMOS Circuits and Technology Meeting) award in 2005. He has received best paper awards for the IEEE ICMTS (International Conference on Microelectronic Test Structures) in 1993 and the IEEE CICC (Custom Integrated Circuits Conference) in 2002. He is an editor of the IEEE Transactions on Electron Devices, and is or has been on the technical program committees for the IEEE BCTM, ICMTS, CICC, and BMAS (Behavioral Modeling and Simulation) conferences.

Contents

PREFACE

xvii

CHAPTER 1

Semiconductors, Junctions, and MOSFET Overview	1
1.1 Introduction	1
1.2 Semiconductors	1
1.2.1 Intrinsic Semiconductors, Free Electrons, and Holes	2
1.2.2 Extrinsic Semiconductors	4
1.2.3 Equilibrium in the Absence of Electric Field	7
1.2.4 Equilibrium in the Presence of Electric Field	10
1.2.5 Nonequilibrium; Quasi-Fermi Levels	12
1.2.6 Relations between Charge Density, Electric Field, and Potential; Poisson's Equation	13
1.3 Conduction	15
1.3.1 Transit Time	15
1.3.2 Drift	17
1.3.3 Diffusion	22
1.3.4 Total Current	25
1.4 Contact Potentials	26
1.5 The <i>pn</i> Junction	32
1.6 Overview of the MOS Transistor	43
1.6.1 Basic Structure	43
1.6.2 A Qualitative Description of MOS Transistor Operation	47
1.6.3 A Fluid Dynamical Analog	49
1.6.4 MOS Transistor Characteristics	52
1.7 Fabrication Processes and Device Features	54
1.8 A Brief Overview of This Book	60
References	62
Problems	63

ix

CHAPTER 2

The Two-Terminal MOS Structure	65
2.1 Introduction	65
2.2 The Flatband Voltage	66
2.3 Potential Balance and Charge Balance	73
2.4 Effect of Gate-Body Voltage on Surface Condition	75
2.4.1 Flatband Condition	75
2.4.2 Accumulation	75
2.4.3 Depletion and Inversion	76
2.4.4 General Analysis	80
2.5 Accumulation and Depletion	86
2.6 Inversion	88
2.6.1 General Relations and Regions of Inversion	88
2.6.2 Strong Inversion	94
2.6.3 Weak Inversion	98
2.6.4 Moderate Inversion	101
2.7 Small-Signal Capacitance	102
2.8 Summary of Properties of the Regions of Inversion	111
References	111
Problems	112

CHAPTER 2

The Three-Terminal MOS Structure	115
3.1 Introduction	115
3.2 Contacting the Inversion Layer	115
3.3 The Body Effect	131
3.4 Regions of Inversion	132
3.4.1 Approximate Limits	132
3.4.2 Strong Inversion	136
3.4.3 Weak Inversion	138
3.4.4 Moderate Inversion	141
3.5 A “ V_{CB} Control” Point of View	141
3.5.1 Fundamentals	141
3.5.2 The “Pinchoff Voltage”	145
3.6 Uses for Three-Terminal MOS Structures	147
References	148
Problems	149

CHAPTER 4

The Four-Terminal MOS Transistor	151
4.1 Introduction	151
4.2 Transistor Regions of Operation	156

4.3	Complete All-Region Model	158
4.4	Simplified All-Region Models	172
4.4.1	Linearizing the Depletion Region Charge	172
4.4.2	Body-Referenced Simplified All-Region Models	173
4.4.3	Source-Referenced Simplified All-Region Models	176
4.4.4	Charge Formulation of Simplified All-Region Models	177
4.5	Models Based on Quasi-Fermi Potentials	181
4.6	Regions of Inversion in Terms of Terminal Voltages	183
4.7	Strong Inversion	186
4.7.1	Complete Strong-Inversion Model	186
4.7.2	Body-Referenced Simplified Strong-Inversion Model	192
4.7.3	Source-Referenced Simplified Strong-Inversion Model	192
4.7.4	Model Origin Summary	203
4.8	Weak Inversion	204
4.8.1	Special Conditions in Weak Inversion	204
4.8.2	Body-Referenced Model	205
4.8.3	Source-Referenced Model	206
4.9	Moderate-Inversion and Single-Piece Models	208
4.10	Source-Referenced vs. Body-Referenced Modeling	210
4.11	Effective Mobility	212
4.12	Effect of Extrinsic Source and Drain Series Resistances	222
4.13	Temperature Effects	224
4.14	Breakdown	226
4.15	The <i>p</i> -Channel MOS Transistor	228
4.16	Enhancement-Mode and Depletion-Mode Transistors	230
4.17	Model Parameter Values, Model Accuracy, and Model Comparison	231
	References	233
	Problems	240

CHAPTER 5

	Small-Dimension Effects	243
5.1	Introduction	243
5.2	Carrier Velocity Saturation	244
5.3	Channel Length Modulation	253
5.4	Charge Sharing	259
5.4.1	Introduction	259
5.4.2	Short-Channel Devices	261
5.4.3	Narrow-Channel Devices	266
5.4.4	Limitations of Charge-Sharing Models	270
5.5	Drain-Induced Barrier Lowering	271
5.6	Punchthrough	275
5.7	Combining Several Small-Dimension Effects into One Model —A Strong-Inversion Example	277
5.8	Hot Carrier Effects; Impact Ionization	280
5.9	Velocity Overshoot and Ballistic Operation	285

5.10	Polysilicon Depletion	288
5.11	Quantum Mechanical Effects	293
5.12	DC Gate Current	295
5.13	Junction Leakage; Band-to-Band Tunneling; GIDL	302
5.14	Leakage Currents—Particular Cases	305
5.15	The Quest for Ever-Smaller Devices	307
5.15.1	Introduction	307
5.15.2	Classical Scaling	308
5.15.3	Modern Scaling	312
	References	316
	Problems	327

CHAPTER 6

1	The MOS Transistor in Dynamic Operation— Large-Signal Modeling	329
---	---	-----

6.1	Introduction	329
6.2	Quasi-Static Operation	330
6.3	Terminal Currents in Quasi-Static Operation	334
6.4	Evaluation of Intrinsic Charges in Quasi-Static Operation	341
6.4.1	Introduction	341
6.4.2	Strong Inversion	342
6.4.3	Moderate Inversion	348
6.4.4	Weak Inversion	348
6.4.5	All-Region Model	350
6.4.6	Depletion and Accumulation	352
6.4.7	Plots of Charges vs. V_{GS}	353
6.4.8	Use of Intrinsic Charges in Evaluating the Terminal Currents	354
6.5	Transit Time under DC Conditions	355
6.6	Limitations of the Quasi-Static Model	357
6.7	Non-Quasi-Static Modeling	363
6.7.1	Introduction	363
6.7.2	The Continuity Equation	364
6.7.3	Non-Quasi-Static Analysis	365
6.8	Extrinsic Parasitics	371
6.8.1	Extrinsic Capacitances	371
6.8.2	Extrinsic Resistances	374
6.8.3	Temperature Dependence	378
6.8.4	Simplified Models	378
	References	379
	Problems	383

CHAPTER 7

Small-Signal Modeling for Low and Medium Frequencies	385
7.1 Introduction	385
7.2 A Low-Frequency Small-Signal Model for the Intrinsic Part	386
7.2.1 Introduction	386
7.2.2 Small-Signal Model for the Drain-to-Source Current	386
7.2.3 Small-Signal Model for the Gate and Body Currents	390
7.2.4 Complete Low-Frequency Small-Signal Model for the Intrinsic Part	393
7.2.5 Strong Inversion	396
7.2.6 Weak Inversion	407
7.2.7 Moderate Inversion	409
7.2.8 All-Region Models	409
7.3 A Medium-Frequency Small-Signal Model for the Intrinsic Part	414
7.3.1 Introduction	414
7.3.2 Intrinsic Capacitances	414
7.4 Including the Extrinsic Part	435
7.5 Noise	436
7.5.1 Introduction	436
7.5.2 White Noise	440
7.5.3 Flicker Noise	450
7.5.4 Noise in Extrinsic Resistances	456
7.5.5 Including Noise in Small-Signal Circuits	456
7.6 All-Region Models	456
References	458
Problems	469

CHAPTER 8

High-Frequency Small-Signal Models	473
8.1 Introduction	473
8.2 A Complete Quasi-Static Model for the Intrinsic Part	474
8.2.1 Complete Description of Intrinsic Capacitance Effects	474
8.2.2 Small-Signal Equivalent Circuit Topologies	478
8.2.3 Evaluation of Capacitances	484
8.2.4 Frequency Region of Validity	491
8.3 y -Parameter Models	492
8.4 Non-Quasi-Static Models	499
8.4.1 Introduction	499
8.4.2 A Non-Quasi-Static Strong-Inversion Model	500
8.4.3 Other Approximations and Higher-Order Models	519
8.4.4 Model Comparison	522

xiv CONTENTS

8.5	High-Frequency Noise	524
8.6	Considerations in MOSFET Modeling for RF Applications	529
	References	538
	Problems	542
CHAPTER 6		
	Substrate Nonuniformity and Other Structural Effects	547
9.1	Introduction	547
9.2	Ion Implantation and Substrate Nonuniformity	548
9.3	Substrate Transverse Nonuniformity	551
9.3.1	Preliminaries	551
9.3.2	Threshold Voltage	555
9.3.3	Drain Current	564
9.3.4	Buried-Channel Devices	566
9.4	Substrate Lateral Nonuniformity	571
9.5	Well Proximity Effect	577
9.6	Stress Effects	581
9.7	Statistical Variability	584
	References	592
	Problems	598
CHAPTER 10		
	Modeling for Circuit Simulation	600
10.1	Introduction	600
10.2	Types of Models	601
10.2.1	Models for Device Analysis and Design	601
10.2.2	Device Models for Circuit Simulation	602
10.3	Attributes of Good Compact Models	606
10.4	Model Formulation	608
10.4.1	General Consideration and Choices	609
10.5	Model Implementation in Circuit Simulators	615
10.6	Model Testing	618
10.7	Parameter Extraction	618
10.8	Simulation and Extraction for RF Applications	635
10.9	Common MOSFET Models Available in Circuit Simulators	638
10.9.1	BSIM	638
10.9.2	EKV	640
10.9.3	PSP	640
10.9.4	Other Models	642
	References	642
	Problems	648

APPENDICES

A	Basic Laws of Electrostatics in One Dimension	653
B	Quasi-Fermi Levels and Currents	659
C	General Analysis of the Two-Terminal MOS Structure	661
D	Careful Definitions for the Limits of Moderate Inversion	666
E	General Analysis of the Three-Terminal MOS Structure	669
F	Drain Current Formulation Using Quasi-Fermi Potentials	674
G	Modeling Based on Pinchoff Voltage and Related Topics	678
H	Evaluation of the Intrinsic Transient Source and Drain Currents	684
I	Quantities Used in the Derivation of the Non-Quasi-Static <i>y</i> -Parameter Model	687
J	Analysis of Buried-Channel Devices	690
K	MOSFET Model Benchmark Tests	700

INDEX

713

Preface

This book provides a unified treatment of the many phenomena encountered in the operation of modern MOS transistors and shows how such phenomena can be modeled analytically. The book is mainly written for use in a senior or first-year graduate course. It is felt that electrical engineering students have much to gain from a course devoted to the subject. The MOS transistor is the dominant VLSI device. A course devoted to it is invaluable to those planning a career in device physics and modeling. For such people, the standard courses on semiconductor devices usually cover too many different devices to do justice to any one of them and do not present the intricacies and trade-offs involved in a detailed modeling effort. A course devoted to the MOS transistor is also extremely valuable for those who want to use the device to design state-of-the-art circuits. Integrated circuit designers have the opportunity to select devices to meet circuit needs, and they can do this most intelligently if they really understand the workings of the devices. One can, of course, design systems by using predesigned circuit building blocks as black boxes, if truly high performance is not important. But when state-of-the-art performance is a must, one has to consider device details. In addition, a deep knowledge of device operation and modeling is needed for understanding the computer simulator models a designer is working with, and for identifying their limitations. Many circuit designers in the industry spend endless hours trying to interpret strange circuit simulation results, not realizing that these are largely due to modeling inadequacies. Without adequate device understanding, valuable time and effort is likely to be wasted on overdesign, brute-force approaches, and design iterations. These authors believe that

no integrated circuit designer's education is complete without detailed exposure to MOS transistor operation and modeling.

In the decade that has passed since the publication of the second edition of this book, there have been significant advances in the understanding and modeling of the MOS transistor. In addition, the requirements for modeling this device on the part of the circuit design community are now much more demanding. For example, the push for ever-smaller dimensions has revealed phenomena previously ignored. In addition, industrial surface potential-based models are now a reality. Thus, the book has been extensively revised, as discussed in the description of the individual chapters below. To give the reader an idea of the extent of the revisions, suffice it to say that, although there are now two coauthors, each has labored more than the first author did when he revised the first edition!

The basic philosophy of the first two editions has been retained. Several aspects of this philosophy are summarized below.

- The book starts with basic concepts. Readers should be able to follow even if they had no prior exposure to the device. The discussions of these concepts are often from a perspective different from the one usually taken, thus making them interesting reading even to those with prior exposure.
- Every effort has been made to give the subject a careful treatment. The reader may at times get the feeling that the authors are “splitting hairs.” The authors would rather be accused of this than fudge. The MOS transistor is a device so complex that once one has decided to fudge, things grow out of hand very quickly, and one ends up with a hodgepodge of careless derivations, conflicting models, and a lot of patchwork. This has been avoided at all costs. Also, the reason for hair-splitting at some points can only be appreciated further along in the book, where those “too fine” details can be seen to make a lot of difference. The use of one name for several quantities, common in some of the literature, is carefully avoided. For example, at least four distinct quantities encountered in MOS transistor work are described by using the single name “threshold voltage.” Although the reader is warned of this practice, the practice itself is avoided in this book.
- The emphasis is on principles. At the same time, to illustrate these principles, relevant models are extensively derived and discussed. Thus, physics and modeling are discussed in parallel throughout the book.
- Analytical results are derived in a logical manner after carefully stating the assumptions made. Empirical modeling is avoided as much as possible. However, there are phenomena for which the only analytical results available are empirical or semiempirical. Such results are presented for completeness after pointing out the necessary hand-waving behind them.
- This book is *not* a survey. In fact, a conscious effort has been made to avoid making it one. A well-connected set of topics has been chosen, and most of these are discussed in significant detail. Nevertheless, for completeness, certain other topics are mentioned, albeit briefly. In such cases, some representative results are shown without proof so that the reader can know what to expect if he or she consults the references provided.

- A great deal of emphasis is placed on providing intuition for the various phenomena discussed. It is rather hopeless to attempt working with a device as complex as the MOS transistor relying only on analytical relations. The emphasis on intuition has made lengthy discussions necessary.
- The pace is unhurried. The authors believe that this actually makes it possible to study the material faster. Thus, whereas the treatment of a given topic may be long in terms of number of pages, it actually should take less time to comprehend it, because of both the detailed derivations and the intuitive discussions. At times, the reader may get the feeling of *déjà vu*, because some points are repeated more than once to make sure they are not missed, especially if the reading of the topics is done out of sequence. In general, the book is written in the style in which the authors would like to have any new subject presented to them. They would very much like to see the new subject “beaten to death,” presented with several points of view to increase perspective and with a significant amount of repetition. The authors have been in the past grateful for treatments of this type and never felt offended by this style. If the reader happens to be “faster” than the authors in this respect, he or she can easily skip some of the discussions.

All chapters in this book, and almost all sections within each chapter, have been extensively revised. Many sections have been rewritten, and new ones have been added. One chapter is entirely new. A list of chapters follows, along with an explanation of their features and of what is new in them in this edition.

Chapter 1: Semiconductors, Junctions, and MOSFET Overview

All preliminary material necessary for the understanding of MOS structures is given here. This material is important to the newcomer, but part of it should also make interesting reading for those with some previous exposure to basics. This includes the material on contact potentials, which is used to advantage in the following chapter. New in this edition is an introduction to energy band concepts. Although treatment without such concepts was favored in the previous editions, it was found that the lack of energy band discussions hindered the students when they attempted to read the literature. Thus, energy bands have been introduced, but in such a way that they do not impede the earlier flow; for the most part, it is possible to skip the corresponding sections if desired. The chapter concludes with an overview of the MOS transistor. This provides a framework for the rest of the book and makes it clear why particular details of the two- and three-terminal MOS structures are studied in the following two chapters.

Chapter 2: The Two-Terminal MOS Structure

Here the reader will find a treatment of the MOS structure with gate and substrate terminals only. Concepts not directly related to the presence of the source and drain in the MOS transistor are treated here. The regions of weak, moderate, and strong inversion are all introduced in this chapter. Energy band concepts have been added in this edition, but a treatment using surface potentials continues to

be emphasized. A new treatment of carrier concentrations has been added in this edition, which helps make clear the contributions of the various types of charge in the semiconductor.

Chapter 3: The Three-Terminal MOS Structure

Here one more terminal is added to the structure of Chap. 2 to connect the inversion layer to the external world. The MOS transistor concepts that are not directly related to current flow are presented in this chapter. This includes the important “body effect,” which is amply treated. The material here has been streamlined and shortened to include only the most dominant approaches. Other parts of this chapter have been moved to an appendix. A recently corrected, widely used expression for carrier concentrations is adopted.

Chapter 4: The Four-Terminal MOS Transistor

The four-terminal MOS transistor is obtained in this chapter by adding one terminal to the structure of Chap. 3. This device is now very easy to understand, on the basis of the concepts already presented for the two- and three-terminal structures. This is the central chapter of the book. A simple introduction to the energy barrier point of view has been added near the beginning of the chapter. Several models are presented in detail. The first of them is the complete all-region model, including drift and diffusion currents (called “charge-sheet model” in the previous editions, a term we have decided not to use in this edition, as basically all models in the book are based on a charge sheet assumption). Thanks to a simplified derivation, this material is brief but thorough. This is followed by new material on simplified all-region models, including both symmetric and source-referenced versions. New material has been introduced on the symmetric linearization model, which is becoming increasingly important.

These models form the basis for deriving several popular strong- and weak-inversion models, which are covered in detail. Some of this material can be skipped without loss of continuity, and this is indicated at the appropriate points. The various models are extensively related to and/or compared with each other, and the way they can all be derived from one master model (the complete all-region model) is pointed out. Sections on effective mobility (expanded), temperature effects, and source/drain resistances are also included. The tradeoffs between accuracy and simplicity are indicated throughout the chapter.

Chapter 5: Small-Dimension Effects

This chapter has been greatly revised and expanded. In addition to updating the material in the second edition on various classical short-channel and narrow-channel effects, new material has been introduced on drain-induced barrier lowering, impact ionization, velocity overshoot and ballistic operation, polysilicon depletion, quantum effects, gate current, and gate-induced drain leakage. The material on classical scaling has been retained due to its teaching value, but a new, invited section on modern scaling practices has been added, written by Prof. Dimitri Antoniadis of MIT.

Chapter 6: The MOS Transistor in Dynamic Operation—Large-Signal Modeling

This chapter is largely devoted to charge modeling. The concept of quasi-static operation is carefully introduced, and general techniques for charge evaluation are presented, illustrated by charge computations for one representative model. Non-quasi-static analysis is then introduced. Because this chapter was considered by instructors and reviewers to be one of the most successful ones in the two previous editions, its basic structure was retained. The chapter includes material on general charge modeling independent of inversion regions, and on transient response in non-quasi-static operation. New to this edition is the material, in the last section of this chapter, on modeling the extrinsic part of the transistor. In previous editions, this material was in the small-signal chapter, but it was felt that it fits better here.

Chapter 7: Small-Signal Modeling for Low and Medium Frequencies

The principles behind small-signal modeling are presented. The discussion is limited to quasi-static behavior. A useful small-signal model is developed for operation at low and medium frequencies. Expressions for small-signal parameters are derived. An expanded discussion of small-dimension effects is included, as is an expanded discussion of extrinsic capacitance modeling. Major changes in this chapter include a detailed discussion of the effects of gate and body currents, and a small-signal model that includes such effects is presented. Noise is then discussed, including an expanded description of flicker noise and of the effects of small device dimensions on noise.

Chapter 8: High-Frequency Small-Signal Models

In this chapter, two kinds of small-signal models are developed. First, complete quasi-static models are introduced, which differ from the models of Chap. 7 in that they include transcapacitances. The nature of these elements is carefully discussed. Techniques are given for the rigorous development of equivalent-circuit topologies from a complete quasi-static description. Then, non-quasi-static models are introduced through a careful development of the transistor's "transmission line" equations, and a useful y -parameter model is derived for high-frequency applications. It is shown that each level of modeling reduces to the next lower one if the frequency is sufficiently reduced. High-frequency noise, emphasizing induced gate noise, is discussed, as are considerations for radio-frequency modeling, including the effects of gate resistance and a discussion of the two common figures of merit for high-frequency performance, the transition frequency and the maximum frequency of oscillation.

Chapter 9: Substrate Nonuniformity and Structural Effects

This chapter is mostly new. In past editions, vertical substrate nonuniformity was treated after Chap. 4, but instructors found this rather distracting. The most important parts of this material have now been moved here, revised to make it easier for an instructor to cover only the parts that discuss the effects of ion implantation on threshold voltage (the detailed development of other aspects of I - V characteristics has

been moved to an appendix). The rest of the chapter is all new and covers lateral non-uniformity (halo regions), the well proximity effect, stress, and statistical variability. Although all these effects are important, they can best be understood after the basic aspects of MOS transistor operation have been covered; this is the reason this material is presented in this chapter.

Chapter 10: Modeling for Circuit Simulation

The material in this chapter is mostly new in the third edition. Having been exposed to the many phenomena in the MOS transistor and to the modeling of such phenomena, the reader will find in this chapter an exposition of the many issues and considerations involved in putting all these together to make an extensive model suitable for circuit simulation. Discussed here are the various types of models, the ways that models of particular phenomena are combined, parameter extraction, desirable properties for simulator models, and common pitfalls in modeling. Benchmark tests for models are discussed in an associated appendix. This chapter provides a starting point for readers who intend to work in modeling, and should save them time and effort by clearly warning of common errors. It is also meant to provide a background for circuit designers, allowing them to understand the limitations of the models they are using and to better communicate their needs to modeling experts.

The book concludes with 11 appendixes containing an introduction to the basic laws of electrostatics as well as several general but complicated results that, it was felt, would distract if put in the main text. For the same reason, some material in the main text was put in footnotes or, as already mentioned, described in the statements of some homework problems. Such material includes certain fine details or alternative points of view. To avoid distraction, the reader may prefer to skip fine-print footnotes and appendices during a first reading; the main text is self-contained. This material can always be consulted at a later time, because its connection with specific points in the text is obvious.

References to the technical literature were extensively updated and expanded. In most cases, a reference was selected for inclusion because it is technically important, or is widely mentioned in the literature, or has historical significance, or is part of controversy that has not yet been resolved.

The subject of this book is definitely among those that are best digested by doing. The homework problems should help to give a feeling for the kind of work involved in modeling. Most of the homework problems fall in one of the following categories:

1. They sketch additional modeling ideas not in the main text and encourage the reader to try them out.
2. They compare several models introduced in the text.
3. They ask for computations and plots to help provide a quantitative feeling and investigate various properties.
4. They ask for detailed derivations that were sketched in the main text, but that were not shown in detail in order to avoid distraction from the main points.

The numerical homework problems of the previous editions have been updated to use modern fabrication process parameter values. New homework problems have been added to parallel the new material in the various chapters.

Students can be encouraged to write subroutines for the calculation of various quantities as they go along and to save them for later use. In this way, they will gradually build a library of useful subroutines that can be helpful not only in new homework assignments but also in a final project if one is assigned. It has been the authors' experience that project work is invaluable, and the more extensive the project, the greater the benefits. The project can take the form of the implementation of models on the computer. Here it is not enough to just copy a model from the book into a computer program. One must worry about, among other things, choosing the right models, appropriately combining them to form general models, and ensuring continuity of calculated quantities with respect to all given parameters. Some examples of projects follow.

- Write a computer program to evaluate the drain current of a device on a uniform substrate, including short- and narrow-channel effects. The current should be continuous with respect to any input parameter (e.g., voltage, geometric dimensions) and so should be the derivatives of the current with respect to each terminal voltage.
- Develop a computationally efficient technique for the evaluation of drain current in a long-channel device, valid in weak, moderate, and strong inversion. This will necessitate the development of efficient numerical techniques, because the general charge sheet approach, if unmodified, will lead to complex computations. Again, continuity of the current and its derivatives with respect to all parameters should be ensured.
- Develop a program for modeling ion-implanted devices in strong inversion, again ensuring continuity.
- Develop a program for the modeling of low- and medium-frequency small-signal parameters in strong inversion, paying special attention to the small-signal output conductance in the saturation region. All small-signal parameters should be continuous with respect to all input parameters.
- Develop a program for the evaluation of all charges and small-signal capacitances. All these quantities should be continuous with respect to any input parameter.
- Develop a program for the extraction of parameter values to be used with a given model. The input to this program is assumed to consist of measured quantities.

In addition, certain long homework problems can easily be expanded into projects. In all cases, the value of the project, and the challenge in it, can be enhanced if the results are compared with measurements, obtained either in the lab or from the technical literature (the references provided should be very helpful in the latter case). Depending on the magnitude of effort foreseen, students can work separately or in teams.

A new feature of this edition is its accompanying Web site, www.oup.com/tsividis_mcandrew. This includes additional results, including ones that result in very complicated expressions, as well as model code. The site should be useful to professional modeling specialists, to graduate students doing research on the topics of the book, and even to students doing class projects. Errata will be posted on the site.

The first two editions of this book have been used for senior or graduate courses at many universities and in industrial short courses. For a one-semester course, a large number of possibilities exist in regard to the topics selected for coverage. Thus, for example, a course emphasizing general principles would cover Chaps. 1 through 3, Chap. 4 including a careful coverage of the complete all-region model, Chaps. 5 to 8, and selected topics from Chaps. 9 and 10, depending on interest and time available. A course emphasizing practical models for digital circuit design may cover Chaps. 1 through 4, deemphasizing the all-regions model, and parts of Chaps. 5 and 6. For parts that are deemphasized, a quick qualitative coverage is possible, based mostly on the figures. The authors would be happy to consider individual teaching needs and suggest specific course outlines to instructors who contact them.

The authors would like to acknowledge the contributions of many individuals to the shaping of this edition. First, we would like to thank our wives, Felicia Tsividis and Françoise Dastous, for their patience and support. Many thanks are owed to Dimitri Antoniadis for writing the section on modern scaling and for useful comments on other parts of the book; some parts of Chap. 5, which he had helped revise for the previous edition, still bear his influence. Thanks are also due to George Eftihivouidis, Gennady Gildenblat, and Charles Zukowski for their very thorough comments. For their comments during the preparation of either this or the first two editions, the authors would like to also thank Pratul Ajmera, Dimitri Anastassiou, Narain Arora, Mehran Bagheri, Peter Bendix, Alexis Birbas, John Brews, Howard Card, Ana Isabela Cunha, Paul Diament, Aloka K. Dutta, Christian Enz, Graham Flower, Venu Gopinathan, Olin Hartin, Renuka Jindal, Ping-Keung Ko, Andy Lish, Gerson Machado, Ranjit Mand, Guido Masetti, Ed Nicolian, Mehmet Ozturk, Demetrios Papageorgiou, Shanthi Pavan, Narendra Rao, Charles Sodini, Ken Shepard, J. P. Sun, Ken Suyama, Yuan Taur, Renan Türkman, Piet Wambacq, Don Ward, Ed Yang, and the anonymous reviewers. The authors also thank their editor, Rachael Zimmermann, for her enthusiasm and support.

Yannis Tsividis and Colin McAndrew

CHAPTER**1**

Semiconductors, Junctions, and MOSFET Overview

1.1 Introduction

The discussion of MOS devices in this book will be based on an understanding of a few basic concepts. These concepts have been collected in this chapter. We begin with an introduction to semiconductors and the evaluation of mobile carrier concentrations in them. We then consider the mechanisms of current transport in semiconductors. We continue with a discussion of contacts between different materials and the electrostatic potentials established in such contacts. One special contact, the *pn* junction, is then discussed. The material here is meant primarily as a review, but has been written in such a way that it can be understood even with no prior exposure to semiconductor electronics. It has thus been kept as simple as possible. A more advanced and detailed treatment can be found in several textbooks.^{1–12}

The chapter concludes with a brief overview of the MOS transistor and of the chapters that follow.

1.2 Semiconductors

In this section we present a highly simplified review of basic semiconductor concepts. Our discussions are limited to simple descriptive models for what are actually very complex physical phenomena, the proper understanding of which requires a background in quantum mechanics and thermodynamics. Extended discussions of semiconductor concepts can be found in the references.^{1–11}

1.2.1 Intrinsic Semiconductors, Free Electrons, and Holes

Semiconductors derive their name from the fact that they can conduct current better than insulators, but not as well as conductors. The most widely used semiconductor material currently is silicon. The following discussion is focused on this material, but the qualitative arguments used are valid for other semiconductors as well. *Throughout this book we assume that no illumination, no radiation, and no magnetic fields are present, and that all points of the semiconductor are at the same temperature (understood to be room temperature unless indicated otherwise).* Until further notice, we also assume that the semiconductor material under discussion is self-contained, with no externally applied voltage or current, and that the electric field is zero in its environment (the assumption of zero electric field will be relaxed later on in this section). Finally, we will assume that all of these criteria have been satisfied for a long time, so that conditions within the semiconductor have settled. The semiconductor is then said to be in equilibrium.

A pure (intrinsic) silicon crystal consists of an orderly three-dimensional array of atoms spaced apart from each other by roughly 0.5 nm .[†] This array is called the *crystal lattice* and contains approximately $5 \times 10^{22} \text{ atoms/cm}^3$. The atoms of the lattice are held together by cooperating electrons, called *valence electrons*, which form *bonds* between the atoms. At absolute zero temperature, all such electrons are firmly held in place, and the total negative charge of the electrons in each atom is canceled by an opposite positive charge contained in the atom's nucleus. At higher temperatures, the lattice vibrates due to thermal energy; this "thermal motion" manages to set some of the electrons loose from the parent atom. These become *free electrons*, in the sense that they are now free to move about the crystal; the name "free electrons" is used to distinguish them from the rest of the electrons that are still part of the bonds between atoms and are not free to wander away. If the motion of free electrons is coordinated, it can cause the flow of current. Since the atoms from which these electrons broke loose were electrically neutral originally, they are now left with a net positive charge.

Consider now two neighboring atoms, *A* and *B*, and assume that an electron was set free from *A*; now there is an electron vacancy in *A*, so *A* is overall positive. A valence electron associated with atom *B* can move and fill this vacancy, thus creating now a vacancy in *B*. Notice that this electron moved from one bond to another, i.e., it did not become free. The result of this valence electron transfer is that now *A* is neutral, whereas *B* has acquired a net positive charge. A valence electron from an atom *C* near *B* can now move, fill the vacancy in *B*, thus making *B* neutral and *C* positive, etc. We see that this mechanism transports a positive charge from *A* to *B* to *C*.

Thus, we encounter *two* mechanisms for carrying charge around the semiconductor: (1) The motion of *free electrons* about the crystal lattice, each such electron carrying a negative charge, and (2) the motion of valence electrons from bond to bond, corresponding to a motion of "vacancies" (and associated positive charges) in the opposite direction. This second phenomenon can be described by modeling it as a motion of fictitious free particles, called *holes*, which carry a positive charge; each

[†] We give rounded-off, easy-to-remember values, for this and other quantities. For values accurate to many significant digits, the reader can consult the references.^{5,13}

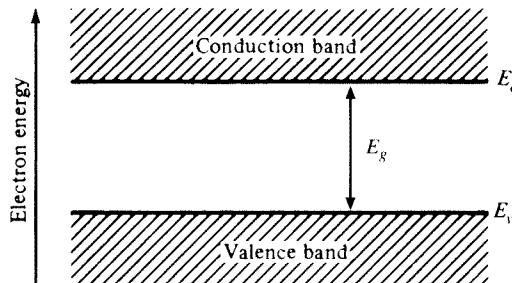


FIGURE 1.1
Energy band model for a semiconductor.

hole can be associated with one vacancy. If the charge of one electron is denoted by $-q$, then the charge of one hole is $+q$. The value of q is

$$q = 1.602 \times 10^{-19} \text{ C} \quad (1.2.1)$$

where C stands for coulomb.

As they wander around the lattice, a hole and an electron can meet and annihilate each other; this is called *recombination*.

Energy Bands The above concepts are usually discussed in solid-state physics using the concept of “energy bands.”^{1–11} We now summarize some of the basic features of this concept with the help of Fig. 1.1. The horizontal axis corresponds to geometrical distance in the semiconductor, whereas the vertical axis corresponds to electron energy. In the *intrinsic* semiconductor we are discussing, electrons bound to their parent atom have energy no larger than E_v ; they are said to “be in the valence band.” An electron with a total energy of at least E_c becomes liberated from the parent atom and is said to “be in the conduction band.” Such an electron leaves “behind” a hole in the valence band. The energy of holes is measured in a direction opposite from that of electrons because of their opposite charge (i.e., hole energy increases downward in Fig. 1.1). If an electron acquires a total energy $E > E_c$ (e.g., because of thermal vibration of the lattice), the difference $E - E_c$ corresponds to net kinetic energy as the electron moves in the crystal lattice. E_c itself represents the potential energy of the free electron. Energy levels between E_v and E_c are not occupied in the intrinsic semiconductor under discussion. Such energies belong to the so-called forbidden band gap, or simply “band gap,” which for silicon has a width $E_g = E_c - E_v$ of 1.12 eV at 300 K, decreasing somewhat with temperature.[†] In an intrinsic semiconductor at room temperature, few electrons in the valence band acquire as much energy as E_g and, therefore, few are liberated. In insulators, the energy gap is even larger, and thus there are practically no electrons available for conduction. In conductors, on the other hand, there is a multitude of such electrons.

In the pure semiconductor we are discussing, since each hole is created by the breaking loose of one electron that becomes free, there is an equal number of holes

[†] An “eV” is the product of the magnitude of the electron charge and 1 V: 1 eV = 1.602×10^{-19} J.

and free electrons. Thus, let the volume concentration of electrons and holes in the intrinsic material be denoted by n_i and p_i , respectively; we will have

$$p_i = n_i \quad (1.2.2)$$

The symbol n_i is often used to denote either concentration and is referred to simply as the *intrinsic carrier concentration*. Its value corresponds to the equilibrium case where the rates of generation and recombination of electron-hole pairs are equal. At room temperature, usually taken as an absolute temperature of 300 K (27°C),[†] the value⁵ of n_i in silicon is approximately 10^{10} per cm^3 (written cm^{-3}) (other values are also in use). This means that roughly 10 electron-hole pairs are to be found in a cube 10 μm on a side. Since there are 5×10^{22} silicon atoms/ cm^3 , only about two out of every 10^{13} of these atoms contribute one electron-hole pair! As might be expected from our previous discussion, n_i increases strongly at higher temperatures.[‡] For example, as temperature is raised from 27°C to 100°C , n_i increases 100 times!

1.2.2 Extrinsic Semiconductors

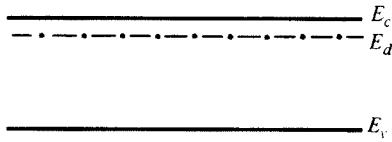
It is possible to make the number of free electrons different from that of the holes by introducing foreign atoms in the silicon crystal. Such atoms are called *impurities*, and the process of introducing them is called *doping*; doped semiconductors are referred to as *extrinsic semiconductors*. If it is desired to enhance the free-electron population, these foreign *dopant* atoms are chosen so as to have available for bonding one electron *more* than the number needed for perfect bonding in the silicon structure. Thus, these atoms form bonds with neighboring silicon atoms by using up all their valence electrons except one. The latter is very loosely held to the parent atom, and at room temperature the thermal vibration of the crystal lattice is enough to set it free. In the energy band model, the loosely held electron corresponds to an energy level E_d only slightly below the conduction band edge E_c , as shown in Fig. 1.2;[§] at room temperature, practically all such electrons (one per impurity atom) have enough thermally acquired energy to enter the conduction band (i.e., to be set free).

In contrast to the case of the intrinsic (pure) semiconductor, the departure of the above electron leaves all valence bonds intact; hence it does not cause a vacancy in them, and thus it does not leave behind a hole. However, since the dopant atom was originally neutral, it is now left with a net positive charge and is said to be *uncovered* or *ionized*. What we have described is shown in Fig. 1.3. Each – sign represents one free electron. Each *circled +* sign represents a dopant atom that has lost one electron, and thus is left with a net positive charge. The circle is used to indicate

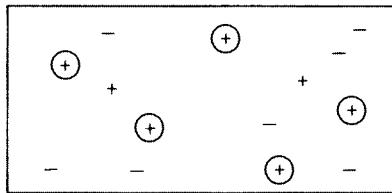
[†] Temperature in degrees Celsius is found from absolute temperature in kelvins by subtracting 273.15 degrees.

[‡] An approximate formula for n_i in silicon as a function of the absolute temperature T is $n_i = A_1 T^{3/2} \exp(-A_2/T)$, where $A_1 = 7 \times 10^{15} \text{ K}^{-3/2} \text{ cm}^{-3}$ and $A_2 = 6600 \text{ K}$.

[§] The difference $E_c - E_d$ is known as the “ionization energy.”

**FIGURE 1.2**

Energy band diagram for an *n*-type extrinsic semiconductor. E_d is the energy level corresponding to donor atoms.

**FIGURE 1.3**

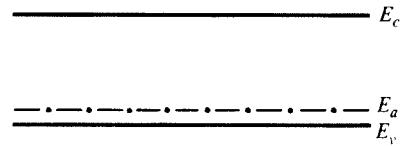
Free electrons, holes, and ionized dopant atoms in *n*-type extrinsic silicon.

that this atom, being part of the crystal lattice, is *immobile* and therefore is not itself available for conduction. At room temperature, practically all dopant atoms are ionized, and thus the number of free electrons created by such ionization is practically equal to the number of these atoms. As already mentioned, each such atom contributes one free electron without creating a hole. However, a few hole-electron pairs are created by the silicon atoms due to mechanisms already described for the intrinsic (pure) case; two such pairs are included in Fig. 1.3 (a hole is represented by an uncircled + sign).

Let us denote the concentration of free electrons by n_o and that of holes by p_o . With so many free electrons moving around the lattice, the chance of their encountering a hole and filling it, or “recombining” with it, is significant; thus p_o decreases compared with its value for the intrinsic case. In fact, if the semiconductor is not extremely highly doped (see below), this chance of recombination is approximately proportional to n_o ; thus p_o decreases by the same factor that n_o has increased (compared with the intrinsic case). Hence, the product $n_o p_o$ remains the same as in the intrinsic case,¹ i.e., it is equal to n_i^2 as seen from (1.2.2). Thus

$$n_o p_o = n_i^2 \quad (1.2.3)$$

Since the dopant atoms in Fig. 1.3 were chosen such as to “donate” one free electron to the silicon crystal, they are called *donors*. Donor materials commonly used to dope silicon are phosphorus, arsenic, and antimony. The donor atoms are introduced into the silicon in minute amounts, but their concentration is usually several orders of magnitude higher than n_i . For example, although a donor atom concentration of 10^{17} cm^{-3} corresponds to two donor atoms in every million silicon atoms, this concentration is ten million times greater than the intrinsic carrier concentration at room temperature. The donor concentration will be assumed uniform unless stated otherwise. At common operating temperatures, practically all donor atoms are ionized, and thus the number of free electrons they contribute is much higher than those contributed by the “intrinsic” mechanism discussed

**FIGURE 1.4**

Energy band diagram for a *p*-type extrinsic semiconductor. E_a is the energy level corresponding to impurity atoms.

previously; hence, the concentration of free electrons, denoted by n_o , is approximately equal to the donor concentration, denoted by N_D :

$$n_o \approx N_D \quad (1.2.4)$$

From this fact and (1.2.3) it follows that

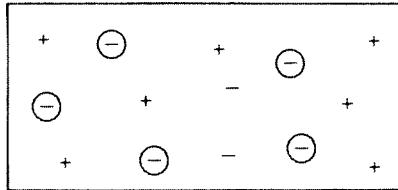
$$p_o \approx \frac{n_i^2}{N_D} \quad (1.2.5)$$

If the doping concentration is very high (higher than about 10^{19} cm $^{-3}$), the accuracy of these relations decreases.^{1,5,10} Semiconductors with very high doping concentration are said to be *degenerate* and they require special treatment.^{‡1,5,10} Also, the preceding two relations will not hold at very low temperatures, where the dopant atoms will not all be ionized, or at very high temperatures, where n_i rises to the point that the assumption $N_D \gg n_i$ is not valid. *Whenever the preceding relations are used in this book, it will be implied that none of these extreme situations is in effect.*

Because in a donor-doped semiconductor n is larger than p , the free electrons are called the *majority carriers* and the holes the *minority carriers*. A semiconductor doped with donor impurities is said to be *n* type, because the majority carriers in it carry a *negative* charge.

Instead of increasing the mobile electron population as described previously, it is possible to increase the *hole* population by introducing into pure silicon impurity atoms that have one valence electron *less* than the number needed for complete bonding with neighboring silicon atoms creating a vacancy there. Thus, when each such atom attempts to form bonds, it will be short one valence electron. It can then “steal” such an electron from a nearby silicon atom. In the energy band model, the vacancy corresponds to an energy level E_a only slightly above the valence band edge E_v (Fig. 1.4), and thus it is easy for an electron from the valence band to fill the vacancy and leave “behind” a hole. This “stealing” of an electron from a neighboring silicon atom has two effects. First, since the impurity atom was electrically neutral originally, now that it has acquired an extra electron, it will have a net negative charge; this charge is associated with one specific atom and is thus immobile in the sense that it is not available for current conduction. Second, this stealing away one electron from a neighboring silicon atom

[†] One of the effects encountered in heavily doped semiconductors is that the energy band gap, E_g , decreases. This is referred to as “band gap narrowing.”⁵

**FIGURE 1.5**

Free electrons, holes, and ionized dopant atoms in *p*-type extrinsic silicon.

left a valence electron vacancy in the latter, and thus created a hole; this hole can move around as in the case of the intrinsic (pure) crystal. Notice, however, that in contrast to the intrinsic case, the hole was created *without* simultaneously creating a free electron.

Since the impurity atoms have stolen or “accepted” one valence electron from the silicon lattice, they are called *acceptors*; typical acceptor materials used to dope silicon are boron, gallium, and indium. Figure 1.5 illustrates the charges in a semiconductor doped with acceptor atoms. The + signs denote holes, the circled – signs denote ionized acceptor atoms (which are immobile), and the uncircled – signs denote free electrons (as in the case of donor-doped silicon, a few hole-free electron pairs are still created by silicon atoms due to the mechanisms described for the intrinsic case). The total charge in Fig. 1.5 adds up to zero, indicating that the semiconductor is macroscopically neutral. Let the acceptor concentration (assumed uniform) be denoted by N_A and assume that $N_A \gg n_i$ (typical values of N_A are 10^{17} to 10^{18} cm^{-3}). Assuming that practically all acceptor atoms are ionized, we will have, since each atom contributes one hole,

$$p_o \approx N_A \quad (1.2.6)$$

As in the case of donor-doped material, (1.2.3) holds, and we have

$$n_o \approx \frac{n_i^2}{N_A} \quad (1.2.7)$$

These approximations will fail at extremely low or high temperatures, or if the doping concentration is extremely high, as explained for the case of donor doping. In an acceptor-doped semiconductor, the holes are the “majority carriers,” and the electrons are the minority carriers. Since majority carriers carry a *positive* charge, semiconductors doped with acceptor impurities are said to be *p type*.

1.2.3 Equilibrium in the Absence of Electric Field

A detailed description of carrier statistics in semiconductors is beyond the scope of this book, but we will summarize here some important results. Although electrons and holes are governed by Fermi-Dirac statistics, these reduce to Maxwell-Boltzmann statistics for nondegenerate semiconductors.¹⁻¹¹ According to Maxwell-Boltzmann

8 OPERATION AND MODELING OF THE MOS TRANSISTOR

statistics, the electron and hole concentrations in an electrically neutral semiconductor in equilibrium can be expressed as follows:

$$p_o = n_i e^{(E_i - E_F)/kT} \quad (1.2.8a)$$

$$n_o = n_i e^{(E_F - E_i)/kT} \quad (1.2.8b)$$

In these equations, n_i is the intrinsic carrier concentration; E_F is a quantity called the “Fermi energy,” which is characteristic of the material;[†] E_i is the “intrinsic energy level” (located very close to the middle of the forbidden energy gap); T is the absolute temperature (measured in kelvins, K); and k is the Boltzmann constant, with value

$$k = 1.3807 \times 10^{-23} \frac{\text{C} \cdot \text{V}}{\text{K}} = 8.62 \times 10^{-5} \frac{\text{eV}}{\text{K}} \quad (1.2.9)$$

The derivation of these equations from first principles is quite complicated; proper understanding of the concepts involved requires quantum mechanics and thermodynamics. Simplified discussions are provided in the references.¹⁻¹¹

For intrinsic semiconductors $E_F = E_i$ (Fig. 1.6a), and the preceding three equations give $n_o = p_o = n_i$ as expected. For n -type extrinsic semiconductors $E_F > E_i$ (Fig. 1.6b), and for p -type extrinsic semiconductors $E_F < E_i$ (Fig. 1.6c). If E_F approaches E_v or E_c within a few kT , the material approaches degenerate behavior and the accuracy of these equations decreases. Yet these equations are often used even for somewhat degenerate materials. From (1.2.8), it can be verified that $n_o p_o = n_i^2$, as expected from (1.2.3).

It is convenient to define a quantity ϕ_F as follows:[‡]

$$\phi_F \equiv \frac{E_i - E_F}{q}, \quad \text{neutral semiconductor}$$

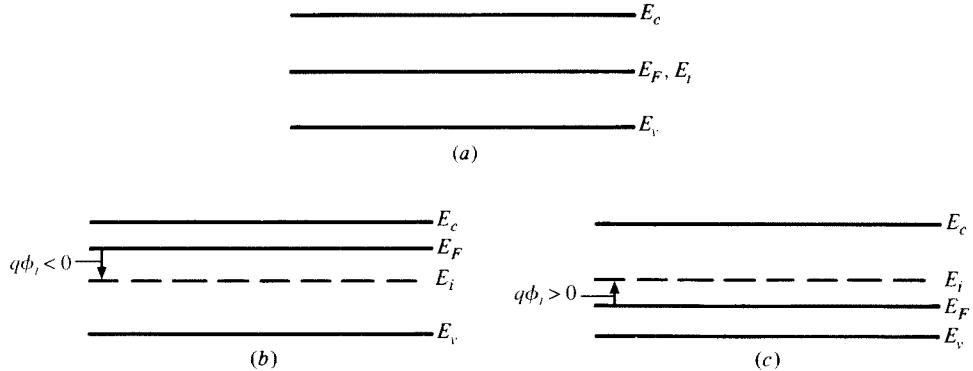
(1.2.10)

The meaning of ϕ_F is indicated in Figs. 1.6b and 1.6c. Once the carrier concentrations are known, the value of ϕ_F can be calculated (always assuming that the material is nondegenerate). Using the definition of ϕ_F in (1.2.10), (1.2.8) can be written as

$$p_o = n_i e^{\phi_F/kT} \quad (1.2.11a)$$

[†] The Fermi energy is a quantity involved in Fermi-Dirac statistics, which give the probability of an electron’s being at a given energy level, if that level is allowed to be occupied. Metals are also described by Fermi-Dirac statistics and thus have a Fermi energy. However, the simplified carrier concentration equations we have given for semiconductors are not valid for metals.

[‡] This definition of ϕ_F is not universal. In some treatments, the negative of our quantity is used. The quantity ϕ_F is sometimes called the “Fermi potential,” but in some treatments this name is reserved for an absolute potential, as opposed to one relative to E_i as in our definition.

**FIGURE 1.6**

Relative position of intrinsic energy level (E_i) and Fermi energy (E_F) for (a) intrinsic, (b) n -type, and (c) p -type semiconductors.

$$n_o = n_i e^{-\phi_t / \phi} \quad (1.2.11b)$$

where we have conveniently defined a quantity ϕ_t as

$$\phi_t = \frac{kT}{q}$$

(1.2.12)

This quantity is the ubiquitous “thermal voltage.” From (1.2.1) and (1.2.9), the value of ϕ_t at room temperature (300 K) is 0.0259 V. From (1.2.11) we obtain

$$\phi_F = \phi_t \ln \frac{p_o}{n_i} \quad (1.2.13a)$$

$$= -\phi_t \ln \frac{n_o}{n_i} \quad (1.2.13b)$$

Finally, if practically all impurity atoms are ionized, we can use (1.2.6) and (1.2.4) in (1.2.13) to obtain:

$$\phi_F \approx +\phi_t \ln \frac{N_A}{n_i} \quad p\text{-type material} \quad (1.2.14a)$$

$$\phi_F \approx -\phi_t \ln \frac{N_D}{n_i} \quad n\text{-type material} \quad (1.2.14b)$$

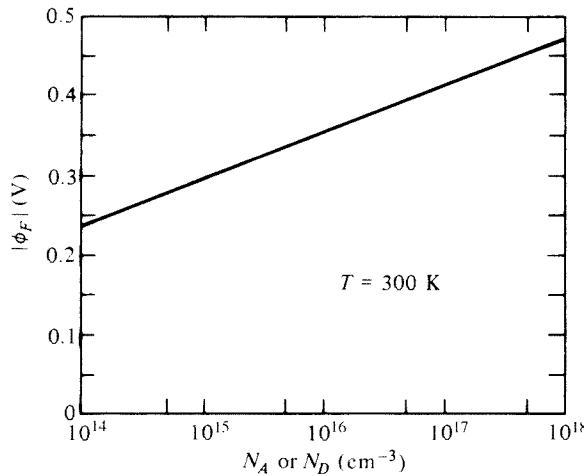


FIGURE 1.7
Magnitude of Fermi potential vs. substrate doping concentration for silicon at room temperature.

Thus, ϕ_F is a quantity that characterizes a semiconductor material at a given temperature. For example, one can talk of a certain *p*-type silicon with $N_A = 10^{17} \text{ cm}^{-3}$ or, equivalently, of *p*-type silicon with $\phi_F = 0.417 \text{ V}$ at 300 K. Values for ϕ_F obtained from these equations at 300 K are plotted vs. doping concentration in Fig. 1.7.

If the doping concentration is too high, E_F is too close to the conduction or valence band and the semiconductor is degenerate. Then the above equations do not hold. For heavily doped silicon, an approximation often used is that $E_F \approx E_v$ (*p* type), or $E_F \approx E_c$ (*n* type). We then have $|E_F - E_i| \approx E_g/2$, where E_g is the band gap energy, roughly taken equal to the nondegenerate value of 1.12 eV at room temperature. For silicon at room temperature, this means that ϕ_F from (1.2.10) is approximately 0.56 V for *p* type and -0.56 V for *n* type. The same approximations are used for degenerate polycrystalline silicon (often referred to simply as polysilicon or poly), a widely used material in modern transistors (Sec. 1.6).

1.2.4 Equilibrium in the Presence of Electric Field

Our discussion so far has assumed zero electric field.[†] If the electric field within the semiconductor is not zero, the free electron and hole concentrations, n and p , will in general be different from n_o and p_o . However, their product is *still equal to* n_o^2 , provided that the semiconductor remains in equilibrium.^{1.5,11} We will encounter such equilibrium cases, which will be characterized by the lack of energy exchange between the semiconductor and the external world and no net current flow; for example, this will be the case for the two-terminal MOS structure

[†] The term *electric field* is reserved here for *macroscopic* electric fields. Fields will always be present at the atomic level, such as between electrons and the nucleus in a given atom.

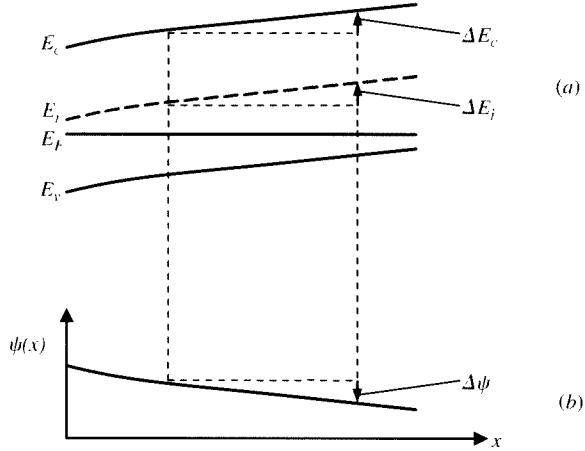


FIGURE 1.8
Semiconductor in equilibrium in the presence of electric field; (a) energy band diagram; (b) potential.

discussed in Chap. 2. For nondegenerate semiconductors, in analogy with (1.2.8), we will have

$$p = n_i e^{(E_i - E_F)/(kT)} \quad (1.2.15a)$$

$$n = n_i e^{(E_F - E_i)/(kT)} \quad (1.2.15b)$$

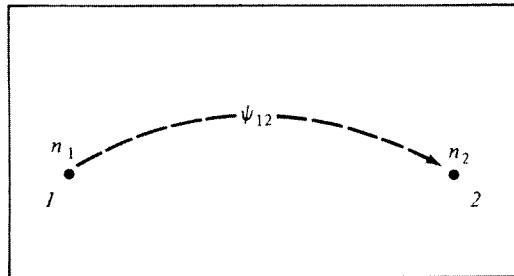
but the difference $E_i - E_F$ can be different from that in the absence of electric fields (see below). From these relations we have

$$np = n_i^2 \quad (1.2.16)$$

In equilibrium, E_F is a single constant throughout a semiconductor. A rough analogy is the water level in a bathtub at rest. If this water level is momentarily disturbed at a point, water movement will occur until the level is restored. Likewise, a momentary disturbance in a semiconductor can change things locally, but carriers will then flow until equilibrium is reached and a constant Fermi level is established.

For a semiconductor in equilibrium, with a constant E_F throughout, the presence of electric field can affect the carrier concentrations. For (1.2.15) to remain valid, E_i has to change, and E_c and E_v will change along with it. We illustrate this with the help of the one-dimensional case shown in Fig. 1.8. (This figure is somewhat arbitrary; we will encounter specific cases later in this book.) The electric field is the negative of the gradient of the electrostatic potential (Appendix A); in turn, electrostatic potential difference is potential energy difference per unit charge. Thus, the presence of electric field implies a variation of electrostatic potential with position and a corresponding variation of potential energy with position. Recall that the potential energy for electrons is given by E_c . Let $\Delta E_c = \Delta E_i$ be the potential energy difference between two points for an electron (which has charge $-q$). The electrostatic potential between these points, $\Delta\psi$, will be

$$\Delta\psi = \frac{\Delta E_c}{-q} \quad (1.2.17)$$

**FIGURE 1.9**

A piece of semiconductor material with different electron concentration at two points.

Thus, to accommodate the constant Fermi level in the presence of electric field, the levels E_c , E_i , and E_v “bend” accordingly. The difference between any two among E_c , E_i , and E_v remains constant with distance. For any point along the horizontal axis in Fig. 1.8, the distance between E_F and E_i must be such as to give the correct values of n and p when (1.2.15) are used. The actual values of n and p can be found from further considerations, which are discussed later.

Consider a region of semiconductor material (intrinsic, n type, or p type) in equilibrium and assume that an electrostatic potential difference ψ_{12} exists between two points 1 and 2, as shown in Fig. 1.9.[†] Using (1.2.15b), the electron concentrations at the two points will be $n_1 = n_i \exp[(E_F - E_{i1})/kT]$ and $n_2 = n_i \exp[(E_F - E_{i2})/kT]$. Dividing these and using (1.2.17) and (1.2.12) gives

$$\frac{n_1}{n_2} = e^{\psi_{12}/\phi_i} \quad (1.2.18)$$

The relation corresponding to (1.2.18) for holes (again, assuming equilibrium) can be obtained in a similar manner starting from (1.2.15a) and is

$$\frac{p_1}{p_2} = e^{\psi_{21}/\phi_i} \quad (1.2.19)$$

Note that (1.2.18) and (1.2.19) together imply $n_1 p_1 = n_2 p_2$, just as would be the case in the absence of any potential difference in equilibrium.

1.2.5 Nonequilibrium; Quasi-Fermi Levels

If there is energy exchange between the semiconductor and the external world (e.g., when a voltage source is connected across a pn junction), the semiconductor is not in equilibrium in general, and it cannot be characterized by a single, constant Fermi

[†] In this book, the potential of a point with respect to another will be indicated by an arrow pointing *from* the former point *to* the latter. If the potential is internal to a device, the arrow will be drawn with a broken line. Potentials between a device's external terminals will be indicated by solid-line arrows.

energy level. Thus, p and n cannot be predicted by (1.2.15). One could attempt to use a modified Fermi energy level that would be allowed to vary with position, but this would not be enough, as the values needed for such a quantity turn out to be different for electrons and for holes in nonequilibrium. Thus, it is convenient to introduce two quantities E_{Fp} and E_{Fn} , called *quasi-Fermi energy levels* (or *imrefs*, with “imref” standing for “*imaginary reference*” and for “Fermi” spelled backward) for holes and electrons, respectively. These are often defined as the quantities that satisfy the relations¹⁻¹¹

$$p = n_i e^{(E_i - E_{Fp})/(kT)} \quad (1.2.20a)$$

$$n = n_i e^{(E_{Fn} - E_i)/(kT)} \quad (1.2.20b)$$

where p and n are the correct hole and electron concentration values, which can be determined from additional considerations (we will see some examples later on). As is obvious by comparing these equations to (1.2.15), in equilibrium we can set $E_{Fn} = E_{Fp} = E_F$, and thus $np = n_i^2$. In nonequilibrium, $E_{Fn} \neq E_{Fp}$ and thus $np \neq n_i^2$. This situation is encountered, for example, in the depletion region of a *pn* junction with external bias applied, as will be seen in Sec. 1.5.

1.2.6 Relations between Charge Density, Electric Field, and Potential; Poisson’s Equation

Let us now consider the charge density (charge concentration per unit volume) in a semiconductor. In the general case, four entities can be responsible for its value: (1) holes, which contribute a charge density of $(+q)p$; (2) free electrons, with contribution $(-q)n$; (3) ionized donor atoms, with contribution $(+q)N_D$; and (4) ionized acceptor atoms, with contribution $(-q)N_A$.[†] The total charge density, denoted by ρ , will be the sum of the individual contributions:

$$\boxed{\rho = q(p - n + N_D - N_A)} \quad (1.2.21)$$

In the cases already discussed, we will have $N_A = 0$ if only donor atoms are present, or $N_D = 0$ if only acceptors are present. There are cases, however, where both N_D and N_A can be nonzero. For example, if it is desired to convert part of a *p*-type region into an *n*-type region, one can introduce donor atoms at a concentration higher than the concentration of acceptor atoms. The region then becomes effectively *n* type, with an “effective” donor concentration of $N_D - N_A$. This process is encountered often in the fabrication of semiconductor devices.

When parts that are macroscopically electrically neutral are brought together to form a semiconductor device, communication between the parts is established and charge can redistribute itself among them. However, charge cannot be created or

[†] If not all the impurity atoms are ionized (e.g., at very low temperatures), N_A and N_D should be replaced with the corresponding densities of ionized atoms, often denoted by N_A^- and N_D^+ ; the calculation of these is discussed in the references.^{5,10,11}

disappear; overall charge conservation must be maintained. Thus, the structure must remain electrically neutral overall, meaning that the algebraic sum of all charges in the structure remains equal to zero; this is the *principle of charge neutrality*.

While the situation just described deals with overall, macroscopic neutrality, there can also be regions within a semiconductor device that are locally neutral. In particular, in a neutral region of a semiconductor at equilibrium, with hole and electron concentrations p_o and n_o , setting (1.2.21) equal to zero gives

$$p_o - n_o = N_A - N_D \quad (1.2.22)$$

Also, from (1.2.3) we have

$$n_o p_o = n_i^2 \quad (1.2.23)$$

The exact values of p_o and n_o can be calculated by solving the system consisting of these two equations, but in most cases simpler approximations are fine. Thus, in the case of *p*-type semiconductors ($N_D = 0$) with $N_A \gg n_i$, the results of such a solution are practically the same as (1.2.6) and (1.2.7).[†]

In the presence of electric fields, the charge density ρ can vary from point to point. We will assume that this density is independent of time. Assuming equilibrium, n and p in (1.2.21) must be such that, for any two points, they are related to the electrostatic potential by (1.2.18) and (1.2.19), which are consequences of semiconductor properties. In addition to these relations, the *total* charge density ρ must satisfy the equations of electrostatics, which are general and are not restricted to semiconductors. Let us consider the “one-dimensional case” where ρ and the electrostatic potential ψ (taken with respect to some arbitrary reference) vary only vertically (along the y direction).[‡] Then the electric field, a vector quantity, can only have a vertical component. It can thus be represented by its signed magnitude $\mathcal{E}(y)$, having a positive value if the field is pointed in the positive y direction and a negative value in the opposite case. The rate of change of the electric field $\mathcal{E}(y)$ with respect to position y is related to ρ by Gauss’s law:

$$\frac{d\mathcal{E}}{dy} = \frac{\rho(y)}{\epsilon_s} \quad (1.2.24)$$

where ϵ_s is the “permittivity” of the material, given by

$$\epsilon_s = k_s \epsilon_0 \quad (1.2.25)$$

with ϵ_0 the permittivity of free space (8.854×10^{-14} F/cm) and k_s the dielectric constant of the material. For silicon, we have approximately $k_s = 11.9$, corresponding

[†] If for example $N_A = 10^{17}$ cm⁻³, those equations give $p_o \approx 10^{17}$ cm⁻³ and $n_o \approx 10^3$ cm⁻³ at room temperature; while these numbers do not make the two sides of (1.2.22) *exactly* equal, they do make them equal within a relative error of 10^{-14} !

[‡] This direction is chosen in anticipation of the direction used in analyses to follow later.

to $\epsilon_s = 1.05 \times 10^{-12} \text{ F/cm}$. By integrating both sides of (1.2.24) from an arbitrary point y_0 to a point y , we obtain

$$\mathcal{E}(y) = \mathcal{E}(y_0) + \frac{1}{\epsilon_s} \int_{y_0}^y \rho(\hat{y}) d\hat{y} \quad (1.2.26)$$

where \hat{y} is a dummy variable of integration. Thus, one can plot the field vs. distance from a plot of charge density; we will have the opportunity to do this for both the *pn* junction and the MOS structure. This equation assumes that a single permittivity value holds in the entire region between y_0 and y ; if the permittivity changes at a point, a jump in the field is observed there (Appendix A).

The electric field $\mathcal{E}(y)$, defined in the positive y direction, is related to the potential, $\psi(y)$, by

$$\mathcal{E}(y) = -\frac{d\psi}{dy} \quad (1.2.27)$$

Integrating this equation from y_0 to y , we obtain

$$\psi(y) = \psi(y_0) - \int_{y_0}^y \mathcal{E}(\hat{y}) d\hat{y} \quad (1.2.28)$$

The above equations allow one to go from knowledge of the charge density to the electric field and from that to the potential. If one is not interested in the electric field, this quantity can be eliminated among these equations. Thus, substituting (1.2.27) in (1.2.24) gives us what is known as *Poisson's equation*:

$$\frac{d^2\psi}{dy^2} = -\frac{\rho(y)}{\epsilon_s} \quad (1.2.29)$$

Our analysis of semiconductor devices will be based on the simple relations we have provided in this section, assuming that the charge density is independent of time. This is the case when charges are immobile, or when they move in such a way that their charge density and the resulting current is independent of time. The analysis will also be valid approximately if a time dependence is involved, as long as it is sufficiently slow; otherwise, the complete Maxwell's equations must be used.

1.3 Conduction

1.3.1 Transit Time

The notion of “transit time” will be used in our discussion of MOS transistors in subsequent chapters. This notion is general and independent of the mechanisms via which current conduction takes place. Thus, it is introduced first before such mechanisms are discussed.

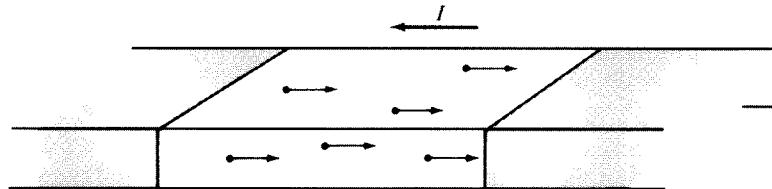


FIGURE 1.10
A piece of material with electrons flowing toward the right.

Consider the piece of material shown between the shaded regions in Fig. 1.10. Assume a steady flow of current in one direction, wholly due to free electrons, has been established. Also, assume that *no* recombination of electrons with holes takes place. Electrons are constantly being supplied from the left side and taken out from the right side at the same fixed rate. At any given instant, then, the magnitude of the total electron charge that happens to be inside the piece is fixed; let $|Q|$ denote its value. We will make the simplifying assumption that it takes each electron the same amount of time to travel the length of the piece; this time will be called the *transit time* and will be denoted by τ .

Let us consider the free electrons found inside the piece at a given instant and follow their motion. After the lapse of one transit time τ , even those initially at the far left will have exited through the right end. Therefore all the electrons inside the piece at the initial instant of observation will have exited through the right-hand end and will have been replaced by new electrons coming from the left. Thus, a *negative* charge of magnitude $|Q|$ exits from the right-hand end in time τ ; this corresponds to a current I in the direction shown, given by

$$I = \frac{|Q|}{\tau} \quad (1.3.1)$$

Note that in the argument leading to (1.3.1) we did not make any assumption as to the detailed mechanisms through which the current is conducted, the distribution of charge inside the material, the presence or absence of electric fields, or the constancy of electron velocity along the current path. Thus, (1.3.1) is quite general.

In the preceding discussion, we made the convenient assumption that all electrons spend the same amount of time τ in traveling the length of the material. However, it is customary to relax this assumption and use

$$\tau = \frac{|Q|}{I}$$

(1.3.2)

as the *definition* of transit time, with $|Q|$ and I as defined above. An obviously similar definition can be given in the case of conduction caused by holes. Expressions for τ will be found for each of two mechanisms of current conduction, which are considered in the next two subsections.

1.3.2 Drift

Let us consider a piece of semiconductor with no external field applied to it. A random “thermal” motion is exhibited by the holes and electrons in all directions; however, on the average, these random charge movements cancel out and there is no net current produced. If an electric field is now applied (for example, by connecting the semiconductor piece across the terminals of a battery), it will exert forces on the charged particles. Thus there will be a net movement along the field lines that can be observed macroscopically as an electric current. This phenomenon is known as *drift*; it would not occur if the particles were not charged.

The movement of electrons and holes during drift is quite complicated since these carriers interact with their environment, for example, with the thermal vibrations of the lattice and with ionized impurity atoms.[†] Such interactions are referred to as *scattering*. For example, as an electron accelerates in the direction dictated by the electric field, it can be scattered and lose energy to its environment, then it accelerates again, etc. Because of this, and because charge is transported by these carriers in discrete amounts, a minute “noise” fluctuation exists in the externally observed current. For the present, such fluctuations will be ignored and attention will be focused on the average current value I , which is nonzero because of the net movement caused by the nonzero electric field. The quantity I can be calculated from the carrier’s *average velocity*, called *drift velocity* and denoted by v_d . For a given electric field, v_d depends on the semiconductor material, the type and concentration of doping, the temperature, and the type of carrier.⁵ For silicon, the magnitude of the drift velocity depends on the magnitude of the electric field \mathcal{E} as shown qualitatively in Fig. 1.11. For high electric fields, the loss of energy of the carriers to the lattice becomes more effective and eventually *velocity saturation* is reached as shown. The maximum velocity value is essentially independent of doping concentration, and for silicon at room temperature is of the order of 10^7 cm/s for both electrons and holes. It is reached at fields roughly above 3×10^4 V/cm for electrons and 10^5 V/cm for holes. Thus, for example, across a length of 100 nm, velocity saturation would be reached with a voltage of roughly 0.3 V and 1 V for electrons and holes, respectively.

We now consider a uniformly doped n -type semiconductor bar with dimensions a, b, c as shown in Fig. 1.12, and a voltage $V > 0$ applied across it. The voltage creates a positive electric field in the direction shown. In the following we will neglect the contribution of hole movement to the electric current since the holes, being the minority carriers in this case, are much fewer than the electrons. Electrons move in a complicated fashion with an *average velocity* of v_d toward the right, as discussed previously. We can calculate the resulting current by considering a simpler, hypothetical picture in which the velocity of all electrons is constant and equal to v_d . Note that with $V > 0$, the quantities \mathcal{E}, v_d , and I , defined in the directions shown in the figure, are positive. The time it takes for an electron to travel the length of the bar is

$$\tau = \frac{a}{v_d} \quad (1.3.3)$$

[†] Our present discussion deals only with drift in the *bulk* of a semiconductor. Additional scattering mechanisms are present at the semiconductor surface in an MOS transistor; these are considered in Chap. 4.

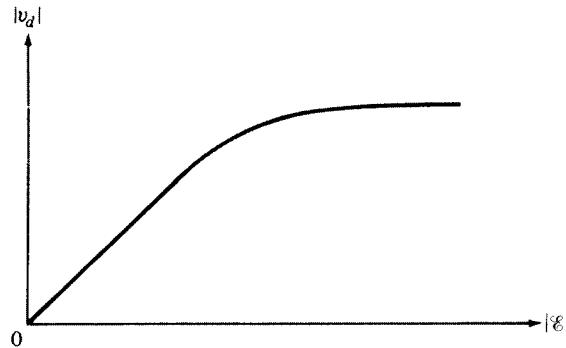


FIGURE 1.11
Magnitude of drift velocity versus magnitude of electric field.

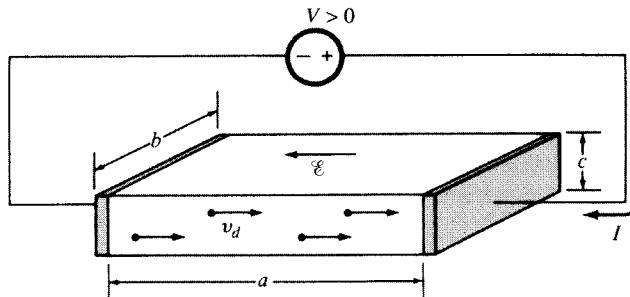


FIGURE 1.12
An *n*-type semiconductor bar with uniform electron concentration under external bias.

The magnitude $|Q|$ of the total free electron charge found inside the bar at a given time instant is given by the charge magnitude of a single electron q times the total number of electrons; the latter is given by the volume of the bar times the free electron concentration per unit volume, n . Thus

$$|Q| = nq(abc) \quad (1.3.4)$$

Using this with (1.3.1) and (1.3.3) gives

$$I = \frac{nq(abc)}{\tau} \quad (1.3.5a)$$

$$= nq(bc)v_d \quad (1.3.5b)$$

This will now be put in a different form that will prove useful later in the discussion of transistors. The area of the bar's top surface in Fig. 1.12 is (ab) . Thus $|Q|/(ab)$ is the *magnitude of the charge per unit area*.[†] This quantity will be denoted by $|Q'|$. Using (1.3.4) we have

[†] The use of this quotient is adequate because the charge is uniformly distributed horizontally; otherwise, the charge per unit area would have to be defined incrementally.

$$|Q'| = \frac{|Q|}{ab} = nqc \quad (1.3.6)$$

Substituting this in (1.3.5b) gives

$$I = b|Q'|\nu_d \quad (1.3.7)$$

The Case of Low Electric Fields The preceding relations assume special, very useful forms in the case of low electric fields. For silicon, “low” means roughly less than 3×10^3 V/cm (i.e., 0.3 V/ μm) for electrons, and less than 6×10^3 V/cm (i.e., 0.6 V/ μm) for holes. For such fields, ν_d is proportional to \mathcal{E} , as suggested by the bottom part of the curve in Fig. 1.11. The constant of proportionality is called *mobility* and is denoted by μ_B . The subscript *B* is used to emphasize that this mobility characterizes the “bulk” of the semiconductor. This is needed to distinguish it from a “surface” mobility that will be introduced in Chap. 4. We thus have

$$\nu_d = \mu_B \mathcal{E} \quad (1.3.8)$$

Electron and hole mobilities for silicon at room temperature are shown in Fig. 1.13 versus doping concentration.¹⁴ Mobilities decrease with temperature under common conditions.^{5,14}

The electric field is independent of position in Fig. 1.12, and is given by

$$\mathcal{E} = \frac{V}{a} \quad (1.3.9)$$

Using this in (1.3.8) and the result in (1.3.7), we have

$$I = \mu_B |Q'| \frac{b}{a} V \quad \dagger, \ddagger \quad (1.3.10)$$

[†] In some treatments, the current is defined from left to right, which is the opposite direction from what we have assumed, in which case a minus sign appears in front of the right-hand side of (1.3.10).

[‡] We have assumed a uniform n throughout the bar. However, we will encounter important cases where this assumption is not valid. If n varies along the x axis, one must instead consider a thin vertical chunk of length Δx centered at x . Let $\Delta\psi$ be the electrostatic potential drop across the chunk, defined from its right side to its left side; letting Δx approach zero, we obtain an equation in which the role of V/a in (1.3.10) is played by $d\psi/dx$:

$$I = \mu_B b |Q'| \frac{d\psi}{dx} \quad (1.3.10')$$

Note that this equation could also have been obtained directly from (1.3.7), by using in it (1.3.8) and expressing the magnitude of the electric field, in analogy with (1.2.27), as $d\psi/dx$ [note that, unlike the case in (1.2.27), here no minus sign is used, as the direction we assume for the field in Fig. 1.12 is in the negative x direction].

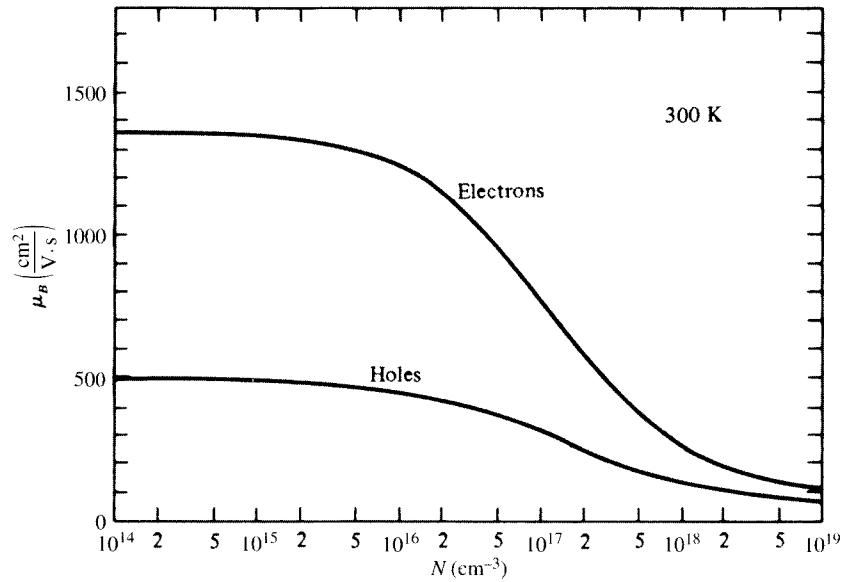


FIGURE 1.13
Electron and hole *bulk* mobility in silicon at 300 K vs. doping concentration.¹⁴

It is seen that the thickness c of the bar does not enter directly; all that counts is the value of the mobile charge concentration per unit area, and the “aspect ratio” b/a . We note that, in the case of the low electric fields we are discussing, the drift current is proportional to the voltage (*Ohm’s law*—see below). It is easy to show that this result is valid even if the electron concentration is nonuniform in the vertical direction, as long as there is uniformity horizontally.[†]

We can obtain the low-field transit time by using (1.3.9) in (1.3.8) and the result in (1.3.3):

$$\boxed{\tau = \frac{a^2}{\mu_B V}} \quad (1.3.11)$$

The presence of the a^2 term reflects the fact that, for fixed μ_B and V , increasing the bar’s length increases the transit time for two reasons: (1) the distance the

[†] To show this, assume that the vertical nonuniformity is kept fixed, and that the flow is laminar, with all electrons moving in the same (horizontal) direction as before. Consider the bar as consisting of thin, parallel horizontal slices within each of which n is constant. Let ΔI be the current and $\Delta Q'$ the charge per unit area for one such slice. Then, from the above results, we have $\Delta I = \mu_B |\Delta Q'| (b/a)V$. For the combination of all slices in parallel, the total current I and charge per unit area Q' are the sums of the corresponding quantities for all slices. Letting ΔI and $\Delta Q'$ become differentials and integrating the resulting equation, we obtain (1.3.10) again.

electrons must travel becomes larger, and (2) the magnitude of the electric field becomes smaller, which slows the electrons down.

Conductance and Conductivity We now connect the preceding results to basic results in electricity. We can obtain an alternative form of (1.3.10) by using (1.3.5b), (1.3.8), and (1.3.9):

$$I = \mu_B n q \frac{bc}{a} V \quad (1.3.12)$$

which is again a statement of Ohm's law, since it is of the form

$$I = GV \quad (1.3.13)$$

In this equation, G is the *conductance*, given by

$$G = \sigma \frac{bc}{a} \quad (1.3.14a)$$

with σ being the *conductivity*, given by

$$\sigma = \mu_B n q \quad (1.3.14b)$$

The inverse of the conductivity, called the *resistivity*, is also used.

Finally, to connect these results to (1.3.10), note that by comparison of that equation to (1.3.13), we have an alternative expression for the conductance:

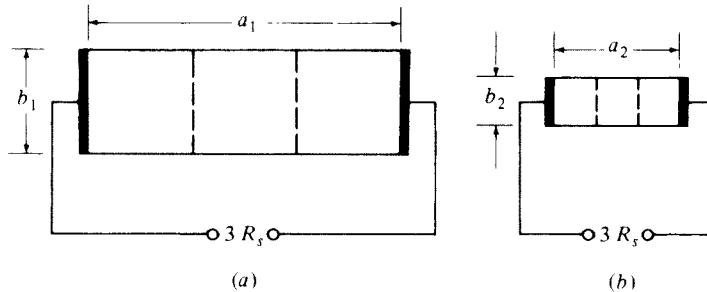
$$G = \mu_B |Q'| \frac{b}{a} \quad (1.3.15)$$

This result can be shown to be valid even if the electron concentration is non-uniform in the vertical direction, as long as there is uniformity horizontally, along the lines of the preceding footnote.

The resistance of the bar $R = 1/G$ is, from (1.3.15),

$$R = R_s \frac{a}{b} \quad (1.3.16)$$

where $R_s = (\mu_B |Q'|)^{-1}$ is called the *sheet resistance*. When $a = b$, i.e., when the bar as seen from the *top* is square, $R = R_s$. Thus R_s is simply the bar's "resistance per square"; R will be given by R_s times the number of squares in the path of the current.

**FIGURE 1.14**

Top view of two bars, each with three “squares” in the path of the current.

This is illustrated in Fig. 1.14. Both bars are assumed to be made out of the same material, and with the same Q' . Since both have a total of $a/b = 3$ squares in the path of the current, they have the same resistance, equal to $3R_s$. Sheet resistance is commonly expressed in “ohms per square.”

Results analogous to the ones in this section can be given in the case of hole conduction. If holes were present in Fig. 1.12, they would move in a direction opposite to that of electrons (i.e., from right to left in the semiconductor). Since the charge associated with holes is positive, the resulting contribution to the total current would be in the direction shown, i.e., in the *same* direction as that for the current owing to electrons. One could then define a conductivity σ for use in (1.3.14a), which would be the *sum* of the electron and hole conductivities.

1.3.3 Diffusion

Drift is only one of two major mechanisms responsible for the flow of electric current in semiconductors. The other mechanism, known as *diffusion*, occurs whenever particles are not distributed uniformly over space, i.e., when there exist “concentration gradients”; then the random motion of the particles tends to make them spread out from regions of high concentration to regions of low concentration. Notice that this phenomenon is not due to electric fields and can thus occur independently of whether or not the particles are charged; for example, particles of smoke exhibit such diffusion. However if the particles *are* charged, as are electrons and holes, diffusion gives rise to movement of charge and thus to electric current.

We consider the origin of diffusion with the help of Fig. 1.15a. Here a piece of semiconductor of rectangular cross section with width b and thickness c is assumed to contain electrons distributed uniformly across any vertical plane (such as the one shown near the middle), but *nonuniformly along the length*. No holes are assumed to be present for now. We assume that the semiconductor communicates with the external world from left and right, so that a fixed electron distribution can be maintained. This distribution can be represented by a fixed plot of n vs. x . For now we will assume this plot to be a straight line as shown in Fig. 1.15b.

Electrons are found on both sides of the plane shown near the middle in Fig. 1.15a, and they exhibit a random thermal motion; however, in a parallelepiped

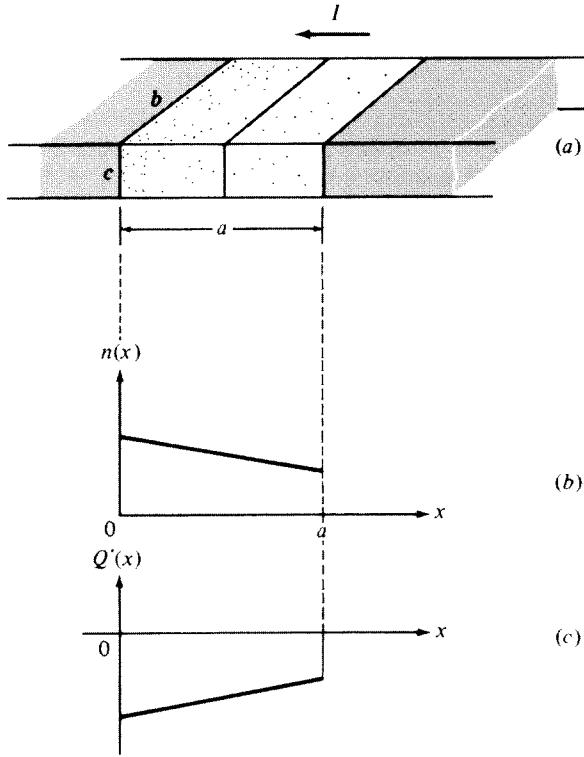


FIGURE 1.15
(a) A semiconductor bar with nonuniform electron concentration along its length; (b) the electron concentration in (a) for a special case of interest; (c) charge per unit area corresponding to (b).

of very small length Δx immediately to the left of the plane, more electrons are to be found than those contained in an identical parallelepiped immediately to the right of the plane. Hence one can expect that in a given time interval more electrons will cross the plane moving from left to right, rather than from right to left. This corresponds to a net flow of electrons from left to right. Since each carries a negative charge, the current defined in the direction shown in the figure will be positive. This current is proportional to the magnitude q of the charge carried by each electron and the cross-sectional area bc . Also, the more negative the slope of n with x , the more positive the current will be; in fact it can be shown that I is proportional to $(-dn/dx)$. We have¹⁻¹¹

$$I = Dq(bc) \left(-\frac{dn}{dx} \right) \quad (1.3.17)$$

where D is a constant of proportionality, called the *diffusion constant*. For nondegenerate semiconductors, this constant is related to the mobility μ_B by the so-called Einstein relationship¹⁻¹¹ (Prob. 1.11):

$$D = \mu_B \phi_t$$

(1.3.18)

where ϕ_t is the thermal voltage given by (1.2.12). Analogous relations hold for the diffusion of holes; of course the values for the diffusion constants of holes and

electrons are different, corresponding to their different mobilities. Note that if the particles shown in Fig. 1.15a were holes, and their concentrations were decreasing toward the right, there would be a hole movement from left to right. Since the charge associated with holes is positive, this movement would correspond to a current in the opposite direction from that shown.

If the plot of n vs. x is not a straight line, dn/dx will vary with position and so will the diffusion current. The latter must now be written as $I(x)$ and will still be given by the right-hand side of (1.3.17). Since, in steady state, the *total* current must be the same at any point x , there must exist in this case a drift current component, varying with position in such a way that the total current (which, as can be shown, will be given by the superposition of diffusion plus drift) is independent of x . In general, current flow in semiconductors is the result of both drift and diffusion; for example, both mechanisms are encountered in a *pn* junction. Often, one of the two mechanisms dominates. In the rest of this section, we assume for simplicity that only diffusion current is present, and thus the plot of n vs. x is a straight line.

Consider a thin vertical slice of the material in Fig. 1.15a of volume $bc \Delta x$ centered around a point at x . The charge in this slice is $(-q)n(x)bc \Delta x$, where the $-$ sign corresponds to the negative electron charge. Dividing the slice charge by the area of the slice $b \Delta x$ as seen from the top, and letting Δx go to zero, we obtain the *charge per unit area*, which here is a function of x :

$$Q'(x) = (-q)c n(x) \quad (1.3.19)$$

This quantity is shown vs. x in Fig. 1.15c. Using (1.3.19) and (1.3.18) in (1.3.17), we obtain

$$I = \mu_B \phi_t b \frac{dQ'(x)}{dx} \quad (1.3.20)$$

Consider now a more general case in which we allow the electron concentration to be *nonuniform in the vertical direction*. Assume that on a horizontal plane at any given depth, the plot of n vs. x is still a straight line (we still assume that n is uniform in the direction perpendicular to the plane of the paper in Fig. 1.15). Since n is now nonuniform vertically, electrons would have a tendency to diffuse in the vertical direction, but we will assume that an appropriate externally applied vertical electric field prevents them from doing so. In other words, we assume that the electrons still move parallel to the x direction, so that *laminar flow is maintained*. Then one can easily show that (1.3.20) will still be valid.[†]

[†] To show this, consider the bar as consisting of thin, parallel horizontal slices within each of which n does not vary vertically. If ΔI and $\Delta Q'(x)$ are the current and charge per unit area for one such slice, we will have $\Delta I = \mu_B \phi_t b d[\Delta Q'(x)]/dx$. The total current I will be the sum of the individual ΔI , and in this summation the $\Delta Q'(x)$ can be grouped together to give the total charge per unit area of the material at x , $Q'(x)$. More precisely, letting ΔI and $\Delta Q'(x)$ become differentials and integrating, we obtain (1.3.20) again (Prob. 1.4).

Since the plot of $Q'(x)$ vs. x is a straight line, its slope will be

$$\frac{dQ'}{dx} = \frac{Q'(a) - Q'(0)}{a} \quad (1.3.21)$$

Use of this in (1.3.20) gives

$$I = \mu_B \phi_t \frac{b}{a} [Q'(a) - Q'(0)] \quad (1.3.22)$$

Since (1.3.20) is valid even if n is a function of both x and the vertical dimension under the assumptions following (1.3.20), then (1.3.22) will also be valid in such cases.

The transit time τ can now be easily calculated. It is easy to show that the total charge of the electrons found at any instant inside the bar will be (Prob. 1.3)

$$Q = ab \frac{Q'(a) + Q'(0)}{2} \quad (1.3.23)$$

Thus the transit time will be, from (1.3.2),

$$\tau = \frac{a^2}{\mu_B(2\phi_t)} \frac{Q'(0) + Q'(a)}{Q'(0) - Q'(a)} \quad (1.3.24)$$

In this book we will often be interested in the case where a “perfect sink” exists at $x = a$, reducing the electron concentration to 0 there; this means

$$Q'(a) = 0 \quad (1.3.25)$$

For this special case (1.3.24) becomes

$$\tau = \frac{a^2}{\mu_B(2\phi_t)}$$

(1.3.26)

Note that this form is very similar to that for drift in (1.3.11). However, in contrast to the drift case, where τ can be made small by applying a large V , here τ is fixed at a comparatively large value, due to the fixed, small value of $2\phi_t$ (about 52 mV at 300 K).

1.3.4 Total Current

In the general case, the total current in a semiconductor is due to both electrons and holes, and each type of carrier contributes both drift and diffusion components. Thus,

there is a total of four possible current components. As we will see, in many instances a simplification can be made because some of those components are negligible.

It can be shown (Appendix B) that the electron current (drift plus diffusion) is proportional to the electron concentration and to the rate of change of the electron quasi-Fermi level E_{Fn} , discussed in Sec. 1.2.5, with respect to position. Similarly, the hole current is proportional to the hole concentration and the gradient of the hole quasi-Fermi level E_{Fp} . In thermal equilibrium, these levels are equal to the Fermi level, which is constant. Thus, thermal equilibrium implies zero electron current *and* zero hole current.

1.4 Contact Potentials

Consider the junction of two different materials $M1$ and $M2$ with no external bias, as shown in Fig. 1.16a; either material can be a semiconductor or a metal. When the two materials are brought together, at first carriers move from one to the other because the energy of these carriers is in general different in $M1$ and $M2$, and no opposing field exists in the initially neutral materials. However, as each charged carrier crosses the junction, it leaves behind a net charge of the opposite polarity, and an electric field is thus established in the vicinity of the junction, which inhibits the movement of carriers. For example, if an electron crosses from $M1$ to $M2$, it leaves a positive charge in $M1$; the contribution of this charge to the electric field is in such a direction as to attract the electron back into $M1$. Eventually, the field intensity increases to the point that it counteracts the tendency of carriers to cross the junction, and a balance is achieved such that there is no further

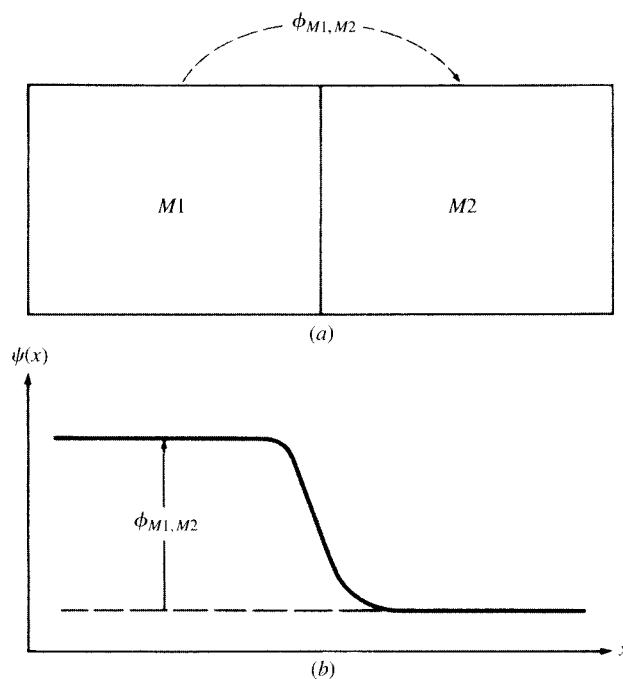


FIGURE 1.16

(a) Two different materials in contact (either can be a metal or a semiconductor); (b) potential vs. distance for (a).

net carrier movement. An electrostatic potential change is then encountered in going from one material, through the junction, to the other material. If $\psi(x)$ is the potential at position x , it can be of the form shown in Fig. 1.16b (arbitrary references are assumed for both ψ and x). Depending on the two materials, the potential change can occur on both sides of the junction, as in the case shown, or mostly on one of them. The total potential drop in going from $M1$ to $M2$ is called the *contact potential* of material $M1$ to material $M2$, and will be denoted by $\phi_{M1,M2}$. The concept of contact potentials will be instrumental in deriving the basic equations of MOS devices.

Contact potentials can be determined using the concept of so-called work functions, which will be explained with the help of Fig. 1.17. In Fig. 1.17a, we show the two materials $M1$ and $M2$ before they are brought into contact. The relevant energy levels are shown in Fig. 1.17b. Here E_R represents the so-called *vacuum energy level*, corresponding to the energy of an electron when it is removed from a material, so that it is not influenced by it. The difference between E_R and a Fermi energy E_F is called the *work function* and is denoted by $q\phi_{W,Mi}$ for material M_i , where q is the electron charge, and $\phi_{W,Mi}$ is the work function divided by q . The reason for the use of the potential $\phi_{W,Mi}$ is just convenience in writing the equations that follow. The value of $\phi_{W,Mi}$ in volts is numerically equal to the value of the work function in electron volts.

The work function is a measure of how difficult it is for an electron to leave its host material. In Fig. 1.17, we have $q\phi_{W,M1} < q\phi_{W,M2}$, and thus electrons find it easier to leave $M1$ and enter $M2$. When the two materials are brought together to form a junction as in Fig. 1.16a, electrons are transferred from $M1$ to $M2$; eventually the potential of $M2$ with respect to $M1$ becomes so negative as to inhibit further net transfer of electrons. This happens when the average energy of the electrons in $M2$ (relative to that in $M1$) has increased by the amount $q\phi_{W,M2} - q\phi_{W,M1}$, in which case the two Fermi energies become equal, as expected for equilibrium. From (1.2.17), this corresponds to a *drop* in electrostatic potential of $(q\phi_{W,M2} - q\phi_{W,M1})/q$; this is the contact potential of $M1$ to $M2$, indicated on Fig. 1.16b. In general, then, the contact potential $\phi_{M1,M2}$ of a material $M1$ to a material $M2$ is given by

$$\phi_{M1,M2} = \phi_{W,M2} - \phi_{W,M1} \quad (1.4.1)$$

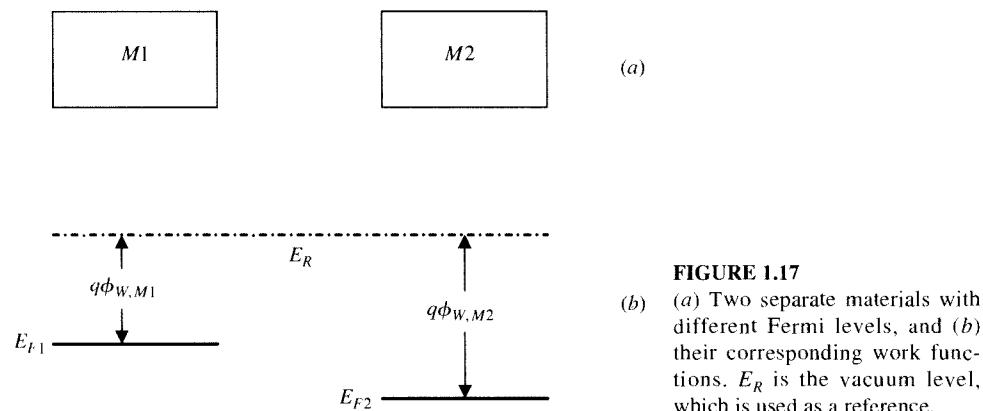


FIGURE 1.17
(a) Two separate materials with different Fermi levels, and (b) their corresponding work functions. E_R is the vacuum level, which is used as a reference.

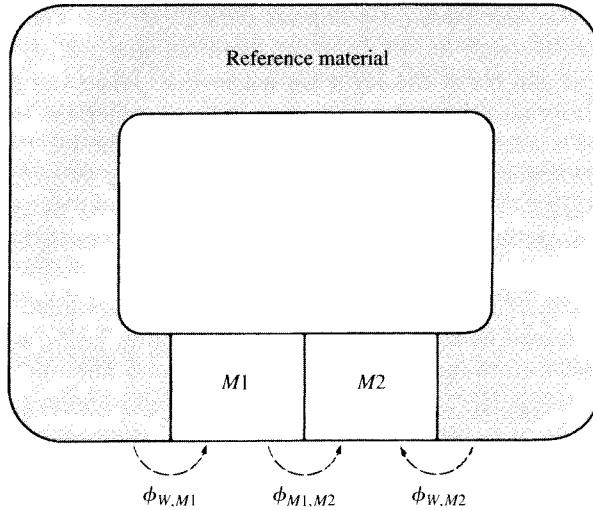


FIGURE 1.18
Two materials in a loop with a reference material.

There is nothing special in using the vacuum as a common reference. In fact, a different reference can be used. This is illustrated in Fig. 1.18. Suppose we place the *M*₁-*M*₂ contact in a loop with a reference material, as shown. Let ϕ_{W,M_i} be the contact potential of the reference material to material *M*_{*i*}, as shown. Such potentials can be measured and tabulated. Going around the loop, the sum of electrostatic potentials must sum to zero. This gives precisely (1.4.1) again. In previous editions of this book, intrinsic silicon was used as the reference; we have now switched to vacuum to be compatible with other treatments.

Work functions and work function potentials are not easy to measure, and sometimes inconsistent values are reported in the literature for various metals. For example, 4.1 V is often used for aluminum, but one also encounters values of 4.0 to 4.3 V; similarly, for copper, 4.5 V is often used, but one also encounters values of 4.45 to 4.8 V. In MOS modeling work, “modified” work function values are used to make measurements agree with theory.

For semiconductors, the work function potential, denoted by $\phi_{W,S}$, can be calculated as is now explained with the help of Fig. 1.19, which shows the cases of both *n*-type and *p*-type semiconductors. The quantity χ is known as the “electron affinity.” The quantity $E_c - E_i$ is very close to half the band gap, as has already been mentioned. The quantity ϕ_F is positive for *p*-type material and negative for *n*-type (Sec. 1.2.3). Thus, from the figure, for both *n*-type and *p*-type cases, the work function potential is

$$\phi_{W,S} = \chi + \frac{E_g}{2q} + \phi_F \quad (1.4.2)$$

The value of χ is 4.05 V for silicon, and E_g/q is 1.12 V at room temperature for nondegenerate silicon. The Fermi potential for nondegenerate silicon can

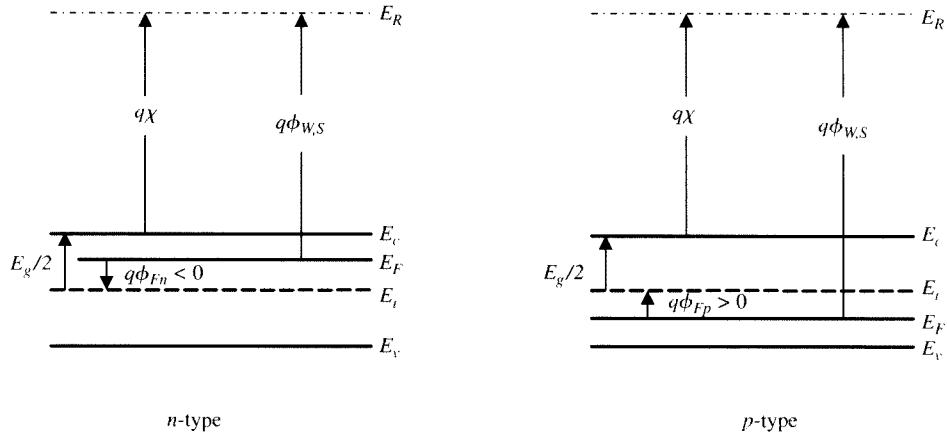


FIGURE 1.19
Semiconductor work function $q\phi_{W,S}$ and its components.

be calculated from (1.2.14), assuming all atoms are ionized. For degenerate *n*-type silicon, as we have mentioned before, we assume that the Fermi energy coincides with E_c , and thus $\phi_{W,S} = \chi$. For degenerate *p*-type silicon, we assume that the Fermi energy coincides with E_v , and thus $\phi_{W,S} = \chi + E_g/q$.

Example 1.1. Find the contact potential of aluminum to *p*-type silicon with a doping of 10^{17} cm^{-3} , at room temperature. Assume that the work function potential of aluminum is 4.1 V.

Let *S* denote the semiconductor material. From (1.2.14a), we have $\phi_F \approx 0.42 \text{ V}$, and (1.4.2) gives $\phi_{W,S} \approx 4.05 \text{ V} + 0.56 \text{ V} + 0.42 \text{ V} = 5.03 \text{ V}$. Thus, from (1.4.1) we get

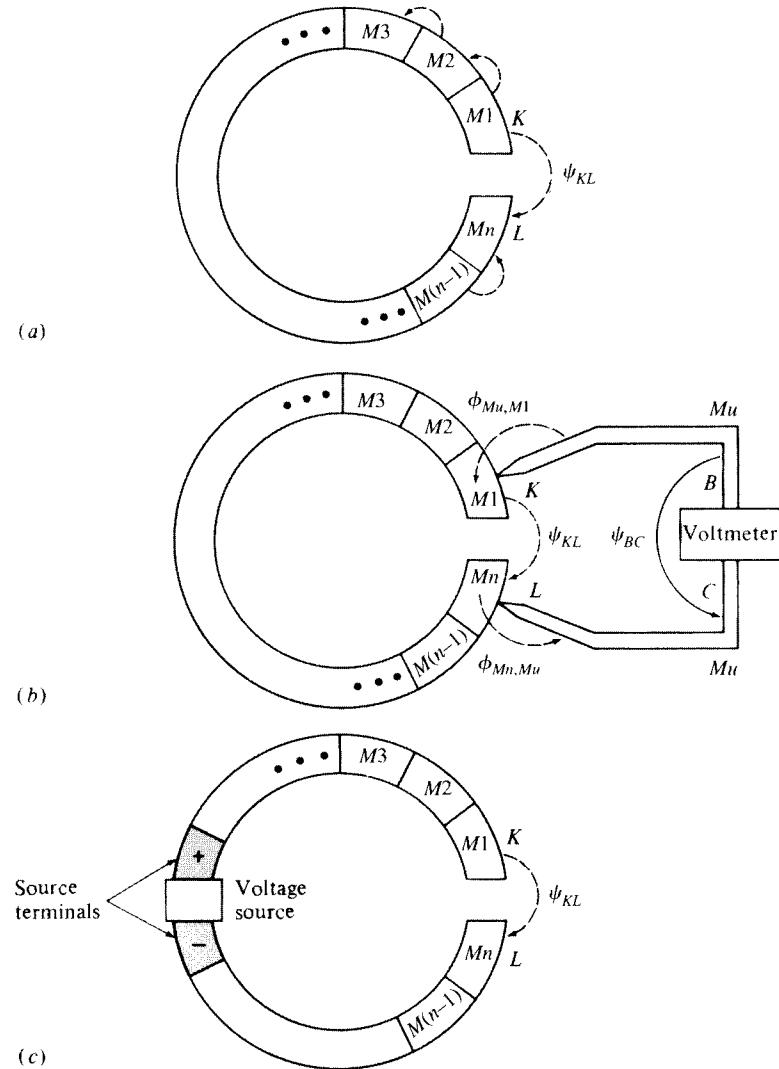
$$\phi_{Al,S} = \phi_{W,S} - \phi_{W,Al} \approx 5.03 \text{ V} - 4.1 \text{ V} = 0.93 \text{ V}$$

Several Materials in Series We can now make an interesting observation concerning several materials in series, as in Fig. 1.20*a*; as before, all materials are assumed to be at the same temperature. If we express the potentials ψ_{KL} in terms of contact potentials in the loop, we have

$$\psi_{KL} = \phi_{M1,M2} + \phi_{M2,M3} + \cdots + \phi_{M(n-1),Mn} \quad (1.4.3)$$

and, using (1.4.1) repeatedly, we get

$$\begin{aligned} \psi_{KL} = & (\phi_{W,M2} - \phi_{W,M1}) + (\phi_{W,M3} - \phi_{W,M2}) + \cdots + (\phi_{W,M(n-1)} - \phi_{W,M(n-2)}) \\ & + (\phi_{W,Mn} - \phi_{W,M(n-1)}) \end{aligned} \quad (1.4.4)$$

**FIGURE 1.20**

(a) Several materials in series; (b) the structure of (a) with a voltmeter attached; (c) the structure of (a) with a voltage source placed in series.

It is clear that, with the exception of ϕ_{W,M_1} and ϕ_{W,M_n} , each ϕ_{W,M_i} appears twice in the sum—once with a plus and once with a minus sign. Therefore

$$\psi_{KL} = \phi_{W,M_n} - \phi_{W,M_1} \quad (1.4.5)$$

A question arises: Can we measure the potential ψ_{KL} with a common voltmeter? The answer is *no*, even if the voltmeter is assumed ideal (i.e., does not draw any current from the circuit under measurement). This can be seen in Fig. 1.20b. Assume

that the voltmeter leads are made by some material Mu . The potential difference ψ_{BC} between the two leads (which is what the volunteer “sees”) is

$$\psi_{BC} = \phi_{Mu,M1} + \psi_{KL} + \phi_{Mn,Mu} \quad (1.4.6)$$

Using (1.4.1) in the previous equation,

$$\psi_{BC} = (\phi_{W,M1} - \phi_{W,Mu}) + \psi_{KL} + (\phi_{W,Mu} - \phi_{W,Mn}) \quad (1.4.7)$$

and, using (1.4.5), we get

$$\psi_{BC} = 0 \quad (1.4.8)$$

In summary then, no matter how many materials are in the loop, the electrostatic potential difference between its two ends depends *only* on the first and the last material and cannot be measured by a common voltmeter.[†] For such measurements, special techniques must be used.^{15,16}

What if a voltage source is inserted in the loop? This situation is illustrated in Fig. 1.20c. The source leads are assumed to be made of the same material. By going around the loop we have

$$\psi_{KL} = V_{\text{source}} + (\phi_{W,Mn} - \phi_{W,M1}) \quad (1.4.9)$$

which should be compared with (1.4.5).[‡]

What would an ideal voltmeter like that of Fig. 1.20b measure if attached to points K and L in Fig. 1.20c? With ψ_{KL} as given by (1.4.9), (1.4.7) is still valid and we have

$$\psi_{BC} = V_{\text{source}} \quad (1.4.10)$$

which should be compared with (1.4.8). As will be seen for a *pn* junction in the next section, a voltage source causes a split of the quasi-Fermi levels; it is this split that a common voltmeter measures.

[†] The leads of a voltmeter are invariably made of the same material. The “academic” case where the leads are made of two different materials can be handled if one considers the various contact potentials within the voltmeter circuit (see Prob. 1.7).

[‡] Leads of voltage sources, like the leads of a voltmeter, are assumed to be made of the same material. The case of a source with leads made of different materials is easily handled by defining V_{source} as the voltage measured by our ideal voltmeter when attached to the terminals of the source. Then (1.4.9) and (1.4.10) will remain valid (see Prob. 1.8).

It is now clear why contact potentials never seem to enter the picture when one works with circuits, provided all contacts are at the same temperature. As seen in the example leading to (1.4.10), such potentials cancel out, cannot be measured by common voltmeters, and do not enter into *circuit* equations. However, in investigating the physics of electronic devices, contact potentials must be taken into account; otherwise, unless a fortuitous cancellation takes place, neglecting contact potentials can lead to serious errors.

1.5 The *pn* Junction

The selective doping of silicon with *n*-type impurities in one region and with *p*-type impurities in an adjacent one produces what is known as a *pn junction*. Such a junction is shown in Fig. 1.21. In this figure, we have assumed that the doping changes abruptly from *n*-type to *p*-type at a boundary, and that the doping concentration is uniform on either side of the boundary; such a junction is referred to as an *abrupt*, or *step*, junction. In drawing Fig. 1.21, we have also assumed that the *n* region is more heavily doped than the *p* region. The two caps on the top and the bottom are made of a metal chosen so as to make “ohmic contacts” (contacts that conduct well in both directions[†]) with the semiconductor.

Zero Bias Let us first consider the case in Fig. 1.21*a* where the externally applied bias is zero. Free electrons are in high concentration in the *n* side and in low concentration in the *p* side. Thus, they tend to diffuse across the boundary between the two sides (downward in the figure), leaving behind positively charged immobile donor atoms as shown. Similarly, holes tend to diffuse from the *p* side, where they are in abundance, upward to the *n* side, leaving behind negatively charged immobile acceptor atoms. For simplicity it is often assumed that these phenomena result in complete depletion of carriers over a region with sharply defined edges, as shown in the figure. This is called the *depletion approximation*, and the region is referred to as the *depletion region*. The more uncovered donor atoms there are on the *n* side, the greater their total charge; therefore, the stronger the electric field they produce, the stronger the attracting force they exert on electrons attempting to diffuse downward, and the stronger the repelling force on holes attempting to diffuse upward. Similarly, the more uncovered acceptor atoms there are on the *p* side, the more they will attract holes that are attempting to diffuse upward, and the more they will repel electrons that are attempting to diffuse downward. Eventually in this process, enough impurity atoms are uncovered on both sides that the strong fields exerted by them inhibit a further net movement of carriers. The depletion region width is then fixed, and the current in the external wire is zero. It is interesting to note that the situation inside the structure will not change if the external wire is now cut. In fact, in this case, it is easy to see from (1.4.4) and the associated discussion that the three contact potentials in the structure cancel each other out, and no electrostatic potential will exist across the two open terminals.

[†] How well current is conducted in either direction is something that cannot be deduced from the contact potential value. The theories for conduction in a metal-semiconductor contact are relatively involved.^{5,17}

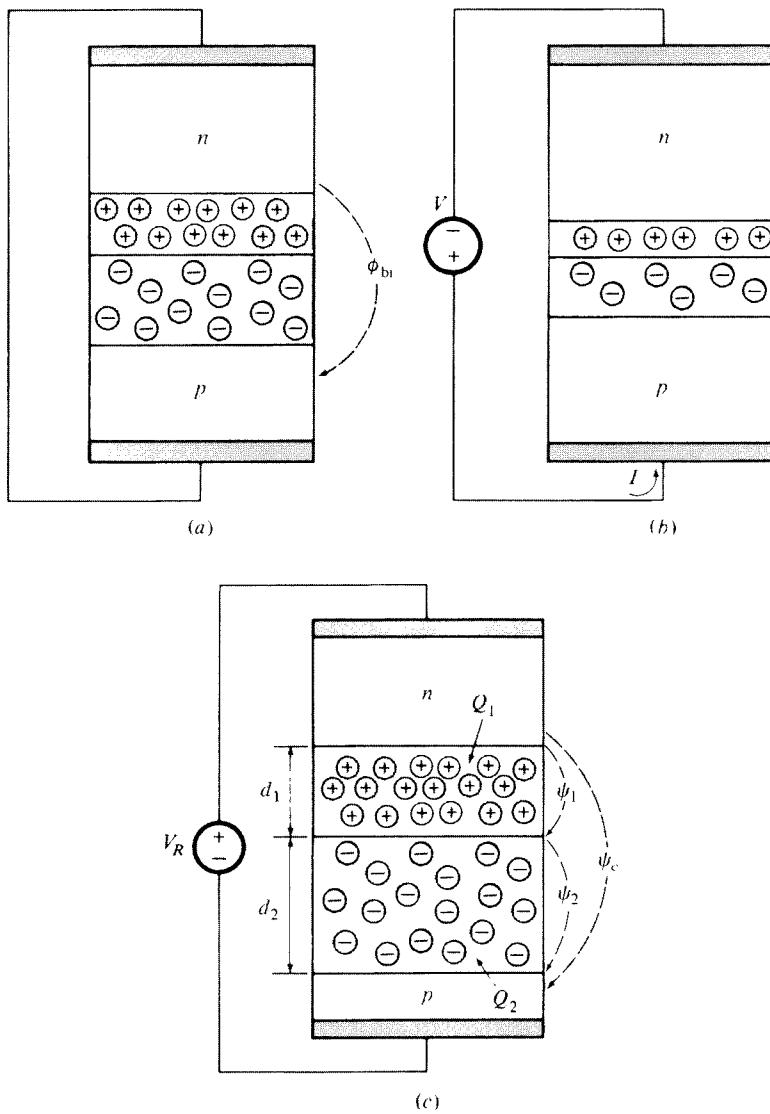


FIGURE 1.21
A *pn* junction with (a) zero bias, (b) forward bias, (c) reverse bias.

For overall charge neutrality, the magnitude of the charge on each side must be the same. Because the *n* region is assumed more heavily doped than the *p* region, the depth over which donor atoms must be uncovered on the *n* side is smaller than the depth over which acceptor atoms must be uncovered on the *p* side, as shown in Fig. 1.21.

In the absence of external bias, the potential across the depletion region (from deep in the *n* side to deep in the *p* side) will simply be the contact potential of the

n-type material to the *p* material. This potential is referred to as the *built-in potential* of the junction, and will be denoted by ϕ_{bi} . To calculate this potential, we will use energy band principles. The energy band diagrams for the *n* and *p* sides *before* they are joined together have been shown in Fig. 1.19. After the materials are brought into contact as in Fig. 1.21*a*, the terminals are shorted together, and equilibrium is established, we obtain the situation shown in Fig. 1.22*a*, where we have used a horizontal orientation consistent with that in Fig. 1.19 (corresponding to a rotation of Fig. 1.21 counterclockwise by 90°). The separation of the electron and hole Fermi levels is reduced to zero, and a single Fermi level prevails. Far away from the junction, the two materials are still neutral, and the relative position of the energy bands there is the same as in Fig. 1.19. This means that the energy levels E_c , E_i , and E_v *rise* from left to right as shown in Fig. 1.22*a*. This corresponds to a *drop* in electrostatic potential, in agreement with (1.2.17), as shown in Fig. 1.22*b*. As seen from the figure, using the Fermi potentials ϕ_{Fn} and ϕ_{Fp} (recall that ϕ_{Fp} is positive and ϕ_{Fn} is negative), gives, for the built-in (contact) potential

$$\phi_{bi} = \phi_{Fp} - \phi_{Fn} \quad (1.5.1)$$

Let us check if this result is consistent with our discussion of contact potentials in Sec. 1.4. Because both sides are silicon, χ and $E_g/2q$ are the same for them and, when (1.4.2) is used in (1.4.1), they cancel out, indeed giving (1.5.1) again.

Example 1.2. Assume that the *n* side is degenerate, and the doping of the *p* side is 10^{17} cm^{-3} . Find the value of the contact potential ϕ_{bi} at room temperature.

From (1.2.14*a*), we have $\phi_{Fp} = 0.42 \text{ V}$. For the degenerate *n* side, we estimate the Fermi potential using the approximation given on page 10: $\phi_{Fn} = -0.56 \text{ V}$. So, from (1.5.1),

$$\phi_{bi} = 0.42 \text{ V} - (-0.56 \text{ V}) = 0.98 \text{ V}$$

Forward Bias We now consider briefly the application of a *forward bias voltage* $V > 0$, as shown in Fig. 1.21*b*. Equilibrium is now destroyed. The polarity of the

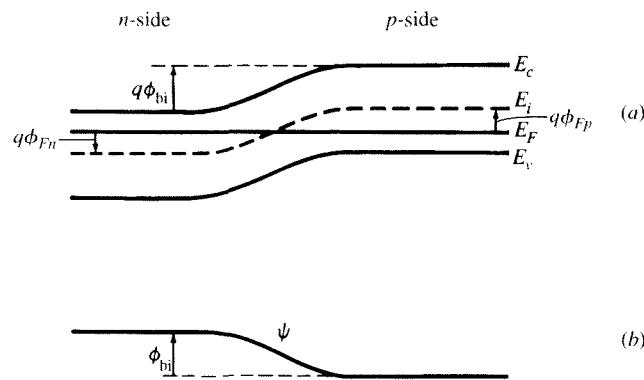
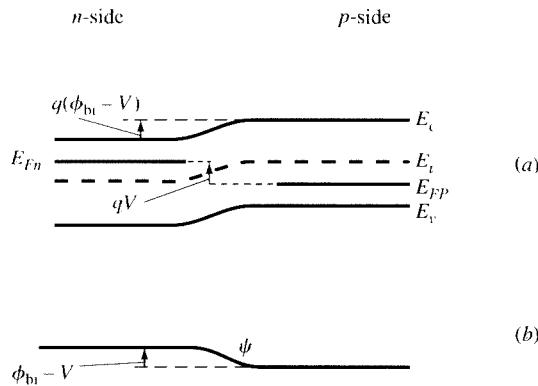


FIGURE 1.22
(b) (a) Energy band diagram and (b) potential for a *pn* junction with zero bias.



(b) **FIGURE 1.23**
(a) Energy band diagram and (b) potential for a *pn* junction under forward bias V .

voltage tends to produce a field opposite from that produced by the uncovered impurity atoms. Assuming that the contact potential of the *ohmic* contacts is not affected, the electrostatic potential across the depletion region will be reduced from ϕ_{bi} to $\phi_{bi} - V$, so that the sum of the electrostatic potentials around the loop remains zero. This reduction in potential means that the total energy band bending will be reduced, as shown in Fig. 1.23. In the regions away from the depletion region, the majority carrier concentrations remain practically at their equilibrium values. However, the electrostatic potential across the depletion region must decrease by V , corresponding to a decrease in the energy band bending by qV , compared with Fig. 1.22. The structure is no longer in equilibrium; we need separate quasi-Fermi levels so that, despite the extra band bending, (1.2.20) can give majority carrier concentrations at the *n* and *p* side, respectively, which have practically the same values as for Fig. 1.22. The two quasi-Fermi levels are in this case said to “split” by the amount qV . It is seen that, in comparison with Fig. 1.22, *the barrier to electrons’ crossing from the n side to the p side has been reduced*. The same holds for holes, crossing in the opposite direction. Thus, carriers find it easier to diffuse across the boundary. Electrons from the *n* side, being majority carriers there and thus in large supply, diffuse downward; holes from the *p* side, which are majority carriers there and therefore again in large supply, diffuse upward. The result of both of these movements is a positive current upward, as shown in Fig. 1.21*b*. For this to make sense, one should recall that the assumption that *all* mobile carriers have left the depletion region, used previously, was just a convenient simplification; mobile carriers actually do exist in the “depletion” region (although their density is much lower than that of the uncovered dopant atoms) and can support the flow of current. Assuming the current is not too large, and that the ohmic drop in the semiconductor outside the depletion region is negligible, it can be shown that the magnitude of the current will be given by^{2-10,12}

$$I = I_0(e^{V/\phi_t} - 1) \quad (1.5.2)$$

where I_0 is a quantity dependent on junction geometry and physical parameters of the semiconductor material and is an increasing function of temperature, and ϕ_t is the thermal voltage given by (1.2.12).

As has already been seen, the total energy band bending under forward bias, (compared with the case of zero bias) is reduced, and thus so is the total potential across the depletion region. This reduction in potential must be supported by a corresponding reduction of the charges in the depletion region, and thus a reduction in the width of this region on both sides of the boundary, as shown in Fig. 1.21b. Some of the previously ionized donor atoms on the *n* side are now “covered” by electrons; similarly, on the *p* side, some of the previously ionized acceptor atoms are now covered by holes. The maximum electric field in the depletion region is reduced as well.

Reverse Bias Consider now the case shown in Fig. 1.21c, where a reverse bias $V_R > 0$ is applied from the *n*-side terminal to the *p*-side terminal. This is the most important case for us, as it will be related to phenomena in the MOS transistor in the following chapters; so it will be analyzed in more detail. The electrostatic potential across the depletion region (ψ_c in Fig. 1.21c) will increase by V_R , so that the potential drops around the loop add up to zero again (the contact potentials at the ohmic contacts are considered unchanged). Thus

$$\psi_c = \phi_{bi} + V_R \quad (1.5.3)$$

Again, the structure will be in nonequilibrium. The energy band diagram is shown in Fig. 1.24. The quasi-Fermi levels split again by the amount corresponding to the externally applied bias, but in the opposite direction from before. The total band bending corresponds to (1.5.3) and is larger than for the shorted junction. This implies that more impurity atoms need to be uncovered, to contribute to a larger field and thus larger potential. Thus the depletion region widens, as illustrated in Fig. 1.21c. This is consistent with the fact that the polarity of the external source is such that it tends to move electrons upward away from the upper edge of the depletion region and holes downward away from the lower edge of that region. Note that the energy barrier to electrons’ crossing from the *n* side to the *p* side is increased, compared with the case in Fig. 1.22.

Let the depths of the depletion region at the two sides of the junction in Fig. 1.21c be d_1 and d_2 , as shown. If the cross-sectional area of the junction (as

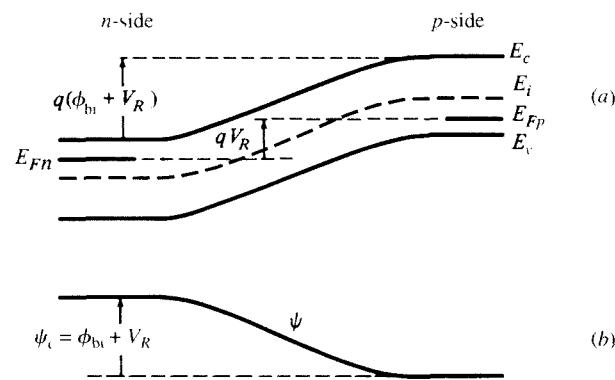


FIGURE 1.24
(a) Energy band diagram and (b) potential for a *pn* junction under reverse bias V_R .

seen from above) is A , the part of the depletion region in the n side has a volume equal to $(d_1 A)$, and therefore a total number of ionized atoms equal to $(d_1 A)N_D$. Because each of these atoms carries a charge $+q$, the total charge there, denoted by Q_1 , is

$$Q_1 = +q(d_1 A)N_D \quad (1.5.4)$$

Similarly, the total charge in the p side of the depletion region (which is due to ionized acceptor atoms) is

$$Q_2 = -q(d_2 A)N_A \quad (1.5.5)$$

For overall charge neutrality, we must have

$$Q_1 = -Q_2 \quad (1.5.6)$$

This implies

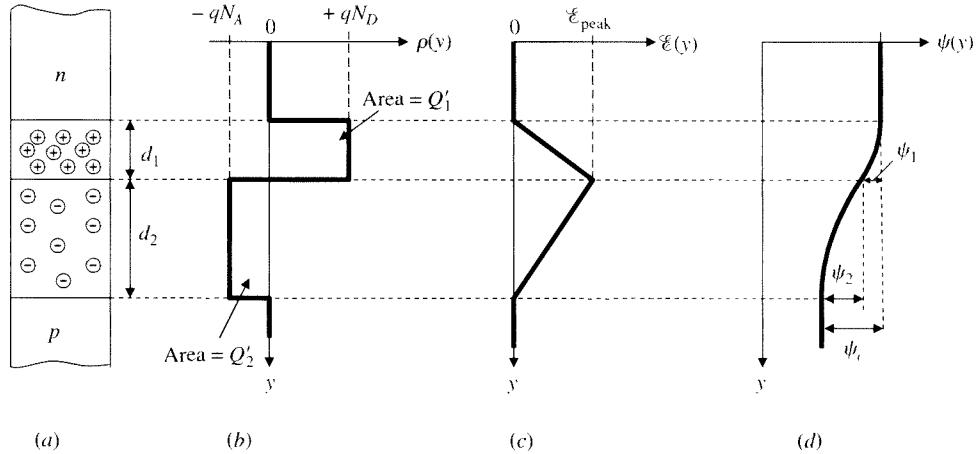
$$\frac{d_1}{d_2} = \frac{N_A}{N_D} \quad (1.5.7)$$

We now calculate the electric field in the depletion region. Part of Fig. 1.21c is repeated as Fig. 1.25a. The corresponding charge density is shown in Fig. 1.25b. We can calculate the electric field from the charge density by using (1.2.26), with y_0 a point above the depletion region; we assume that there is no external electric field there. As the charge density is piecewise constant, the integration gives straight-line segments for the field plot, as shown in Fig. 1.25c. Because of (1.5.7), the field ends up with a value of zero below the depletion region, the value that it started with above it. This is consistent with charge neutrality outside the depletion region. As follows from (1.2.26), the peak field $\mathcal{E}_{\text{peak}}$ is simply the area under the positive part of the charge density plot, divided by ϵ_s . This, together with (1.5.7) gives

$$\mathcal{E}_{\text{peak}} = \frac{qN_D d_1}{\epsilon_s} = \frac{qN_A d_2}{\epsilon_s} \quad (1.5.8)$$

Note that we could have obtained the last part of this equation directly, if we had integrated the charge density in the opposite direction, starting from a point below the depletion region. The distances d_1 and d_2 will be determined shortly.

From the electric field, we can obtain the potential by using (1.2.28). The result is shown in Fig. 1.25d; integration of the electric field straight-line segments results in two parabolic segments for the potential. The potential variation on each side is

**FIGURE 1.25**(a) A reverse-biased *pn* junction; (b) charge density; (c) electric field; (d) potential.

simply the area of the corresponding triangle in Fig. 1.25c. Thus, for the two potentials noted in the figure, we have

$$\psi_1 = \frac{\epsilon_{\text{peak}} d_1}{2} = \frac{q N_D d_1^2}{2 \epsilon_s} \quad (1.5.9)$$

$$\psi_2 = \frac{\epsilon_{\text{peak}} d_2}{2} = \frac{q N_A d_2^2}{2 \epsilon_s} \quad (1.5.10)$$

and these potentials sum up to the total potential across the depletion region, given by (1.5.3):

$$\psi_1 + \psi_2 = \psi_c \quad (1.5.11)$$

The case handled so far is that of a “two-sided” step (abrupt) junction, in which N_A and N_D can have comparable values.[†] Let us now restrict our attention to the case

[†] From (1.5.8)–(1.5.11), it can be calculated that (Prob. 1.9)

$$d_1 + d_2 = \sqrt{\frac{2\epsilon_s}{q} \frac{N_A + N_D}{N_A N_D} \psi_c}$$

Thus, if desired, d_1 and d_2 can be calculated from this equation and (1.5.7). Using then d_2 in (1.5.5) one can find the charge per unit area on the *p* side, $Q'_2 = Q_2/A$, to be

$$Q'_2 = -\sqrt{2q\epsilon_s} \sqrt{\frac{N_A N_D}{N_A + N_D} \psi_c}$$

of a “one-sided” step junction of the type n^+p , i.e., a step junction for which

$$N_D \gg N_A \quad (1.5.12)$$

This, along with (1.5.7), implies

$$d_1 \ll d_2 \quad (1.5.13)$$

which means that practically all the depletion region extends into the p side. This relation, from the first part of (1.5.9) and (1.5.10), implies

$$\phi_1 \ll \phi_2 \quad (1.5.14)$$

Thus, from (1.5.11):

$$\phi_2 \approx \psi_c \quad (1.5.15)$$

i.e., the total potential drop across the depletion region in a n^+p junction (n^+ standing for heavily doped n) appears almost entirely on the p side. Using this equation in (1.5.10) and solving for d_2 gives the depth of the depletion region on the p side as

$$d_2 = \sqrt{\frac{2\epsilon_s}{qN_A}} \sqrt{\psi_c} \quad (1.5.16)$$

where, in this and the equations to be derived from it, we are assuming that the equality is exact for simplicity. Let us now define the charge *per unit area* on the p side, Q'_2 , as follows:[†]

$$Q'_2 = \frac{Q_2}{A} \quad (1.5.17)$$

Using (1.5.5) in this equation, we see that Q'_2 is represented graphically by the area under the lower rectangle in Fig. 1.25b, as indicated in that figure.[‡] Using (1.5.5) and (1.5.16) in (1.5.17), we obtain

$$Q'_2 = -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_c} \quad (1.5.18)$$

From (1.2.1) and (1.2.25), $\sqrt{2q\epsilon_s}$ has a value of $5.8 \times 10^{-16} \text{ C} \cdot \text{V}^{-1/2} \cdot \text{cm}^{-1/2}$ for silicon (or, equivalently, $5.8 \times 10^{-16} \text{ F} \cdot \text{V}^{1/2} \cdot \text{cm}^{-1/2}$).

[†] The results derived here will later be extended to cases where d_2 changes with horizontal position; in such cases, one must define Q' at each horizontal position incrementally as $\Delta Q'_2 / \Delta A$.

[‡] Similarly, the area under the top rectangle is $Q'_1 = Q_1 / A$.

In the case of reverse bias we are considering, the polarity of the externally applied bias no longer aids the flow of majority carriers from each side toward the junction boundary, as it did in the case of forward bias. Instead, on the *n* side it tends to pull electrons (majority carriers) upward, away from the junction; likewise, on the *p* side it tends to pull holes downward, also away from the junction. Nevertheless, a small junction current does flow, as will now be explained. Some of the few holes (minority carriers) near the depletion region edge in the *n* side are swept downward by the field; similarly, some of the few electrons near the depletion region edge in the *p* side are swept by the field upward. This flow of minority carriers from each side causes a net external current in the opposite direction of that in Fig. 1.21*b*; and, since minority carriers are in short supply, this current is very small. Outside the depletion region, the current is carried mostly by majority carriers.[†]

Equation (1.5.2) remains theoretically valid for the reverse-bias case, with $V = -V_R$. For V_R larger than a few ϕ_i , the exponential becomes negligible and the current is seen to “saturate” at the value $-I_0$. For this reason, I_0 is often referred to as the *saturation current*. In practice, other phenomena (e.g., electron-hole pair thermal generation in the depletion region) contribute to make the actual reverse-current magnitude larger than I_0 and indeed somewhat dependent on V_R .[‡] The reverse-current magnitude is strongly dependent on temperature, and can double for every 8°C or even less of a temperature change. If V_R exceeds a certain value V_{RX} , the theory^{4-9,12} used to derive (1.5.2) fails. The junction conducts a large current from the *n* to the *p* side and is said to be in “reverse breakdown.” V_{RX} is called the *reverse-breakdown voltage*, and, for the junctions encountered on MOS integrated circuits, it is typically from a couple of volts in common integrated circuits to over 100 V in special high-voltage chips. A complete *I-V* characteristic of a *pn* junction, including breakdown, is shown in Fig. 1.26. Most *pn* junctions in MOS integrated circuits are kept reverse-biased.

Small-Signal Capacitance We now consider the concept of the small-signal capacitance of the reverse-biased *pn* junction. If the reverse bias is increased by a small amount ΔV_R , the depletion region width must increase on both sides. The charge that must be moved for this to happen must flow through the external circuit. Specifically, additional donor atoms will be uncovered in the *n* side, increasing the total charge there. For this to happen, electrons must be removed by the external circuit from near the depletion region’s top edge. Thus, a negative charge, say $-\Delta Q$, leaves

[†] The quasi-Fermi levels $E_{f,n}$ and $E_{f,p}$ are almost constant with position in the neutral regions, but not exactly [the total electron and hole currents can be shown to be proportional to the corresponding carrier concentrations and to the gradients of the corresponding quasi-Fermi potentials with position (Appendix B)]. Because the currents are very small whereas the carrier concentrations are very large, the corresponding quasi-Fermi level gradients are very small in the neutral regions. At other places, the shape of the quasi-Fermi levels can be deduced from considerations of the carrier concentrations and the detailed mechanisms associated with the reverse-bias current. For the purposes of this book, such considerations are not essential. It is only remarked here that assumptions commonly made lead to the conclusion that the quasi-Fermi levels continue practically horizontal over the depletion region, and then, outside it, gradually merge into single levels well into the neutral regions (see, for example, Refs. 4 and 5).

[‡] In junctions that are part of MOS transistors, additional leakage mechanisms are present (Chap. 5).

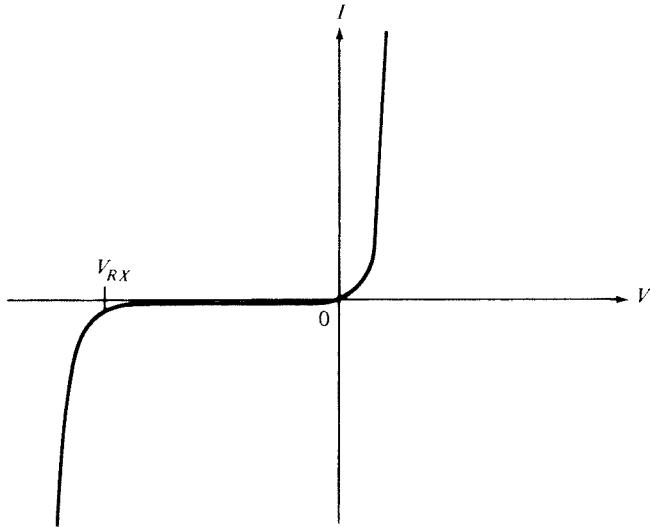


FIGURE 1.26
Current vs. voltage for a pn junction.

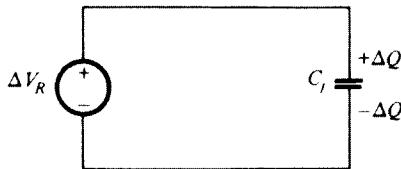


FIGURE 1.27
Small-signal equivalent circuit of a reverse-biased pn junction driven by a voltage source.

the junction from the top terminal. This can equivalently be described by saying that an opposite *positive* charge $+ΔQ$ flows *into* the junction through the top terminal. Similarly, since the reverse bias has increased, additional acceptor atoms must be uncovered in the p side. For this to happen, the external circuit must cause positively charged holes to be removed from near the depletion region's bottom edge. This implies a positive charge $+ΔQ$ flowing out of the bottom terminal or, equivalently, a *negative* charge $-ΔQ$ flowing *into* the bottom terminal. Thus the changes of the charges Q_1 and Q_2 in Fig. 1.21c will be

$$\Delta Q_1 = +ΔQ \quad (1.5.19)$$

$$\Delta Q_2 = -ΔQ \quad (1.5.20)$$

Assume now that a voltage source of value equal to the *change* of the reverse bias $ΔV_R$ is connected across a capacitor as in Fig. 1.27, and that its capacitance value is chosen so that the charges on the two plates are $+ΔQ$ and $-ΔQ$ (i.e., they are equal to the incremental charges that entered the junction through the top and the bottom terminal, respectively, when the reverse bias was changed from V_R to $V_R + ΔV_R$).

Then this capacitance represents the *small-signal capacitance* of the reverse-biased *pn* junction. If this capacitance is denoted by C_j , we have

$$C_j = \frac{\Delta Q}{\Delta V_R} \quad (1.5.21)$$

Using (1.5.20), and refining the definition of C_j by letting the finite differences become differentials, we have

$$C_j = -\frac{dQ_2}{dV_R} \quad (1.5.22)$$

Finally, dividing both sides by the cross-sectional area A , we obtain the *small-signal capacitance per unit area* C'_j/A , denoted by C'_j :

$$C'_j = -\frac{dQ'_2}{dV_R} \quad (1.5.23)$$

Using (1.5.3) and (1.5.18) in the previous equation gives

$$C'_j = \frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{V_R + \phi_{bi}}} \quad (1.5.24)$$

which, using (1.5.3) and (1.5.16), can also be written as

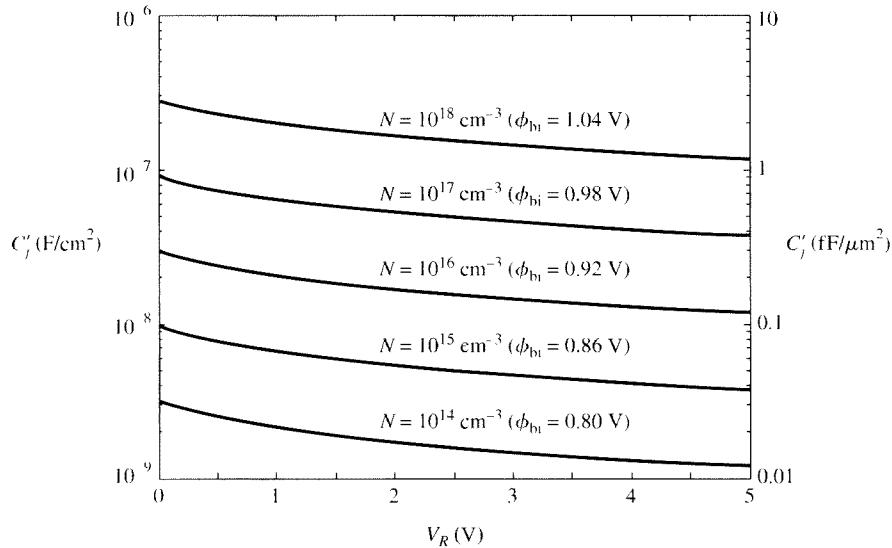
$$C'_j = \frac{\epsilon_s}{d_2} \quad (1.5.25)$$

This equation is plotted in Fig. 1.28 where approximate values for ϕ_{bi} are given⁵ assuming the heavily doped side is degenerate. If C'_{j0} denotes the value of C'_j for $V_R = 0$, (1.5.24) can be easily converted to the following:

$$C'_j = \frac{C'_{j0}}{\sqrt{(V_R/\phi_{bi})+1}} \quad (1.5.26)$$

Again it is emphasized that these results have been derived for a one-sided, abrupt junction. For more complex doping transitions from the *n* to the *p* side, it can be shown⁵ that C'_j can be approximated by

$$C'_j = \frac{C'_{j0}}{\left[(V_R/\phi_{bi})+1\right]^{\alpha_j}} \quad (1.5.27)$$

**FIGURE 1.28**

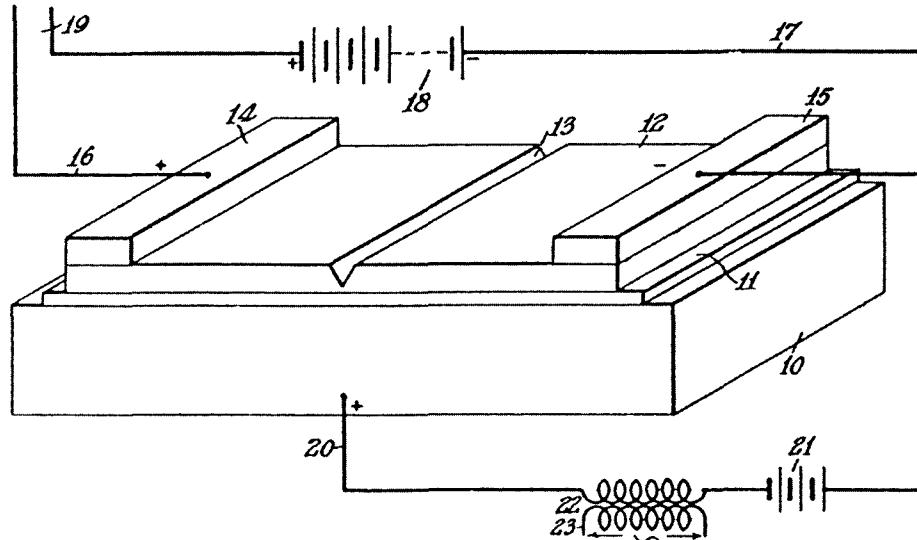
One-sided step-junction small-signal capacitance per unit area vs. reverse-bias voltage, with doping concentration of the lightly doped side as a parameter. The heavily doped side is assumed degenerate; corresponding approximate values of built-in voltage are indicated. $T = 300$ K.

where the expression for C'_{j0} will depend on the details of the transition from n type to p type (these details form the so-called *profile* of the junction). The value of α_j is 1/2 for the one-sided abrupt junction we have discussed, and 1/3 for the linearly graded junction (one in which the net impurity concentration varies linearly with vertical position, being zero at the boundary between the two sides). For practical profiles (which are neither exactly abrupt nor exactly linearly graded), one often chooses the values of the parameters α_j , C'_{j0} , and ϕ_{bi} so as to obtain a best fit of (1.5.27) to measurements. The values obtained this way are said to be “extracted” from measurements.

1.6 Overview of the MOS Transistor

1.6.1 Basic Structure

The basic idea on which the MOS transistor is based is very old and is due to J. E. Lilienfeld, who obtained the first patents in the early 1930s;¹⁸ see Fig. 1.29. Other patents were obtained by O. Heil.¹⁹ Laboratory studies were performed in the late 1940s (see, for example, Ref. 20) but the device remained in the laboratory for over a decade following that. Then in the early 1960s the MOS transistor “took off” following the demonstration of working devices by Kahng and Atalla,²¹ the development of techniques for reliably growing oxides,²² and the establishment of basic theories of operation.^{23–26} A historical review of early work, including the development of practical fabrication techniques, can be found elsewhere.²⁷ Modern fabrication techniques are described in a number of books.^{28–30}

**FIGURE 1.29**

A figure from J. E. Lilienfeld's U.S. patent 1,900,018 (filed March 28, 1928; granted 1933). The structure should be turned upside down to correspond to the orientation used in this book (e.g., Fig. 1.30).

We now offer a simple preview³¹ of the MOS transistor, which will help introduce the detailed material in the rest of this book. A simplified structure of an *n*-channel MOS transistor is shown in Fig. 1.30 (the names *n*-channel and *MOS* will be discussed shortly). The transistor is formed on a *p*-type silicon *body*. Typical doping concentrations for the body are 10^{16} for older or higher-voltage technologies to several times 10^{18} cm^{-3} for modern technologies. The dopant concentration will be assumed uniform throughout the body, except where noted. The body is often also called the *substrate*; however, sometimes the name "substrate" is reserved for the underlying region on which devices sit (this region will be seen in the upcoming discussion of CMOS processes). In this book, the names "body" and "substrate" will be used interchangeably, unless indicated otherwise.

The center part of the structure is covered by an insulator (typically silicon dioxide, which is often referred to simply as *oxide*, or a hafnium-based dielectric for thinner insulators), the thickness of which can be from 1.2 nm for high-performance technologies to several tens of nanometers for high-voltage transistors. The body interface to the oxide is often called the *surface*.

A low-resistivity electrode, called the *gate*, is formed on top of the oxide.³² Although this is made of metal in some advanced fabrication processes, common processes use polycrystalline silicon (*polysilicon*, or *poly*, for short) for the gate. This

³¹ This electrode is actually much taller than suggested by Fig. 1.30 and many other figures in this book, which compress its height to economize on space. Realistic views of MOS transistors are shown in Sec. 1.7.

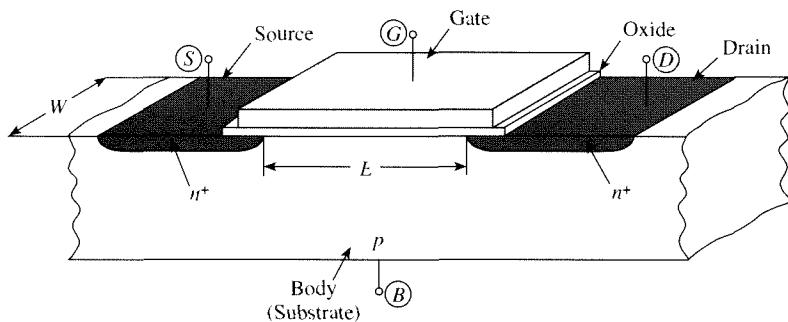


FIGURE 1.30
Simplified structure of an *n*-channel MOS transistor.

material, although silicon, is not a single crystal. Rather, it consists of many regions within each of which there is a regular array of atoms, and this regularity is broken at the boundaries between adjacent regions. The polysilicon material is heavily doped *p* or *n* type (e.g., 10^{20} cm^{-3}). The two regions shown on the sides are formed by implanting donor atoms, with the gate acting as a mask against the implant; this mask receives the donor atoms itself and prevents them from landing under it. Thus the gate is heavily doped and exhibits low resistivity. Donor atoms land in the substrate just outside the “shadow” of the gate and form the two *n*⁺ (heavily doped *n*) regions indicated as *source* and *drain* in Fig. 1.30; these regions can be as shallow as 10 nm in modern technologies. The heavy doping results in low resistivity for these regions, since the abundance of free electrons in them is available for conduction. Subsequent high-temperature fabrication steps cause a diffusion of the dopant atoms both vertically and laterally. This *lateral diffusion* causes the source and drain regions to extend slightly under the gate as shown in the figure. The resulting overlap distance can be as small as 10 nm in modern technologies, larger for older ones and high-voltage processes.

The region between the source and drain is called the *channel*. The channel width *W* and length *L* of individual transistors can vary greatly (from a few tens of nanometers to several hundred micrometers), depending on circuit design needs. In digital circuits, *L* is normally kept at the minimum value possible.

As we will see, if the gate potential is made sufficiently positive with respect to other parts of the structure, electrons can be attracted directly below the insulator (near the “surface” of the body). These electrons can come through the *n*⁺ regions, where they exist in abundance, and can fill the channel between them; for this reason, the device in Fig. 1.30 is referred to as an *n-channel*, or *n*MOS, device (the opposite-type device, called *p-channel*, or *p*MOS, has holes in its channel and will be considered later). The number of electrons in the channel can be varied through the gate potential. This can cause a variation of the “strength” of the connection between the two *n*⁺ regions, resulting in transistor action. If the two *n*⁺ regions are biased at different potentials, the lower-potential *n*⁺ region acts as a source for electrons, which then flow through the channel and are drained by the higher-potential *n*⁺ region. It is thus common to call the lower-potential *n*⁺ region *source* and the higher-potential one *drain*.

The source, gate, drain, and body regions can be contacted through terminals attached to them, as shown schematically in Fig. 1.30. We will denote these terminals by S , G , D , and B , respectively.

The first successful MOS transistor used metal for the gate material and silicon dioxide for the insulator. It is for this reason that the device was named *MOS* transistor, with *MOS* standing for *metal oxide semiconductor*. Other acronyms are *MOST* (for MOS transistor), *MOSFET* (for MOS field-effect transistor), and *IGFET* (insulated-gate field-effect transistor). The latter name originated to distinguish the device from the junction-gate field effect transistor, in which the gate is separated from the rest of the structure by a *pn* junction. Among the three acronyms, *IGFET* is the most general, since it does not specify the material used for the gate or the insulator. This acronym, though, is not in wide use. Today, the popular acronyms *MOST* and *MOSFET* have come to mean the same as *IGFET* and do not imply that metal and silicon dioxide are necessarily used for the gate and the insulator.

The actual values of channel width and length after fabrication (shown as W and L in Fig. 1.30), can be different from those specified during the chip “layout” process, which takes place on a computer screen during chip design. Several reasons contribute to this difference. For example, during high-temperature fabrication steps the source and drain regions diffuse not only downward in Fig. 1.30, but also laterally, thus changing the initial distance between them. In general, we have:

$$W = W_m - \Delta W \quad (1.6.1)$$

$$L = L_m - \Delta L \quad (1.6.2)$$

where W and L are the *actual* channel width and length, respectively (sometimes called *effective*, or *electrical*, width and length), W_m and L_m are the corresponding layout values (often referred to as *mask* or *drawn* values), and ΔW and ΔL are the corrections that must be applied to arrive at the true dimensions. These corrections are normally positive and can be up to several tenths of a micrometer each. In this book, the terms *channel width* and *channel length* will exclusively refer to *actual* values of these quantities (i.e., W and L), unless specifically mentioned otherwise.

As already mentioned, the source-body and drain-body junctions are normally kept reverse-biased. This guarantees that only a very small reverse-bias “leakage” current flows in the *pn* junctions (typically less than 1 pA for minimum-size regions at room temperature; note, though, that this current increases with temperature, roughly doubling every 8°C to 10°C). Other leakage currents can be present, related to the presence of high electrical fields near transistor drains, especially for very small devices; however, through judicious fabrication techniques and choice of operating voltages, these can be held in check. The same can be said for currents through the oxide. All these phenomena will be discussed later (Chap. 5). Until then, we will assume that the voltages between any two terminals of the transistor are sufficiently low not only to prevent breakdown but also to maintain leakage currents at negligible levels.

One last assumption that will hold for all the discussions in this book is that all parts of a transistor operate at the same temperature. Unless otherwise mentioned, this will be assumed to be “room temperature,” taken as 300 K.

1.6.2 A Qualitative Description of MOS Transistor Operation

An *n*-channel MOS transistor with external voltages applied is shown in Fig. 1.31. The condition shown is for a sufficiently positive gate-source voltage V_{GS} , so that positive charges are placed on the gate. These repel the holes from the surface, leaving the latter depleted of holes. The resulting depletion region contains a number of negatively charged acceptor ions, a situation similar to that on the *p* side of the *pn* junction in Fig. 1.21. Further, V_{GS} is assumed to be sufficiently positive to even make the surface attractive to electrons; the latter can easily enter through one or both of the two n^+ regions, where they are in large supply. This situation is called *inversion*, since the surface of the *p*-type body, which normally would have a large concentration of holes, now has plenty of electrons. The layer of electrons at the surface is called an *inversion layer*.

The source and the drain form two *np* junctions with the body. As we have assumed, these are reverse-biased. As has been shown in Fig. 1.21, in a *pn* junction the resulting depletion region extends to both the *n* and *p* sides. However, in the *n*MOS transistor of Fig. 1.31 the part inside the *n* regions is much shallower, since the doping there is much heavier (see Sec. 1.5); thus, that part is not shown for simplicity. In drawing Fig. 1.31, we have assumed that the drain potential is more positive than the source potential. Thus, the reverse bias across the drain-body *np* junction is larger, and the depletion region shown for that junction is deeper. As a result, there are larger numbers of negatively charged acceptor atoms around the drain than there are near the source. This means that fewer electrons are needed in the

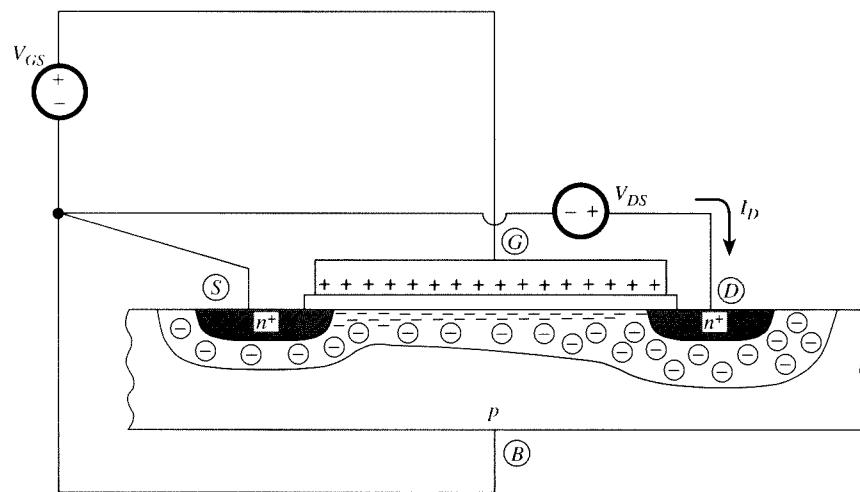


FIGURE 1.31
An *n*-channel MOS transistor under bias in the inversion region.

channel near the drain to balance the positive charges on the gate. It is for this reason that the concentration of electrons is shown to decrease as the drain is approached in Fig. 1.31. The largest electron concentration is found near the source. The larger the value of the gate potential, the more the electrons and the “heavier” the inversion is at that point. A few volts of variation of the gate potential can vary the population of electrons there by several orders of magnitude. Although such variation is continuous, we often say that as the gate potential is raised, we go from *weak* inversion to *moderate* inversion, and eventually to *strong* inversion. It will be seen that, when appropriately done, this division into three regions is convenient, as distinct types of behavior are observed in each of the three regions.

The potential difference V_{DS} between the drain and the source is positive and appears across the inversion layer. It causes electron movement: electrons enter from the source, pass through the channel, and are drained by the drain. The electrons start relatively slowly at the source, and as they approach the drain they speed up; in this way, constant current can be maintained along the channel, although the electron concentration varies along the channel’s length. The movement of negative charges toward the right corresponds to positive current I_D from the drain, through the channel, to the source, as shown by the arrow in Fig. 1.31.

Assume now that V_{DS} starts at zero and is gradually increased. The resulting drain current will also increase, as shown qualitatively in Fig. 1.32. For small values of the drain potential, the effect of the drain potential on the drain current is large; for large-enough values of the drain potential, though, the current gradually tends to saturate as indicated by the part of the curve toward the right in Fig. 1.32. This happens when the drain potential is so large that it drains all electrons that can be supplied by the channel, for a given gate potential. Two regions can thus be distinguished, marked *nonsaturation* and *saturation* in the figure. Although the transition between nonsaturation and saturation is smooth, their boundary is sometimes taken to be at a specific point for convenience.

In Fig. 1.31, a short is shown between the body and source terminals. Assume now that this connection is broken, and that a voltage source is inserted in such a way as to make the body potential negative with respect to the source, as shown in Fig. 1.33. This will cause the population of electrons in the channel to decrease, which will be manifested externally as a decrease in I_D . This phenomenon is referred to as *body effect*. It can be understood qualitatively as follows: The body is a conductive

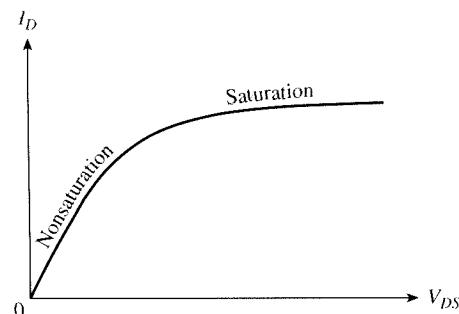
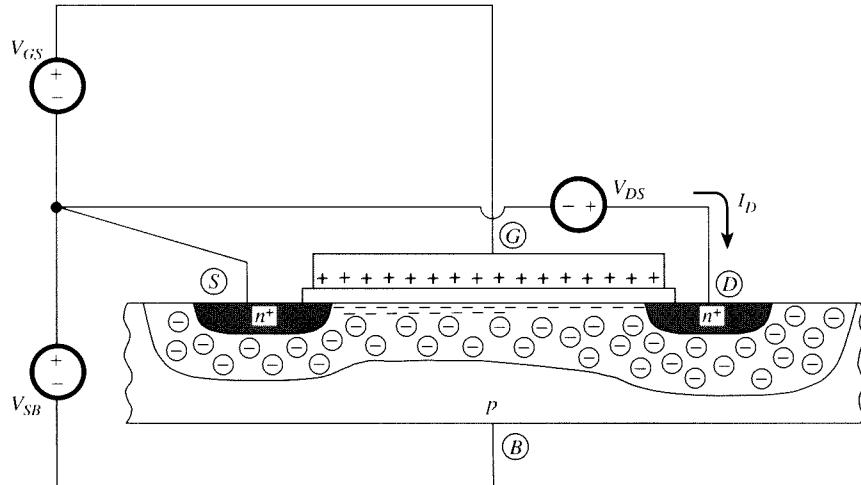


FIGURE 1.32
Typical behavior of drain current vs. drain-source voltage for a fixed gate-source voltage.

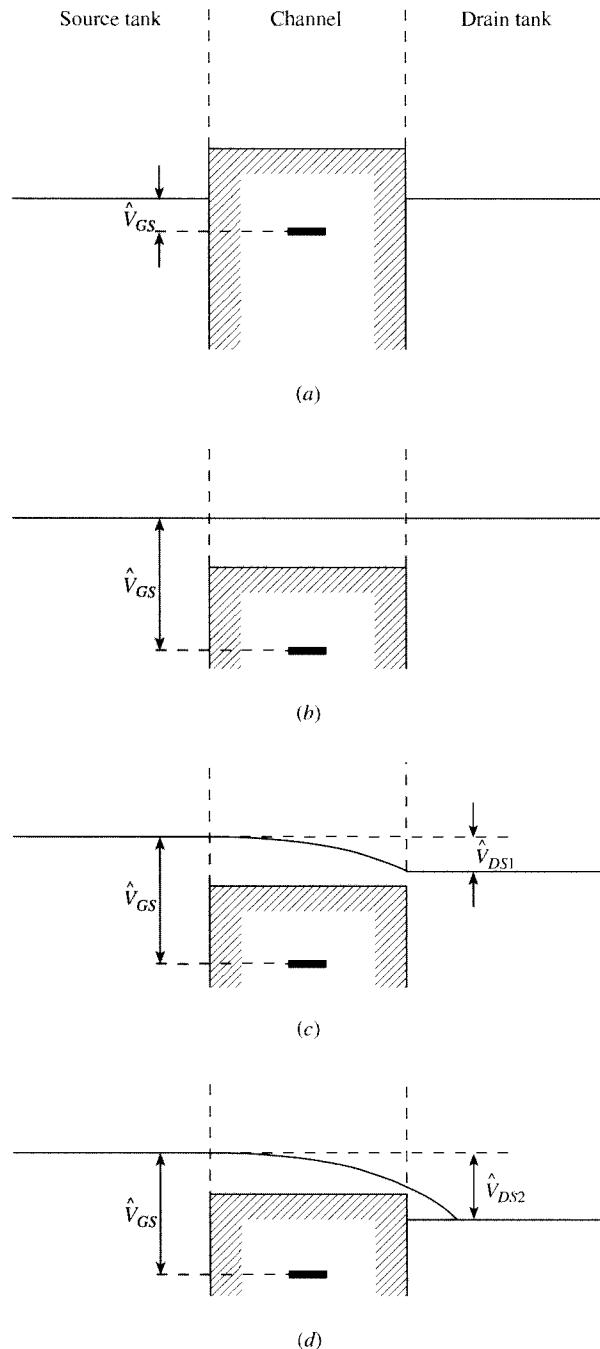
**FIGURE 1.33**

The connection of Fig. 1.31, modified to insert a negative bias between body and source.

structure, separated from the channel by an “insulating” region (the depletion region), just as the gate is separated from the channel by the oxide. The body, then, qualitatively acts as another “gate” (in fact, it is sometimes referred to as a *back gate*): applying a negative potential to it reduces the population of electrons, just as would be the case if a negative potential were applied on the gate. A more rigorous explanation of the body effect must await our detailed discussions in subsequent chapters.

1.6.3 A Fluid Dynamical Analog

To help increase intuition about the MOS transistor, we now present a fluid dynamical analog.^{31–33} In our analog, electrons correspond to water molecules. Electric current corresponds to net flow of water. The source and drain correspond to two large tanks, filled with water up to a certain level each. As water is moved from one tank to the other, the level of water in each tank is maintained constant through external means. This corresponds to the potentials at the source and drain of a transistor being held constant despite current flow. We will assume that the level of water in the “source” tank is fixed throughout our discussion. The two tanks are separated by a piston corresponding to the gate, as shown in Fig. 1.34a. A “handle,” shown by a thick short line at the center, is attached to the piston. The depth at which this handle is with respect to the “source” level will be denoted by \hat{V}_{GS} and corresponds to the gate-source voltage V_{GS} . In this water analog, quantities corresponding to potentials increase *downward* in the figure. In particular, the lower the handle (and thus the piston), the larger the value of \hat{V}_{GS} . Similarly, the level of water in the “drain” tank with respect to the source level will be denoted by \hat{V}_{DS} and corresponds to the drain-source voltage. The lower the “drain” level, the larger the value of \hat{V}_{DS} . The level of the water can be thought of as representing potential energy for electrons. As (1.2.17) indicates, the higher the potential, the lower the potential energy will be, due to the electrons’ negative charge.

**FIGURE 1.34**

Fluid dynamical analog of MOS transistor operation in cutoff and strong inversion.³² (a) Cutoff; (b) with channel strongly inverted, but no current since $V_{DS} = 0$; (c) nonsaturation with nonzero V_{DS} ; (d) saturation. The water level corresponds to electron potential energy. (Figure reprinted with permission from Y. Tsividis, *Mixed Analog-Digital VLSI Devices and Technology*, World Scientific, 1996.)

In Fig. 1.34a, \hat{V}_{GS} is very low. The “channel” is cut off, and no communication exists between source and drain. Now assume \hat{V}_{GS} is increased considerably (i.e., the piston is moved *downward*), as shown in Fig. 1.34b. The channel is now filled with the water. Communication between the source and drain is now possible, but no flow is observed in the steady state, since $\hat{V}_{DS} = 0$. Now if \hat{V}_{DS} is increased as in Fig. 1.34c, flow is observed as shown. The flow increases as \hat{V}_{DS} is increased further, until “saturation” is reached; it is easy to see that in saturation (Fig. 1.34d), further increases in \hat{V}_{DS} do not affect the flow. The water enters slowly from the source and, as it approaches the drain, moves faster, in order to maintain a fixed flow despite the fact that the amount of water is less near the drain. Again, an analogy exists here to the simplified picture of transistor operation in saturation. Increasing the drain potential lowers the electron potential energy at the drain; thus, electrons tend to flow from a region of high potential energy (the source) to a region of lower potential energy (the drain).

It is evident from the figure that, for a given \hat{V}_{DS} , the flow will increase if \hat{V}_{GS} is increased and will decrease if \hat{V}_{GS} is decreased. If \hat{V}_{GS} is decreased to the point

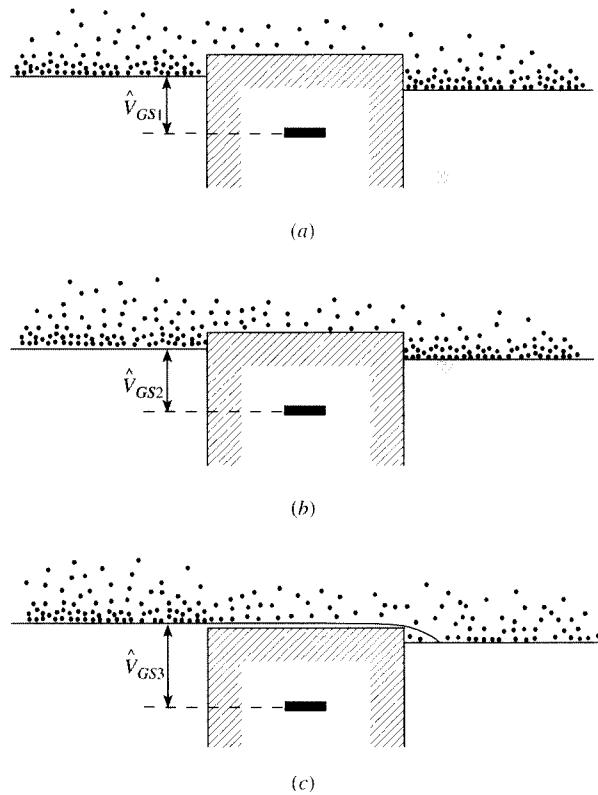


FIGURE 1.35

Fluid dynamical analog of MOS transistor operation in (a and b) weak inversion and (c) moderate inversion.^{31,33} (Figure reprinted with permission from Y. Tsividis, *Mixed Analog-Digital VLSI Devices and Technology*, World Scientific, 1996.)

that the top surface of the piston is at, or above, the source surface level, direct flow of the fluid is prevented. Water molecules can still, though, flow from left to right. This can be seen if one takes into account the diffusion of water vapors,³³ as illustrated in Fig. 1.35a. This situation corresponds to weak inversion. As shown, the vapor concentration is maximum at the water surface and decreases as one moves vertically away from it (in fact, it can be shown that it decreases exponentially). Thus, if the water surface in the drain tank is lower, as shown in Fig. 1.35a, at any given horizontal plane above the piston the vapor concentration will be decreasing as one goes from left to right over the channel. Vapor thus diffuses from left to right, carrying a minute “current,” even if the position of the middle shaft is slightly above the source water surface. However, a very small drop in the position of the middle shaft (i.e., an increase in \hat{V}_{GS}) can drastically increase the “current,” since the vapor concentration over the source, at points slightly above the level of the shaft, will increase exponentially (Fig. 1.35b). One thus might expect an exponential dependence of the “current” on \hat{V}_{GS} .

For large \hat{V}_{DS} (low drain water level), the concentration on the right at the level of the shaft’s surface becomes negligible, and the “current” assumes a value largely independent of \hat{V}_{DS} , reaching saturation.

If the top surface of the piston is only *very* slightly *below* the source level, both the fluid and its vapors can contribute significantly to the flow (Fig. 1.35c). Notice that vapors would also exist in the case of very positive \hat{V}_{GS} shown in Fig. 1.34. However, they were not considered there because their contribution was negligible in comparison to the much larger flow of liquid water.

1.6.4 MOS Transistor Characteristics

Let us finally look at a set of typical transistor characteristics. In Fig. 1.36a, we show plots of I_D (on a *logarithmic* axis) vs. V_{DS} , with V_{GS} as a parameter, and with the body shorted to the source, for a specific transistor. The logarithmic axis is used to reveal the several orders of magnitude of I_D , over which control is possible through V_{GS} . The regions of inversion have been roughly marked in terms of V_{GS} (the limits between the regions will be discussed later in this book). As V_{GS} or V_{DS} is reduced, the current in the channel can eventually become so small that it is masked by the leakage current of the reverse-biased drain-body junction (or even the leakage from the inversion layer to the substrate). This can be seen near the bottom of the plot. The boundary between the regions of nonsaturation and saturation is indicated by a broken line; saturation is to the right of that line.

The same characteristics are shown in Fig. 1.36b using a linear I_D axis. Obviously this type of plot does not do justice to weak inversion. This region was in fact unknown for several years, until suspicious “leakage currents” in dynamic memories prompted researchers to take a more detailed look. The weak inversion region is now rather well understood and is an important region of operation for many applications, especially those for which very small power dissipation, or low voltage operation, is desired.

These plots have been obtained for a long-channel device. For this device, the increase of I_D with V_{DS} in saturation is not easily seen on the scale used. For devices

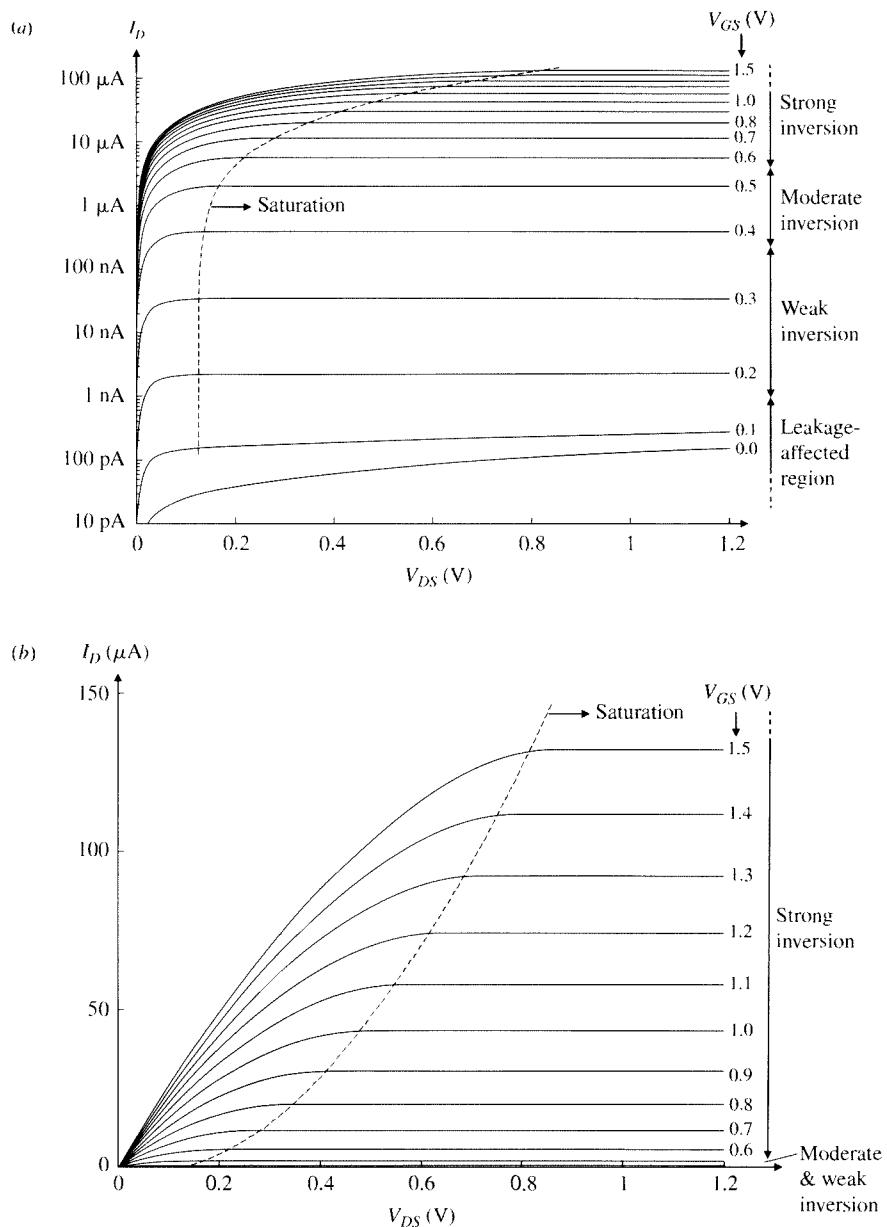


FIGURE 1.36
 I_D - V_{DS} characteristics with V_{GS} as a parameter for a specific device in the connection of Fig. 1.33 and $V_{SB} = 0$. (a) Logarithmic I_D axis; (b) linear I_D axis.

with short channels, the effect of V_{DS} on I_D in saturation is more evident. Devices with short channels will be considered in detail later in this book.

A careful look at transistor characteristic curves reveals a distinctly different behavior in each region of inversion. Let us begin with the weak inversion region, which is marked in Fig. 1.36a. In this region, the current turns out to be due to the diffusion of carriers (Sec. 1.3.3) and corresponds to the case of Fig. 1.35a and b. The spacing of consecutive curves in Fig. 1.36a can be seen to be almost equal, for equal V_{GS} increments. Given that the vertical axis is logarithmic, this means that I_D is, for all practical purposes, exponentially related to V_{GS} in weak inversion (Prob. 1.14). This behavior reminds one of the type of dependence of collector current on base-emitter voltage in a bipolar transistor. Typically, the range of V_{GS} values over which weak inversion behavior is observed can be a few tenths of a volt; nevertheless, because of the exponential behavior in this region, this is sufficient for a drain current variation of three or four orders of magnitude at room temperature.

In strong inversion, marked in Fig. 1.36a and b, the current turns out to be due to drift (Sec. 1.3.2). This mode of operation corresponds to the case in Fig. 1.34. In the saturation part of strong inversion, I_D is found to be approximately quadratic in V_{GS} for large-dimension devices. This can be verified from the spacing of consecutive curves in Fig. 1.36b (Prob. 1.14).

In moderate inversion (marked in Fig. 1.36a and b), both drift and diffusion currents contribute significantly. This situation corresponds to Fig. 1.35c. In this region, I_D is neither exponential nor polynomial; rather its behavior changes gradually from one form of functional dependence to the other, as V_{GS} is raised by a few tenths of a volt. Within this region, the current can vary by a couple of orders of magnitude.

1.7 Fabrication Processes and Device Features

CMOS Processes Both *n*MOS and *p*MOS devices are available on chips made using CMOS (complementary MOS) technology. Partial views of several types of CMOS chips are shown in Fig. 1.37. In Figure 1.37a, we show the result of the so-called local oxidation of silicon (LOCOS) CMOS process.^{28–31,34} We present this process for historical purposes, and because it is the most common process for technologies with minimum transistor channel lengths of 0.25 μm and greater, which are still in use for high-voltage applications. Characteristic of this process is the gradual transition of the oxide from thin (under the gates) to thick, resulting in the so-called bird's beak shape. The thick oxide is needed so that wiring running above it (not shown) cannot accidentally create a parasitic channel underneath it. Heavily doped "channel stop" regions (not shown) are used under the thick oxide to further suppress the formation of parasitic channels.

Figure 1.37b shows a different type of CMOS process, known as *shallow-trench isolation* (STI).^{28–30} This is the most prevalent isolation process for technologies with minimum channel lengths of 0.18 μm and below. As shown, a "trench" filled with oxide isolates devices from one another. This trench can be formed with vertical walls, and does not have the lateral spread of the bird's beak of the LOCOS process, thereby allowing the devices to be packed closer to each other.

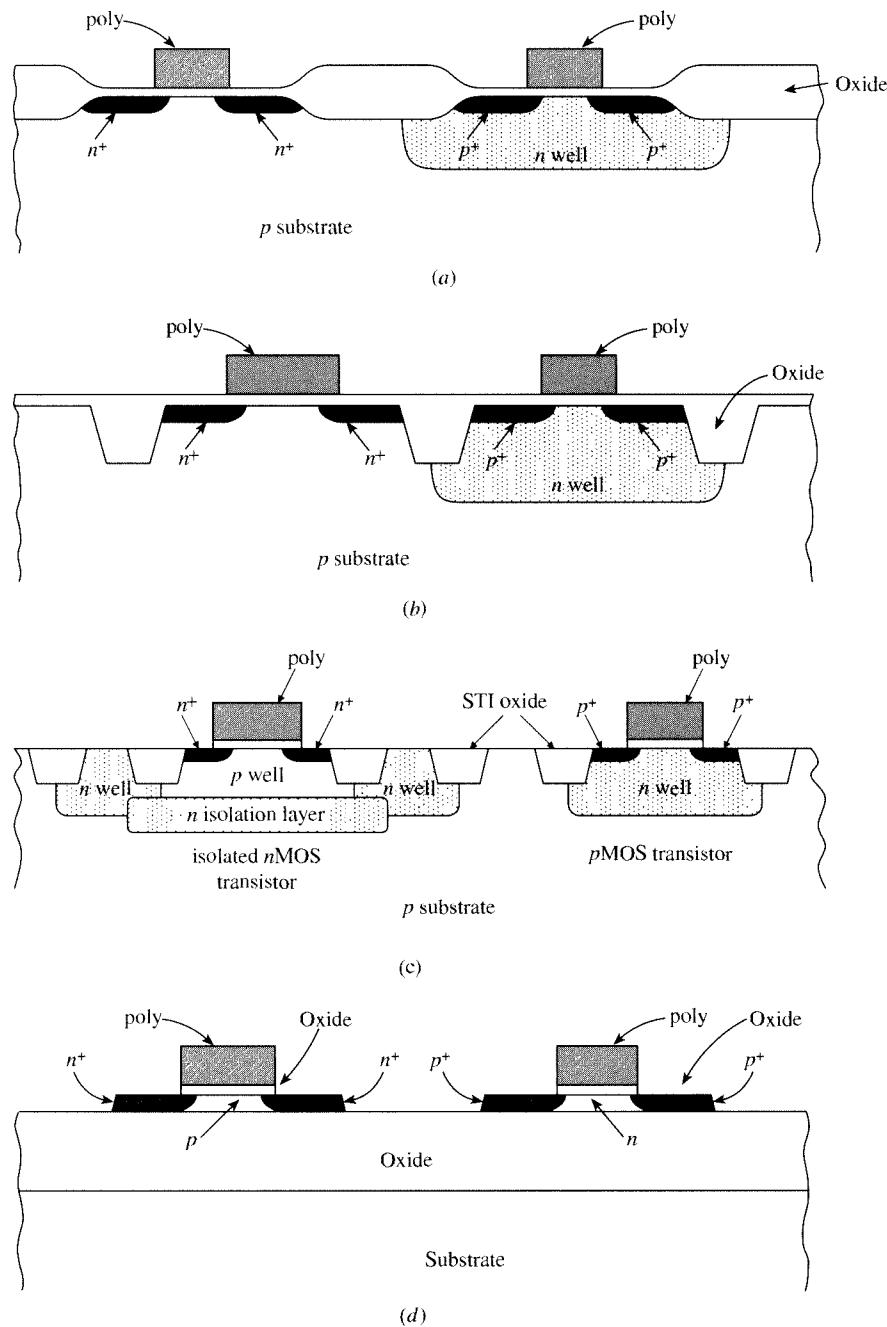


FIGURE 1.37
CMOS processes (simplified). (a) LOCOS, (b) single-well STI, (c) triple-well STI, (d) SOI. Contacts to the various regions are not shown.

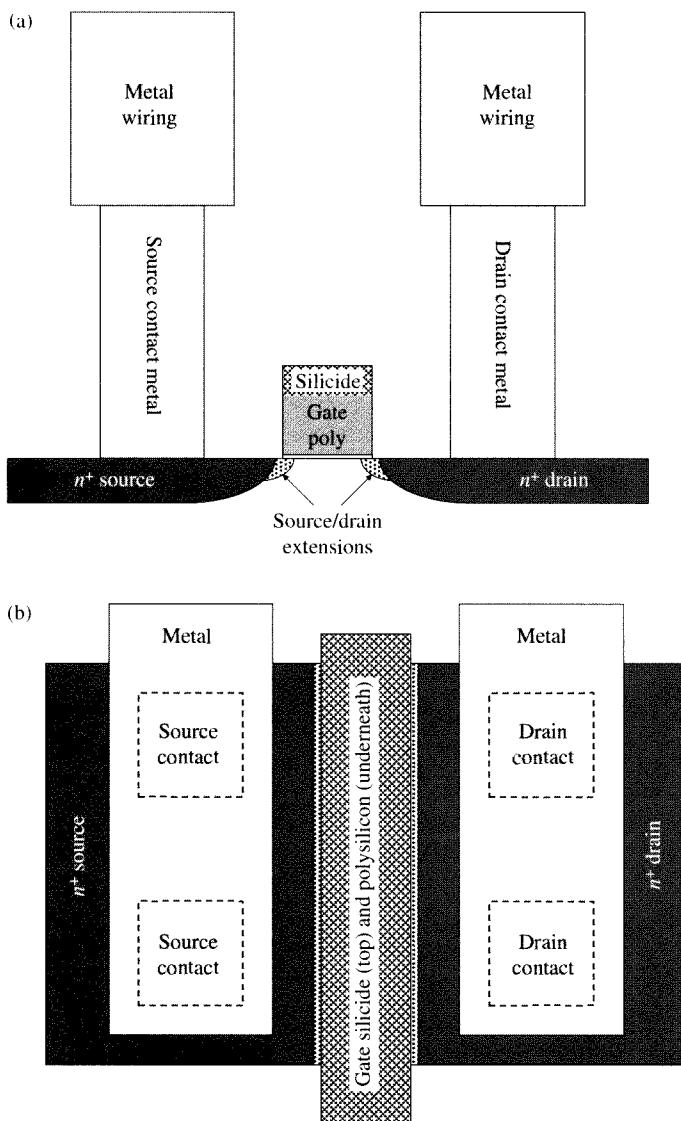
In the CMOS processes shown in Figs. 1.37a and b, the *p* substrate is common to all *n*MOS devices on the chip and serves as an isolation region between them. The *pn* junctions formed between the *p* substrate and the *n* sources and drains are normally kept reverse-biased, by connecting the substrate to the most negative potential in the circuit. The *p*MOS devices are contained within *n*-type wells as shown; for this reason, these versions of the CMOS fabrication process are referred to as *n-well processes*. To keep the junctions formed between the *n* well and the *p* sources and drains reverse-biased, the well must be connected to an appropriate potential, which, for digital circuits, is the most positive potential available in the circuit.

To allow individual *n*MOS and *p*MOS device optimization, most modern processes start with a lightly doped substrate, and separate *n* wells and *p* wells are then formed, in which, respectively, *p*MOS and *n*MOS transistors are subsequently built. This is called a *twin-tub* or a *twin-well* process. Triple-well processes are also in use; these make possible connections to the bodies of individual *n*MOS devices. An STI triple-well process is shown in Fig. 1.37c.

Extra care is taken to avoid a condition known as “latchup,”³⁴ in which some of the diodes formed by *pn* junctions may become accidentally forward-biased, thereby initiating bipolar transistor action and activating a type of positive feedback involving parasitic *pnpn* structures (e.g., see Fig. 1.37a). If latchup occurs, such structures can be locked into an on condition, and excessive currents can flow, which can interfere with proper operation and can destroy the chip. A sufficient distance between the various regions must be maintained to prevent latchup. It is also desirable to keep the substrate resistivity low to prevent the formation of significant voltage drops due to parasitic currents in it, because such voltage drops can give rise to, and can sustain, latchup. However, this requires a high substrate doping concentration, which can have an undesirable effect on MOS transistor operation, as we will see. This conflict is resolved in some processes by using a shallow, lightly doped *p*-type “epi” region (so called because it is formed by a process called *epitaxy*^{28–30}), on top of a heavily doped (*p*⁺) substrate. Thus, the body of the MOS devices is made of lightly doped material, whereas the material underneath it is heavily doped to keep the parasitic resistances (and also the current gains of the parasitic bipolar devices) low.

A high degree of isolation and more dense packing of devices is possible in silicon on insulator (SOI) processes,³⁵ in which each device is formed in its own separate region on top of an oxide layer and is thus completely isolated from its neighbors, as shown in Fig. 1.37d. Although SOI technologies offer improved performance over STI technologies, they are more costly to manufacture and so are used primarily for high-performance microprocessors. In SOI processes, the substrate of each device is normally left floating, and this requires special modeling considerations. In this book, we do not cover the modeling of floating-substrate SOI devices; the interested reader is referred to the literature.³⁵ Also, we do not discuss devices made using special processes for high-voltage and power applications.³⁶

Other processes, making different device structures possible, are under development. For example, it has been proposed to use undoped or lowly doped multigate structures, in which the gate surrounds the channel on three sides, rather than being simply on top of it (Chap. 5). There is considerable research on the technology and modeling of such devices.³⁷ At the time of this writing such technology is not mature

**FIGURE 1.38**

A device drawing showing realistic relative dimensions for the various features; (a) cross section, (b) top view.

enough, so multigate structures are not covered in this book. However, the knowledge provided in this book can be applied toward the understanding of such structures, as described in the research literature.

Detailed Device Features In the various drawings of MOS transistors so far, we have only shown their most basic features for simplicity. Also, we have shown the

gate as very thin (a practice used in much of this book, to economize on space). A more realistic drawing of a modern MOS transistor is shown in Fig. 1.38. A cross section is shown in *a* and a corresponding top view in *b*. As seen in *a*, the gate height is much greater than the insulator thickness. In this device, the gate is made of polysilicon; to lower the gate resistance, its top is covered by a layer of “silicide,” a material formed by the reaction of a metal (e.g., titanium or cobalt) and silicon. Keeping the gate resistance low is essential for high-frequency operation, as this resistance conspires with parasitic capacitances to form RC circuits that can limit the device speed (Chaps. 6–9).

In Fig. 1.38*a*, regions called “source/drain extensions” are used as bridges from the deeper source and drain regions to the “intrinsic” device formed under the gate. The use of the shallow extension regions limits undesirable two-dimensional field effects under the channel (Chap. 5), whereas the deeper regions allow low-resistance connections to the external world. Contacts to the source and drain regions are made as shown, through oxide cuts called *contact windows*. Contacts are also made to the gate, as well as to the substrate and well regions (not shown). To connect the transistors to form circuits, multiple (up to nine) layers of metal, separated by insulating dielectrics, are then formed (part of the lower metal layer, connecting to the source and drain contacts, is shown in Fig. 1.38*a*). The metal and insulator layers are interleaved, and different metal layers are connected by “vias.” An oxide layer (not shown), called a *passivation layer*, is deposited on top of the devices to protect them from their environment.[†]

The gate is contacted through contact windows as well. This can be done on one side as shown in Fig. 1.39*a* or, to lower the gate resistance, on both sides, as shown in Fig. 1.39*b*. In the latter case, the two gate contacts are shorted together, thus effectively placing the two gate halves (above and below the horizontal axis of symmetry) in parallel and reducing the gate resistance by a factor of 4.

Actual MOS transistors in high-performance fabrication processes are even more complicated than the structure of Fig. 1.38. Figure 1.40 shows scanning electron microscope (SEM) images of the cross sections of modern *n*MOS and *p*MOS transistors (left and right, respectively), on an Intel 45-nm process.³⁹ The rectangle in the center of each image is the gate; in these devices the gate is made of metal rather than polysilicon, making possible very low gate resistance. The doping concentration in the body is generally nonuniform in both vertical and lateral directions, to optimize device performance (Chap. 9). In the *n*MOS device, the dark areas on the upper left and upper right are contacts to the source and drain regions. The material between the contacts and the gate is an insulating dielectric. The striping seen in this region is from layers that are introduced to deliberately create mechanical expansive (or “tensile”) stress, which enhances *n*MOS transistor performance. In the *p*MOS device, the source and drain regions are elevated with respect to the silicon surface under the gate, and two “chevron”-shaped regions can be seen pointing toward the center.

[†] The steps up to contact formation are termed front-end-of-line (FEOL) processes. The steps done after contact formation, involving the metal layers, the vias, and the insulating and passivation layer, are termed back-end-of-line (BEOL) processes.

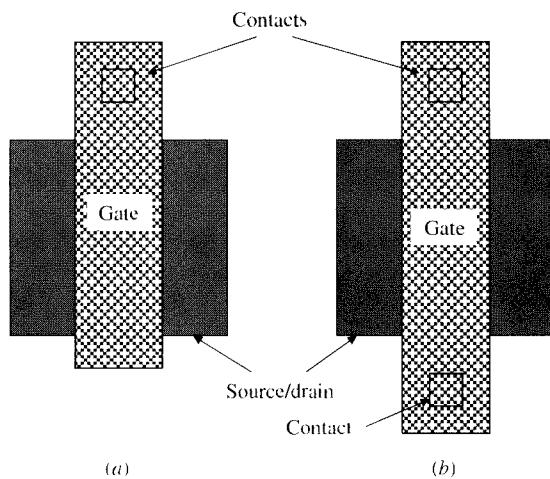


FIGURE 1.39
Contacting the gate (a) on one side, (b) on both sides.

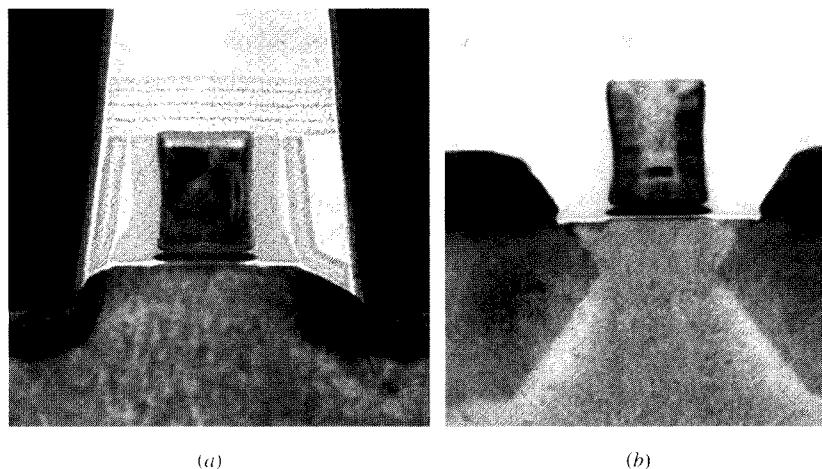


FIGURE 1.40
SEM images of (a) an nMOS and (b) a pMOS transistor in Intel's 45-nm technology³⁹ [(a) Copyright © 2008 by IEEE. (b) Courtesy of Intel Corporation].

These regions are formed by depositing a silicon-germanium alloy crystalline material. The lattice spacing (distance between neighboring atoms) of SiGe is greater than that of silicon, so the SiGe source and drain regions compress the channel region. This significantly enhances the performance of the pMOS device. Mechanical stress (expansive or compressive) will be discussed in Chap. 9.

Scaling The historical trend for CMOS technologies has been a continuous reduction in minimum channel length and gate oxide thickness over time (this process is called “technology scaling,” and it also involves reductions in the source and drain junction depth, an increase in the dopant concentration of the body, and a reduction of maximum operating voltages for the transistors). Such scaling leads to increased

performance and reduced cost per transistor (and also per integrated circuit, or “die”); even though total wafer processing cost increases with technology scaling, this is more than offset by the increased number of chips per wafer from the reduction in die size).

The minimum possible feature dimensions for a given technology are determined by the capabilities of fabrication equipment and by reliability and manufacturability considerations. In the past, technology generations were typically named after the minimum dimension of some structure that can be formed in the technology. By convention this has been the “half-pitch” of a dynamic random access memory (DRAM) bit cell, which is one-half of the center-to-center spacing between DRAM cells in a memory and is different from the minimum channel length. For example, in “45-nm technology,” the minimum effective channel length is significantly smaller than 45 nm (see the discussion in Sec. 1.6.1 concerning effective gate length).³⁸ From each technology generation to the next, it is customary to reduce area by a factor of about two, which means that linear dimensions are reduced by a factor of about $\sqrt{2}$; thus, for example, after 65-nm technology came 45-nm technology, named so because $65/\sqrt{2} \approx 45$.

As technology advances and devices are made smaller and smaller, it has historically been attempted to make the oxide under the gate thinner and thinner as well. However, very thin oxides do not behave as good insulators; there is current flow through them, due to a phenomenon called *tunneling* (Chap. 5). The resulting tunneling current increases drastically with decreasing gate oxide thickness and has become a limiting factor to technology scaling. The solution has been to move to different gate dielectric materials (so-called high- k dielectrics) with higher permittivities; this allows the gate dielectric thickness to be increased, hence reducing the tunneling current, without reducing the effectiveness of the control of the gate potential in attracting electrons below the insulator. All dielectric thicknesses in this book will be specified as an equivalent thickness of silicon dioxide; it should be kept in mind that the physical thickness for high- k dielectrics is greater than this. The combination of high- k dielectrics and metal gates does not alter the fundamental structure or behavior of MOS transistors; it has re-engineered the manufacturing processes so that the industry can keep moving down its historic scaling path.

The continued shrinking of the minimum channel length of MOS transistors has caused several previously negligible effects to now have a significant impact on device performance, enough to cause circuits to fail to operate if these effects are not taken into account during circuit design. These include an increased doping concentration near well edges (a by-product of fabrication techniques used), as well as mechanical stress introduced by the STI regions. These effects, which will be discussed in Chap. 9, make the characteristics of transistors dependent on their spacing from adjacent well edges and isolation regions.

1.8 A Brief Overview of This Book

In order to derive quantitative results for MOS transistor behavior, the study of many phenomena becomes necessary. In the following chapters, we will make clear which phenomena are responsible for which behavior, by studying them separately.

In Chap. 2, we will consider only the middle part of the transistor in Fig. 1.30, i.e., a gate-oxide-body structure with only gate and body terminals. Several fundamental mechanisms and results will be studied for this structure. In Chap. 3, we will add one more terminal (corresponding to the source in Fig. 1.33). In the resulting three-terminal structure, the third terminal can be used to contact the inversion layer. We will consider the additional phenomena that occur because of the presence of this third terminal, notably the body effect. These two chapters cover phenomena whose manifestation does not rely on the presence of current flow; this approach makes the study of these phenomena easy. Nevertheless, the phenomena studied are also present when current is flowing, and thus the results derived in Chaps. 2 and 3 will be found to be directly applicable to the cases discussed in the rest of the book.

In Chap. 4, we will add the fourth terminal (the drain), which will make possible the flow of current in the complete transistor. The transistor will be assumed to be of large dimensions and on a uniform substrate. It will be seen that one-dimensional and pseudo-one-dimensional approaches are sufficient for this case. A fundamental approach to deriving expressions for the drain current will be discussed, one that results in a model valid in all regions of operation. Simpler, approximate models valid in particular regions will also be considered, as will several refinements.

In Chap. 5, we will consider what happens when the dimensions of the transistor are made small. We will see how our description of the various phenomena, and the resulting models, must be modified for small devices. Here we will see the limits of some analytical approaches, and we will explain the need for two-dimensional and pseudo-two-dimensional approaches.

Up to the end of Chap. 5, our interest will be to study the dc drain current as a function of the terminal voltages, the latter being assumed constant. In Chap. 6, we will allow those voltages to become functions of time, and we will consider the resulting varying charges and terminal currents. Such dynamic behavior cannot be described adequately by the results of the previous chapters. Depending on the speed of variation of the terminal voltages, we will see that models of differing complexity become possible. We remark that results in this and the previous chapters deal with *total* voltages and currents only.

The next two chapters deal with the relation of *small increments* of the terminal voltages to the resulting increments of the terminal currents. This “small-signal” behavior is of importance in the design not only of analog circuits, but also of digital circuits of high performance. Chapter 7 discusses approaches that are simple and satisfactory at low and medium frequencies, whereas Chap. 8 discusses additional considerations that become necessary at very high frequencies.

In Chap. 9, we will allow for substrate nonuniformity, both vertical and horizontal, made possible by the process of “ion implantation.” We will discuss how device performance can be enhanced by this process and how it can be modeled. We will also discuss several structural effects, such as those due to well-edge proximity and stress, and we will consider the effects of parameter variability on device performance.

The last chapter of the book, Chap. 10, considers the process of “compact” modeling in which the several phenomena discussed separately must be put together

in order to describe the transistor in a variety of situations. Such models are indispensable in computer-aided circuit design. Several considerations in the development of compact models, as well as ways to evaluate such models, will be discussed.

References

1. R. B. Adler, A. C. Smith, and R. L. Longini, *Introduction to Semiconductor Physics*, SEEC Series, vol. 1, John Wiley, New York, 1964.
2. P. E. Gray, D. DeWitt, A. R. Boothroyd, and J. F. Gibbons, *Physical Electronics and Circuit Models of Transistors*, SEEC Series, vol. 2, John Wiley, New York, 1964.
3. J. P. Colinge and C. A. Colinge, *Physics of Semiconductor Devices*, Springer, New York, 2002.
4. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley, New York, 1967.
5. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
6. R. S. Muller, T. Kamins, and M. Chan, *Device Electronics for Integrated Circuits*, 3rd ed., John Wiley, New York, 2002.
7. M. Shur, *Introduction to Electronic Devices*, John Wiley, New York, 1996.
8. H. C. Casey, Jr., *Devices for Integrated Circuits*, John Wiley, New York, 1999.
9. E. S. Yang, *Microelectronic Devices*, McGraw-Hill, New York, 1988.
10. J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
11. R. F. Pierret, *Advanced Semiconductor Fundamentals*, 2nd ed., Prentice Hall, New York, 2002.
12. G. W. Neudeck, *The PN Junction Diode*, 2nd ed., Prentice Hall, New York, 1988.
13. The NIST Reference on Constants, Units, and Uncertainty [online]. Available at <http://physics.nist.gov/cuu/Constants/index.html>. Accessed Sept. 29, 2009.
14. W. E. Beadle, J. C. C. Tsai, and R. D. Plummer (editors), *Quick Reference Manual for Silicon Integrated Circuit Technology*, Wiley-Interscience, New York, 1985.
15. B. E. Deal, E. H. Snow, and C. A. Mead, "Barrier energies in metal-silicon dioxide-silicon structures," *Journal of Physics and Chemistry of Solids*, vol. 27, p. 1873, 1966.
16. S. Kar, "Determination of Si-metal work function differences by MOS capacitance technique," *Solid-State Electronics*, vol. 18, pp. 169–181, 1975.
17. E. H. Rhoderick, *Metal-Semiconductor Contacts*, Oxford University Press, London, 1978.
18. J. E. Lilienfeld, U.S. Patents 1,745,175 (1930, filed October 26, 1926), 1,877,140 (1932, filed December 8, 1928) and 1,900,018 (1933, filed March 28, 1928).
19. O. Heil, British Patent 439,457 (filed and granted 1935).
20. W. Shockley and G. L. Pearson, "Modulation of conductance of thin films of semiconductors by surface charges," *Physical Review*, vol. 74, pp. 232–233, 1948.
21. D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field induced devices," *Solid-State Device Research Conference*, Pittsburgh, June 1960.
22. E. H. Snow, B. E. Deal, A. S. Grove, and C. T. Sah, "Ion transport phenomena in insulating films," *Journal of Applied Physics*, vol. 36, pp. 1665–1673, 1965.
23. H. K. J. Ihantola, "Design theory of surface field-effect transistor," Stanford Electronics Laboratories, *Technical Report No. 1161-1*, Stanford University, California, September 1961.
24. S. R. Hofstein and F. P. Heinman, "The silicon insulated-gate field effect transistor," *Proceedings of the IEEE*, vol. 51, pp. 1190–1202, 1963.
25. H. K. J. Ihantola and J. L. Moll, "Design theory of a surface field-effect transistor," *Solid-State Electronics*, vol. 7, pp. 423–430, 1964.
26. C. T. Sah, "Characteristics of the metal-oxide-semiconductor transistors," *IEEE Transactions on Electron Devices*, vol. ED-11, pp. 324–345, 1964.
27. C.-T. Shah, "Evolution of the MOS transistor—From conception to VLSI," *Proceedings of the IEEE*, vol. 76, no. 10, pp. 1280–1326, 1988.
28. S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd ed., John Wiley, New York, 2001.
29. W. S. Ruska, *Microelectronics Processing*, McGraw-Hill, New York, 1987.
30. S. K. Ghandi, *VLSI Fabrication Principles—Silicon and Gallium Arsenide*, 2nd ed., John Wiley, New York, 1994.

31. Y. P. Tsividis, *Mixed Analog-Digital Devices and Technology—An Introduction*, World Scientific, Singapore, 2002.
32. C. H. Sequin, “A fluid model for visualizing MOS transistor behavior,” Sec. 1.15 in C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, Mass., 1980, pp. 29–33.
33. C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, Mass., 1989.
34. J. P. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology: Fundamentals, Practice, and Modeling*, Prentice Hall, New York, 2000.
35. J. P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI*, 3rd ed., Springer, New York, 2004.
36. D. A. Grant and J. Gowar, *Power MOSFETs—Theory and Application*, John Wiley, New York, 1989.
37. J. P. Colinge, *FinFETs and Other Multi-Gate Transistors*, Springer, New York, 2007.
38. The International Technology Roadmap for Semiconductors, 2007 [online]. Available at <http://www.itrs.net/Links/2007ITRS/Home2007.htm>. Accessed Sept. 29, 2009.
39. C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf, P. Hentges, S. Jaloviar, S. Joshi, J. Klaus, K. Kuhn, D. Lavric, M. Lu, H. Mariappan, K. Mistry, B. Norris, N. Rahhal-orabi, P. Ranade, J. Sandford, L. Shifren, V. Souw, K. Tone, F. Tambwe, A. Thompson, D. Towner, T. Troeger, P. Vandervoorn, C. Wallace, J. Wiedemer, and C. Wiegand, “45 nm High-k * metal gate strain-enhanced transistors,” *Proceedings of the Symposium on VLSI Technology*, pp. 128–129, Honolulu, June 2008.

Problems

- 1.1.** Calculate the electron and hole concentrations in *p*-type silicon with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$ at $T = 280 \text{ K}$, 300 K , and 330 K .
- 1.2.** This problem is intended to give an idea of the order of magnitude involved for the various quantities discussed in Sec. 1.3. Consider the bar of Fig. 1.12 with $a = 100 \mu\text{m}$, $b = 10 \mu\text{m}$, $c = 2 \mu\text{m}$, and *n*-type silicon with doping concentration of 10^{15} cm^{-3} ; $V = 1 \text{ V}$. Find the value of the conductivity, the mobility, the conductance, the sheet resistance, the total mobile charge, the mobile charge per unit area, the field intensity, the drift velocity, the transit time, and the current.
- 1.3.** Prove that (1.3.23) is valid for the case of Fig. 1.15c.
- 1.4.** (a) Prove mathematically that (1.3.10) and (1.3.15) are valid even if the electron concentration varies with depth as long as it is uniform horizontally.
(b) Prove mathematically that (1.3.20) and (1.3.22) are valid for the conditions stated in the paragraph following (1.3.20).
- 1.5.** Verify (1.2.19).
- 1.6.** Calculate the contact potential of copper to *n*-type silicon with $N_D = 10^{17} \text{ cm}^{-3}$ at 300 K , assuming that the work function potential for copper is 4.5 V .
- 1.7.** A voltmeter with both of its leads made out of a metal *X* measures a voltage *V* across a battery. Show that if one of the voltmeter’s leads is replaced by a different material *Y*, the electrostatic potential across the voltmeter will not be affected.
- 1.8.** Assume that the two terminals of a voltage source are made out of different materials. Show that (1.4.9) and (1.4.10) are valid if V_{source} is defined as the voltage measured by an ideal voltmeter when it is attached to the terminals of the source.
- 1.9.** For a two-sided step *pn* junction with neither of its sides degenerate, prove that the built-in potential ϕ_b is given by $\phi_b \ln(N_A N_D / n_i^2)$, and that the total length of the depletion region and the charge per unit area are given by formulas provided in the footnote to (1.5.11).

- 1.10. Plot the junction capacitance vs. reverse-bias voltage (from 0 to 2 V) for a silicon n^+p junction of area $200 \mu\text{m}^2$ with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$. Assume $\phi_{Fn} = -0.56 \text{ V}$.
- 1.11. Consider a semiconductor in equilibrium. Express the fact that the *total* current (drift plus diffusion components) must be zero, by using (1.3.10'), (1.3.6), and (1.3.17). Applying (1.2.18), show that the Einstein relation (1.3.18) results.
- 1.12. Study the material on basic laws of electrostatics in Appendix A and provide detailed derivations for results (A.4), (A.6), and (A.7) given there.
- 1.13. Provide detailed derivations for the results in Fig. 1.25.
- 1.14. (a) Show that, within the weak inversion saturation region marked in Fig. 1.36a, I_D is of the form $I_1 \exp(V_{GS}/V_i)$, and deduce the values for I_1 and V_i .
(b) Show that within the strong inversion saturation region marked in Fig. 1.36b, I_D is of the form $k_1(V_{GS} - V_T)^2$, and deduce the values for k_1 and V_T .

The Two-Terminal MOS Structure

2.1 Introduction

In our gradual development toward the MOS transistor, we consider in this chapter the two-terminal MOS structure. This structure is often referred to as a *MOS capacitor* and is shown in Fig. 2.1. The acronym *MOS*, standing for metal-oxide-semiconductor, is used independently of whether the gate is actually made of metal or whether the insulator is silicon dioxide. The reasons are historical; in the early days, gates were invariably made of metal, the insulator was invariably silicon dioxide, and *semiconductor* unambiguously referred to the body. Such usage continues today, even though polysilicon became a widely used gate material and silicon dioxide is not the only material used for the insulator.

The two-terminal MOS structure has been the subject of extensive studies over many years. Such studies have resulted in a detailed understanding of the structure and have identified the sources of a number of undesirable effects that plagued early devices. This led to the development of better fabrication methods that greatly reduced such effects and that made possible MOS transistors with high performance. A detailed study of the two-terminal MOS structure and a related history can be found elsewhere.¹ Here we will only discuss those aspects directly relevant to the objective of this book. We will consider the various potentials and charges developed in the two-terminal MOS structure when a voltage is applied between gate and body. We will also consider the capacitance properties of the structure.

Most of the foundation for our later study of the complete MOS transistor is laid in this chapter. To lay this foundation adequately, a considerable amount of cross-referencing and page turning will be necessary in order to properly derive

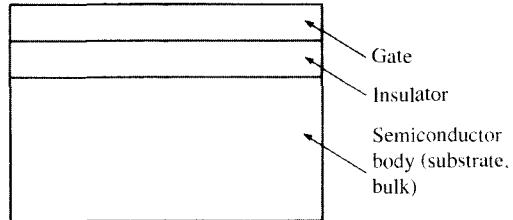


FIGURE 2.1
A two-terminal MOS structure.

several results and relate them to each other. We ask for the reader's patience in this respect.

2.2 The Flatband Voltage

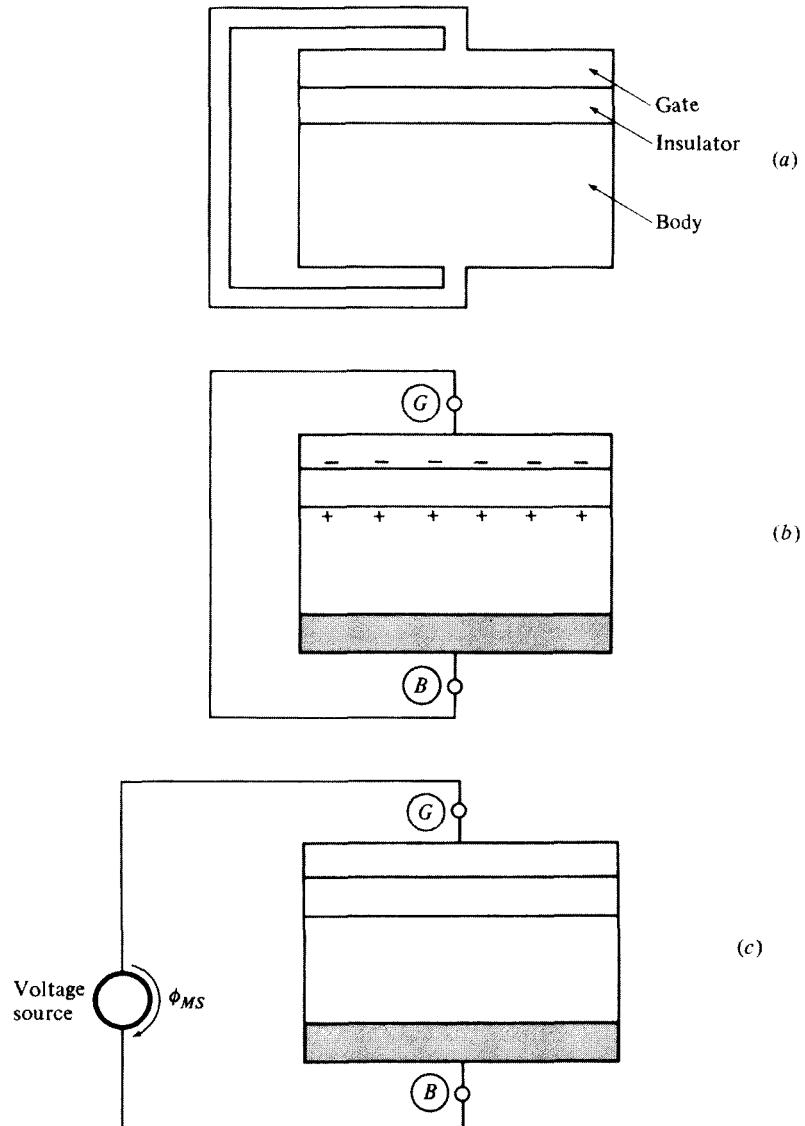
We begin our discussion with an "academic" case, shown in Fig. 2.2a. Here we assume that the gate is made of the same crystalline material as the body (in this case *p*-type silicon), with the same doping concentration. We also assume that somehow the same material is used to connect the gate to the body, in a sense making one an extension of the other. No charges are shown in the silicon, since to each positively charged hole of the *p*-type material there corresponds a negatively charged acceptor atom, from which the hole has originated, and there is no reason for the holes to pile up in any particular region. The charges, therefore, cancel and the material is shown as electrically neutral everywhere. The structure of Fig. 2.2a is basically symmetric; assuming no parasitic charges have been introduced during fabrication, no field can exist in the insulator or the semiconductor, and there is no reason for carriers to be attracted toward the insulator-body interface.

Consider now a realistic case, as shown in Fig. 2.2b. The gate is made out of a certain material, not necessarily the same as the substrate material. A metal is used to contact the gate material and form the *gate terminal G* as shown. In the following, the subscript *G* will be used to indicate the *gate terminal* rather than the gate itself. The body (substrate) is contacted through a back metal plate as shown. This plate, in turn, is contacted through some metal, which thus forms the *body terminal B*. Again, the subscript *B* will be used to denote the *body terminal*, as shown, rather than the body itself.

Let us now short-circuit the gate terminal to the body terminal by using a wire as shown. Consider going from the gate material through the external connection to the body. In doing so we encounter several contact potentials. However, from our discussion in Sec. 1.4, we know that their sum will only depend on the first and the last material, being independent of any material in between. Applying (1.4.5) we have

$$\begin{aligned} \text{Sum of all contact potentials} \\ \text{from gate, through external} \\ \text{connection, to bulk} &= \phi_S - \phi_M \end{aligned} \quad (2.2.1)$$

where ϕ_S and ϕ_M are the work function potentials of the body and gate materials, respectively (historically, "S" stood for the semiconductor substrate and "M" for the

**FIGURE 2.2**

(a) A two-terminal MOS structure with gate, body, and short-circuiting external connection all made out of the same semiconductor material; (b) a MOS two-terminal structure (with gate and body made of different materials, and a bottom metal contact) with zero effective interface charge and with the gate terminal short-circuited to the body terminal; (c) the structure of (b) with a voltage source placed in the loop, so that the surface charge becomes zero; (d) effect of effective interface charge Q_o ; (e) the structure in (d) with additional external bias so that the surface charge becomes zero.

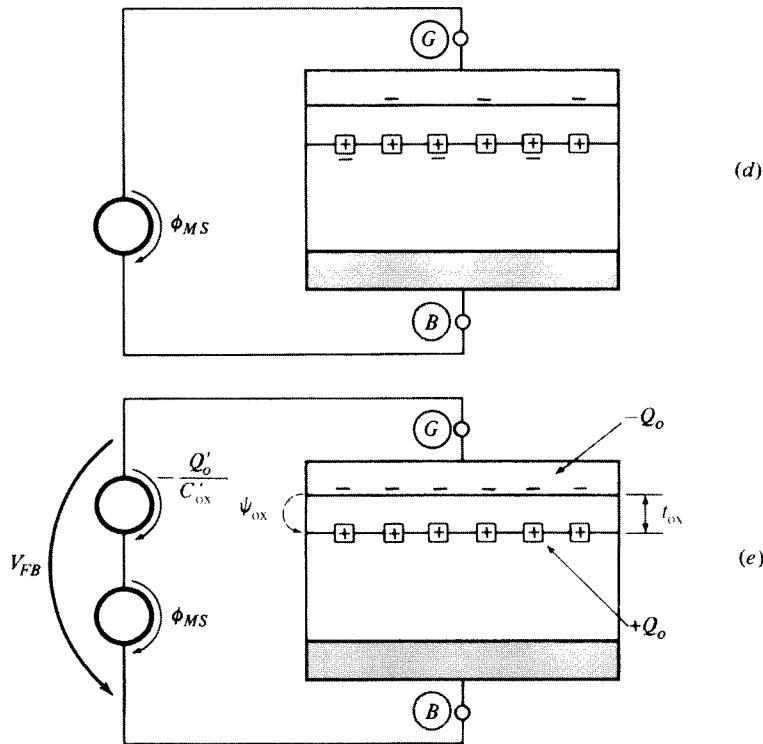


FIGURE 2.2 (continued).

metal gate). The existence of a nonzero potential between the gate material and the bulk causes net charges to appear on both sides of the oxide. If, e.g., this potential is negative, the polarity of the charges will be as shown in Fig. 2.2b.

Consider now the following question: Can an external voltage be applied in such a way that the net charges disappear? This case is shown in Fig. 2.2c. To make the total potential from the gate material through the external connection to the bulk equal to zero, the voltage of the external source must precisely cancel the sum of the contact potentials. From (2.2.1), we see that it must have a value given by

$$\phi_{MS} \equiv \phi_M - \phi_S \quad (2.2.2)$$

where ϕ_{MS} is a widely used symbol; the value of ϕ_{MS} can be calculated from this equation using the approach in Sec. 1.4.

Example 2.1. Find expressions for ϕ_{MS} at room temperature for a gate made of aluminum, n^+ polysilicon, and p^+ polysilicon, and a general silicon body for which its Fermi potential is given. For the work function potential of aluminum, use 4.1 V.

Using (2.2.2) and the approach in Sec. 1.4 involving (1.4.1) and (1.4.2), we obtain

For aluminum gate: $\phi_{MS} = -0.51 \text{ V} - \phi_F$

For n^+ polysilicon gate: $\phi_{MS} = -0.56 \text{ V} - \phi_F$

For p^+ polysilicon gate: $\phi_{MS} = +0.56 \text{ V} - \phi_F$

where in all cases ϕ_F is the Fermi potential of the substrate (positive for p type and negative for n type), given by (1.2.14). Note that the last two numbers are equivalent to the built-in potential between the silicon body and the silicon gate, and so they could have been obtained directly as discussed in Sec. 1.5.

The effect of contact potentials is not the only one that can cause a net concentration of charges in the body in the absence of external bias. Another cause is a "parasitic" charge that exists within the oxide as well as at the oxide-semiconductor interface. This charge consists of four parts:¹

1. An *oxide fixed charge* exists very close to the oxide-semiconductor interface due to the mechanisms of oxide formation at the time such formation is completed. This charge is found to be rather independent of oxide thickness, body doping type (n or p), and doping concentration.
2. A so-called *oxide trapped charge* can exist throughout the oxide, but usually close to either of its interfaces to the body or the gate. This charge can be acquired through radiation, photoemission, or the injection of high-energy carriers from the body.
3. A *mobile ionic charge* can exist within the oxide due to contamination by alkali ions (often sodium) introduced by the environment during fabrication. This charge can move within the oxide under the presence of an electric field.
4. An *interface trap charge* (also called *fast surface-state charge*) exists at the oxide-semiconductor interface. It is caused by defects at that interface, which give rise to charge "traps"; these can exchange mobile carriers with the semiconductor, acting as donors or acceptors.

The preceding charges initially inhibited the development of high-performance MOS devices because their nature was not well understood and they were difficult to control. In the early days, for example, the mobile ionic charge was large and drifted about the oxide, depending on the externally applied voltage and the temperature. This charge, often introduced into the oxide during fabrication through the hands of the people handling the device, resulted in very uncontrollable characteristics. Today, after many years of work, these four types of charge have been greatly reduced through appropriate fabrication techniques.¹ Throughout this book, we assume that the devices we are dealing with have been fabricated by using such techniques. Until further notice, we will assume that all parasitic charge is located at the oxide-semiconductor interface and that its value, denoted by Q_o , is fixed.[‡] Devices in which immobile charge exists *within* the oxide (uniformly distributed along the horizontal dimension in Fig. 2.2) can be modeled by assuming no such

[‡] The case of devices in which a significant fraction of Q_o is not fixed but depends instead on the externally applied potential (due to the interface traps) will be considered later.

charge exists and by adjusting Q_o accordingly (Prob. 2.16); thus Q_o can be taken to be the *effective interface charge*. This charge is almost always positive for both *p*- and *n*-type substrates. In modern devices, the effective interface ion density is about 10^{10} to 10^{11} ions/cm², corresponding to an effective interface charge density of 1.6×10^{-9} to 1.6×10^{-8} C/cm².

The effective interface charge Q_o is shown inside the little squares in Fig. 2.2d. A battery of value ϕ_{MS} is used in this figure to cancel the effect of the contact potentials discussed previously, so that we can study the effect of Q_o by itself. The charge Q_o will cause a total charge $-Q_o$ to appear in the system as demanded by charge neutrality. As shown in Fig. 2.2d, part of that charge will appear on the gate and part in the body. Assume we want to eliminate the latter. One can note that, if all the required balancing charge $-Q_o$ were provided on the gate, no charge would be induced in the semiconductor. To provide a charge $-Q_o$ on the gate, a battery can be connected in series with the external circuit, with the (-) terminal toward the gate. The following is now clear from basic electrostatics (Appendix A): Since at the gate and body ends of the oxide we must have charge $-Q_o$ and Q_o , respectively, the potential drop across the oxide, ψ_{ox} , defined from the gate through the oxide to the body, must be equal to $-Q_o/C_{ox}$, where C_{ox} is the total capacitance between the two ends of the oxide. This is exactly the voltage that must be provided by the battery as in Fig. 2.2e. In terms of quantities per unit area Q'_o and C'_{ox} , we have, for ψ_{ox} :

$$\psi_{ox} = -\frac{Q'_o}{C'_{ox}} \quad (2.2.3)$$

The oxide capacitance per unit area is given by (Appendix A)

$$C'_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$$

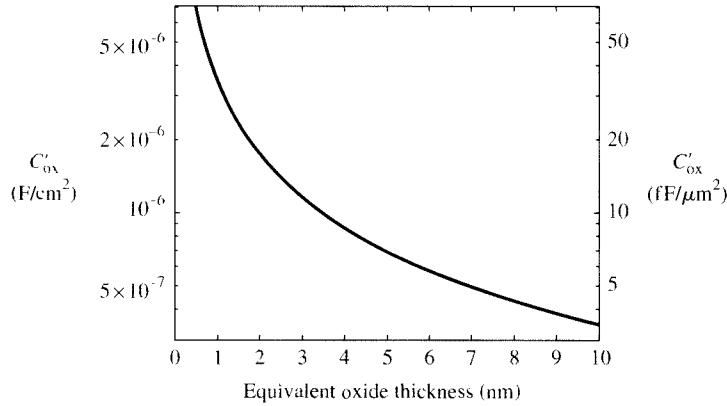
(2.2.4)

where t_{ox} is the thickness of the oxide and ϵ_{ox} is its permittivity, given by

$$\epsilon_{ox} = k_{ox}\epsilon_0 \quad (2.2.5)$$

with ϵ_0 the permittivity of free space (8.854×10^{-14} F/cm) and k_{ox} the dielectric constant of the insulator; for SiO₂, $k_{ox} = 3.9$. For simplicity, we are assuming that the gate insulator is made of only SiO₂. In modern technologies, other insulator materials are used as well. In such cases, (2.2.4) is still used, with t_{ox} understood to be the “equivalent oxide thickness” (EOT), defined as the thickness of a SiO₂ insulator that would be needed in that equation to give the correct value for the capacitance per unit area. If, for example, the insulator is made out of a single material with permittivity ϵ_{ins} and thickness t_{ins} , it is easy to see that the EOT to use in (2.2.4) would be $(\epsilon_{ox}/\epsilon_{ins})t_{ins}$. Equation (2.2.4) is plotted in Fig. 2.3.

We have therefore seen that an external voltage can be used between the gate and substrate terminals to keep the semiconductor everywhere neutral by canceling the effects of the contact potentials and Q'_o . This voltage is called the *flatband voltage* and is denoted by V_{FB} . The reason for this name is that with the gate-to-body

**FIGURE 2.3**

Capacitance per unit area vs. equivalent oxide thickness (EOT). For a SiO_2 ("oxide") insulator, EOT is the actual insulator thickness; for an insulator with permittivity ϵ_{ins} and thickness t_{ins} , EOT is $(\epsilon_{\text{ox}}/\epsilon_{\text{ins}}) t_{\text{ins}}$.

voltage V_{GB} equal to V_{FB} , the energy bands in the body, if plotted versus depth, are flat, as expected of a neutral body (see below). From Fig. 2.2e, we have the expression for the flatband voltage:

$$V_{FB} = \phi_{MS} - \frac{Q'_o}{C'_\text{ox}} \quad (2.2.6)$$

In modern fabrication processes, the second term in the preceding equation is often negligible.

Example 2.2. Calculate the flatband voltage for a *p*-type body with $N_A = 10^{18} \text{ cm}^{-3}$, a SiO_2 insulator with a thickness $t_{\text{ox}} = 2 \text{ nm}$, and an *n*-type polysilicon gate with $N_D = 10^{20} \text{ cm}^{-3}$. The interface charge Q'_o is $10^{-8} \text{ C}/\text{cm}^2$.

The Fermi potential of the body is, from (1.2.14), equal to 0.476 V. From Example 2.1, we have

$$\phi_{MS} = -0.56 \text{ V} - 0.477 = -1.036 \text{ V}$$

From (2.2.4), $C'_\text{ox} = 1.73 \times 10^{-6} \text{ F}/\text{cm}^2$, and therefore $-Q'_o/C'_\text{ox} = -0.006 \text{ V}$. Therefore, from (2.2.6),

$$V_{FB} = -1.036 \text{ V} - 0.006 \text{ V} = -1.042 \text{ V}$$

Although we have presented the previous concepts using potentials around the loop, an equivalent description is possible using energy bands, and is in fact more common. Consider a two-terminal MOS structure with a *p*-type body, as shown in Fig. 2.4a. Assume that the effective interface charge per unit area, Q_o , is zero. Let ϕ_M and ϕ_S be the work function potentials of the gate and body materials, respectively; see Fig. 1.17 for a definition of the work function potential. As an example,

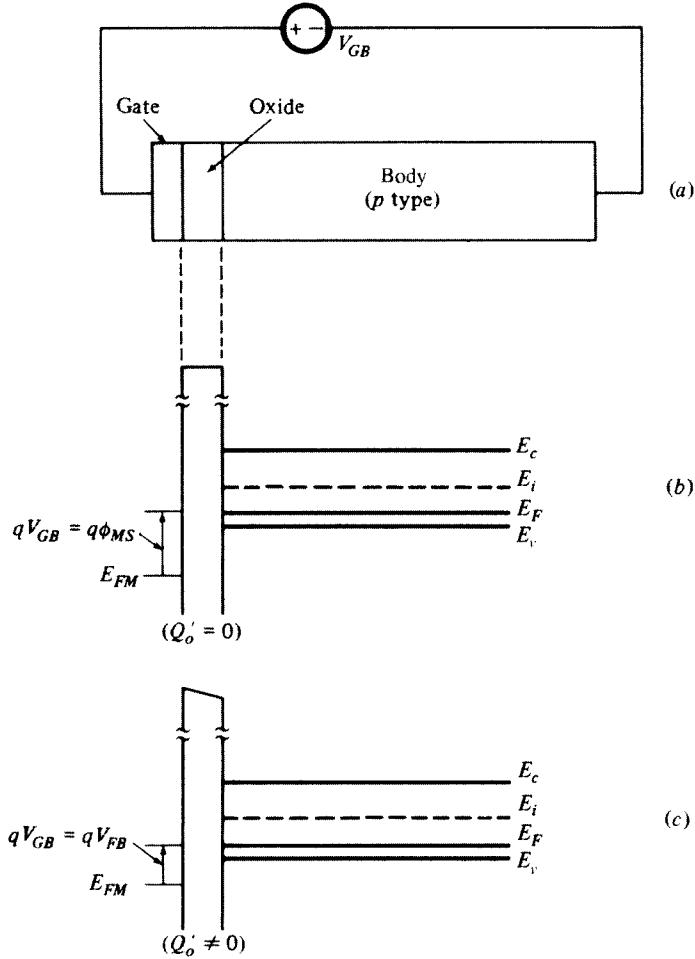


FIGURE 2.4

(a) A two-terminal MOS structure with a *p*-type body; (b) energy band diagram for the flatband condition, assuming $\phi_{MS} > 0$ and $Q'_o = 0$; (c) energy band diagram for the flatband condition, assuming $\phi_{MS} > 0$ and $Q'_o \neq 0$.

assume $\phi_M > \phi_S$. If the MOS structure is short-circuited ($V_{GB} = 0$), electrons in the body, where the work function is smaller, will find it attractive to leave and, traveling through the external wire, enter the gate. This will leave behind positive charges in the body, as has been shown in Fig. 2.2b, and will cause in it a potential drop. Because the system is in equilibrium, the other energy levels will bend until the Fermi levels in the gate and body can line up (as in the case of the *pn* junction in Fig. 1.22a). If it is desired to prevent this from happening and, instead, keep the bands “flat” in the body region next to the oxide, no potential drop and no net charges should exist there. This can be achieved by inserting in the external circuit a battery of value $V_{GB} = \phi_M - \phi_S$. When the connection is first established, the electrons in the substrate looking toward the external circuit no longer face the gate with

its electrons of lower energy directly, but rather face the *negative* terminal of the battery. The battery has increased the electron energy at that point by $q(\phi_M - \phi_S)$. Now the environment the electrons would face if they attempted to leave the body is no more attractive than their present environment (the body itself). Thus there is no reason for electrons to leave and for charges to pile up. The external voltage source keeps the Fermi levels of the gate (E_{FM}) and of the body (E_F) separated by $E_F - E_{FM} = q(\phi_M - \phi_S)$ and achieves the so-called *flatband* condition shown in Fig. 2.4b, which corresponds to Fig. 2.2c. A conduction band can be defined for the oxide (approximately 3.1 eV higher than the conduction band of silicon) and is shown to be horizontal in Fig. 2.4b as there is no field in the oxide under our assumption of $Q_o = 0$. If now $Q_o \neq 0$, then to keep the semiconductor bands flat, one has to adjust the external voltage to the value $V_{GB} = V_{FB}$ as given by (2.2.6). Now the conduction band in the oxide will not be horizontal because there will be a nonzero oxide field (see Fig. 2.2e) and a nonzero potential across it. The corresponding band diagram is shown in Fig. 2.4c.

2.3 Potential Balance and Charge Balance

We will now discuss how the body is affected when the externally applied voltage V_{GB} assumes values different from the flatband voltage V_{FB} . Consider as an example a MOS structure with a *p*-type body, shown in Fig. 2.5a. An arbitrary value of V_{GB} will in general cause charges to appear in the semiconductor. Practically all of these charges will be contained within a region adjacent to the top surface of the semiconductor, which is shaded in Fig. 2.5a. Outside this region, the substrate is practically neutral (we assume that the body is deeper than the extent of the charge region in it). We define the *surface potential* ψ_s as the total potential drop across the region, defined from the *surface* to a point in the *bulk* outside that region.[†] We will be more specific about the width of the region later on.

Four kinds of potential drops are encountered in the loop, as seen in Fig. 2.5:

1. The voltage of the external source V_{GB} .
2. The potential drop across the oxide ψ_{ox} .
3. The surface potential ψ_s .
4. Several contact potentials. Their sum, when going *clockwise*, is ϕ_{MS} , as seen from (2.2.1) and (2.2.2).

Going around the loop, we can write:

$$V_{GB} = \psi_{ox} + \psi_s + \phi_{MS} \quad (2.3.1)$$

[†] The top surface of the semiconductor is commonly referred to simply as “the surface” in MOS literature. The term *surface potential* is widely used.²⁻⁶ However, in some treatments,⁷ this term is used for a quantity that differs from our ψ_s by the Fermi potential ϕ_F ; such usage is uncommon.

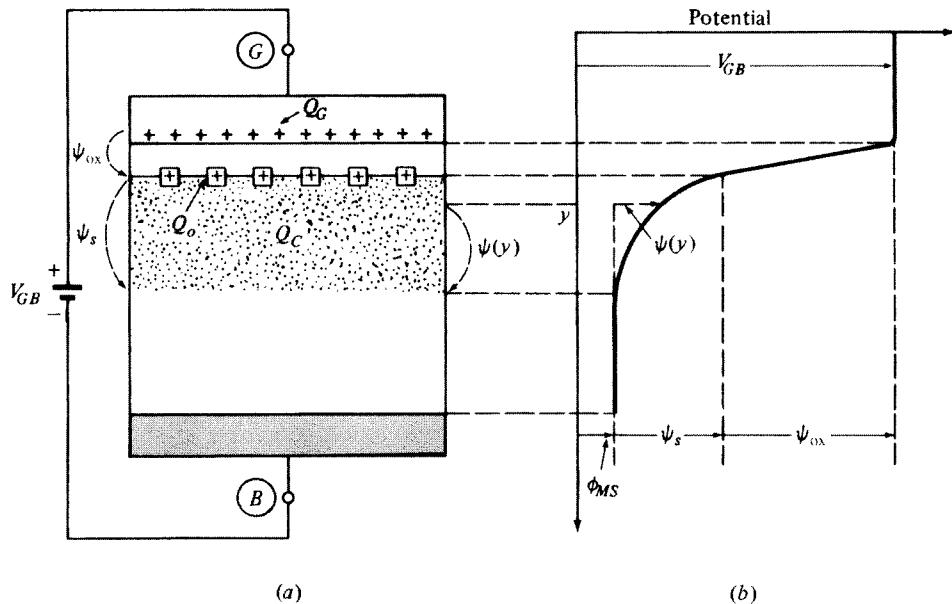


FIGURE 2.5
(a) A *p*-body, two-terminal MOS structure under general gate-body bias; (b) potential distribution, assuming the gate, the body-metal contact, and the external wires are all made of the same material. The special case of $\psi_s > 0$ has been assumed in drawing this plot.

The “potential balance” expressed by the preceding equation is illustrated in Fig. 2.5b. If the gate, the body, contact, and the wires were all made of the same material (e.g., aluminum), the only contact potential involved would be between the body and its metal contact. If more than one contact is involved, (2.3.1) and Fig. 2.5b still hold, but several contact potentials would be involved, adding up to the value given by (2.2.2).

Note that in (2.3.1) ϕ_{MS} is a known constant; therefore, any changes in V_{GB} must be balanced by changes in ψ_{ox} and ψ_s :

$$\Delta V_{GB} = \Delta \psi_{ox} + \Delta \psi_s \quad (2.3.2)$$

Consider now the charges in the system. In Fig. 2.5, we encounter three kinds of charges:

1. The charge on the gate Q_G .
2. The effective interface charge Q_o .
3. The charge in the semiconductor under the oxide Q_C .

These charges must balance one another for overall charge neutrality in the system:

$$Q_G + Q_o + Q_C = 0 \quad (2.3.3)$$

or, in terms of charges per unit area,

$$Q'_G + Q'_o + Q'_C = 0 \quad (2.3.4)$$

From now on, we will be using *charges per unit area* rather than charges most of the time. For brevity, though, we will often omit the words “per unit area”; these will be implied by the context.

Notice that if Q'_G is changed, the balance required by the foregoing equation will be achieved through a change in Q'_C , since the equivalent interface charge Q'_o is for the present assumed fixed:

$$\Delta Q'_G + \Delta Q'_C = 0 \quad (2.3.5)$$

The potential balance equation (2.3.1) and the charge balance equation (2.3.4) have been deduced from general fundamental physical laws. As we will see, particular properties of the MOS system will impose additional relations between the quantities appearing in the preceding equations.

2.4 Effect of Gate-Body Voltage on Surface Condition

Let us now consider the effect of V_{GB} on the condition of the region containing Q_C in Fig. 2.5. A *p*-type body is assumed. Depending on whether V_{GB} is equal to, less than, or greater than the flatband voltage V_{FB} , three cases are distinguished.

2.4.1 Flatband Condition

This case has already been discussed in detail in Sec. 2.2 and has been illustrated in Fig. 2.2e. From the associated discussion, we have

$$V_{GB} = V_{FB} \quad (2.4.1)$$

$$Q'_C = 0 \quad (2.4.2)$$

$$\psi_s = 0 \quad (2.4.3)$$

2.4.2 Accumulation

Consider the case in which V_{GB} decreases below V_{FB} (e.g., for the device of Example 2.2, this means that V_{GB} is *more negative* than -1.043 V). The negative change of V_{GB} (relative to flatband) will cause a negative change in Q'_G which, according to (2.3.5), must be balanced by a positive change in Q'_C above the value given by (2.4.2). Thus, holes will accumulate at the surface to provide a net positive charge. This condition is called *accumulation* and is illustrated in Fig. 2.6a. The negative change in V_{GB} causes negative changes in ψ_{ox} and ψ_s , and (2.3.1) will remain valid. In accumulation, therefore, we have

$$V_{GB} < V_{FB} \quad (2.4.4)$$

$$Q'_C > 0 \quad (2.4.5)$$

$$\psi_s < 0 \quad (2.4.6)$$

2.4.3 Depletion and Inversion

Assume now the case in which V_{GB} increases above V_{FB} . The total charge on the gate Q_G will become more positive than the value at flatband (that value is $-Q_o$ in Fig. 2.2e). An example is shown in Fig. 2.6b, where it is assumed that the resulting Q'_G is positive. The positive change in Q'_G (relative to flatband) must be balanced by a negative change in Q'_C so that (2.3.3) remains valid. Also, the positive change in V_{GB} will cause positive changes in ψ_{ox} and ψ_s , with (2.3.1) remaining valid. Accordingly, we have

$$V_{GB} > V_{FB} \quad (2.4.7)$$

$$Q'_C < 0 \quad (2.4.8)$$

$$\psi_s > 0 \quad (2.4.9)$$

Let us now consider the nature of the negative charge Q'_C .

Depletion If V_{GB} is not much higher than V_{FB} , the positive potential at the surface with respect to the body will simply drive holes away from the surface, leaving it depleted. This condition is called *depletion* and it is illustrated in Fig. 2.6b. More precisely, as V_{GB} is raised above V_{FB} , the hole density will keep decreasing well below the doping concentration value N_A . For practical purposes, then, the charge Q_C is due to the uncovered acceptor atoms, each of which contributes a charge $-q$, and we can assume the presence of a “depletion region” as in the *p* side of the *pn* junction in Sec. 1.5.

Inversion As V_{GB} is increased further, more acceptor atoms are uncovered, and ψ_s becomes sufficiently positive to attract a significant number of free electrons to the surface; each of these electrons will also contribute a charge $-q$ to Q_C . Note that in the two-terminal MOS structure we are considering, these electrons come from the relatively slow process of electron-hole generation in the depletion region, caused by the thermal vibration of the lattice (assuming no radiation is present). Eventually, with a sufficiently high V_{GB} the density of electrons will exceed that of holes at the surface. This is a situation opposite from that normally expected in a *p*-type material; we now have surface *inversion*. The situation is illustrated in Fig. 2.6c.[†]

[†] We should warn the reader that the term *region* is often used with two different meanings in MOS work. Thus, for example, “depletion region” could refer to the physical region containing the ionized acceptor atoms in Fig. 2.6b; or it could refer to the region of V_{GB} (or ψ_s) values in which we have depletion as has been defined. Which of the two meanings is employed will usually be clear from the context.

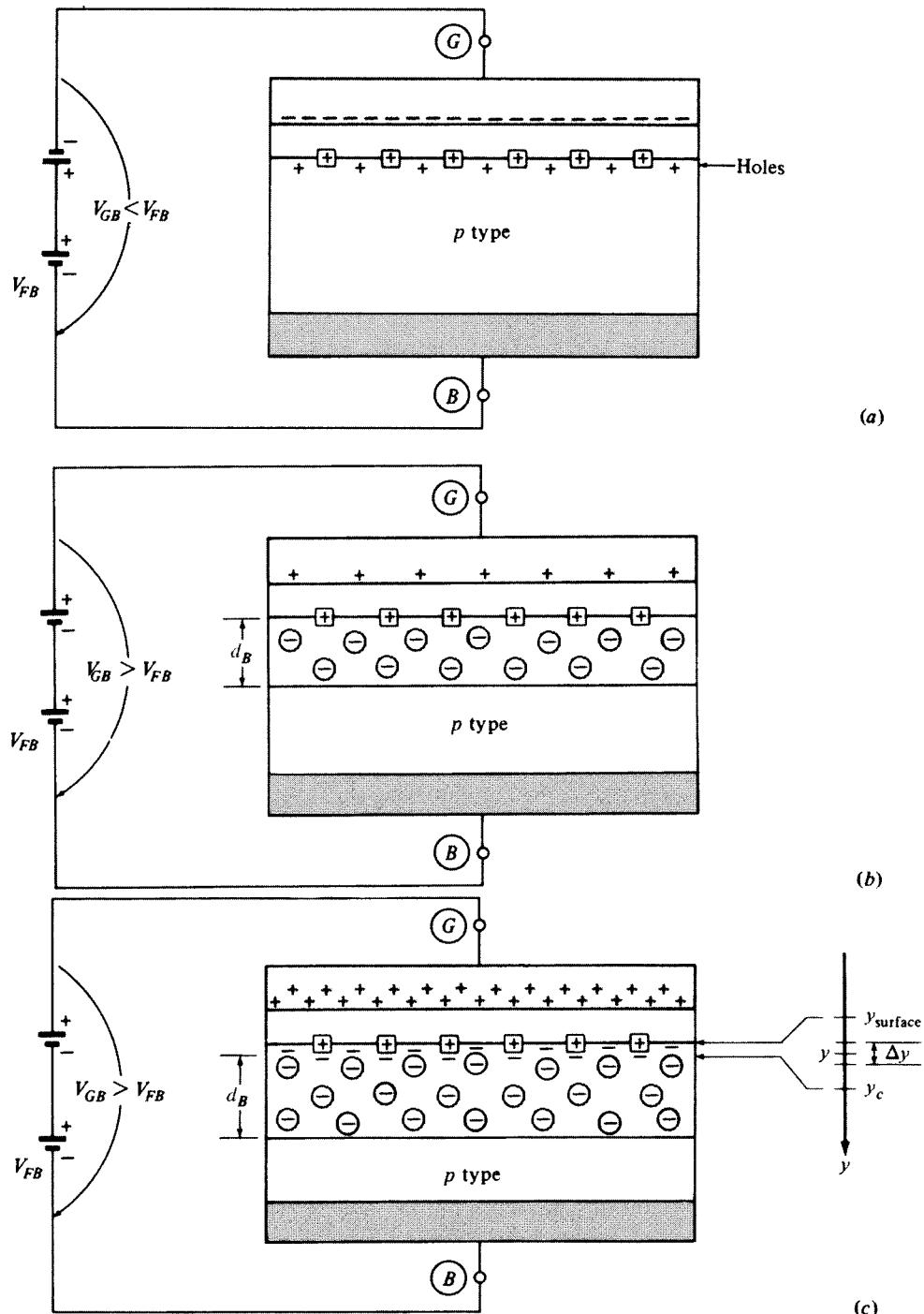


FIGURE 2.6
A MOS two-terminal structure in (a) accumulation; (b) depletion; and (c) inversion.

The charge, electric field, and potential for the cases of Fig. 2.6 can be determined using general relations for the charge densities and the laws of electrostatics.

Let us see what the energy bands look like for the various cases we considered previously. For simplicity in the rest of the discussion, we will assume both $\phi_{MS} = 0$ and $Q_o = 0$. Then the flatband condition corresponds to the band diagram shown in Fig. 2.7a. Now $V_{FB} = 0$, so accumulation will be caused by $V_{GB} < 0$. This increases the electron energy in the gate, and the Fermi energy there becomes higher as shown in Fig. 2.7b. The bands bend accordingly as shown. Recall that the potential varies in the opposite direction from the energy bands, as in (1.2.17), and thus the surface potential is negative. At the surface, the valence band edge approaches the Fermi energy, and thus the condition there is very favorable for holes (see Sec. 1.2.4). This case corresponds to Fig. 2.6a. If, instead, we apply a small $V_{GB} > 0$, we have the case shown in Fig. 2.7c. Now the bands bend downward as we go from the bulk toward the surface, which means that the potential of the surface with respect to the bulk becomes positive. The situation at the surface is no longer favorable for holes, and we get a depletion region as in Fig. 2.6b.[†] If V_{GB} is made even more positive, we get further band bending as shown in Fig. 2.7d, and we are in inversion. Indeed, now E_F is higher than E_i at the surface, a situation normally encountered in *n*-type semiconductors, despite the fact that the substrate is clearly *p*-type. The surface potential is now large enough to attract a significant number of electrons at the surface, as we have seen in Fig. 2.6c.

Carrier Concentration The oxide is assumed to be thick enough to block the flow of current, and thus equilibrium is maintained in the substrate. Deep in the neutral bulk outside the depletion region, the values of the hole and electron concentrations will be p_o and n_o , as given by (1.2.6) and (1.2.7). We can relate the electron concentration at the surface to that in the bulk by using (1.2.18):

$$n_{\text{surface}} = n_o e^{\psi/\phi_i} \quad (2.4.10)$$

Using (1.2.11), the preceding can be written as follows:

$$n_{\text{surface}} = n_i e^{(\psi_i - \phi_F)/\phi_i} \quad (2.4.11a)$$

$$= p_o e^{(\psi_i - 2\phi_F)/\phi_i} \quad (2.4.11b)$$

and, since $p_o \approx N_A$ from (1.2.6), we have

$n_{\text{surface}} \approx N_A e^{(\psi_i - 2\phi_F)/\phi_i}$

(2.4.12)

[†] We remind the reader that the depletion region is just an approximation; some holes and electrons are still present in it, and if the band bending is known their density can be calculated from (1.2.8). The band bending itself can be determined if the potential variation is known from (1.2.17). This can be done using Poisson's equation, as we will discuss in Sec. 2.4.4.

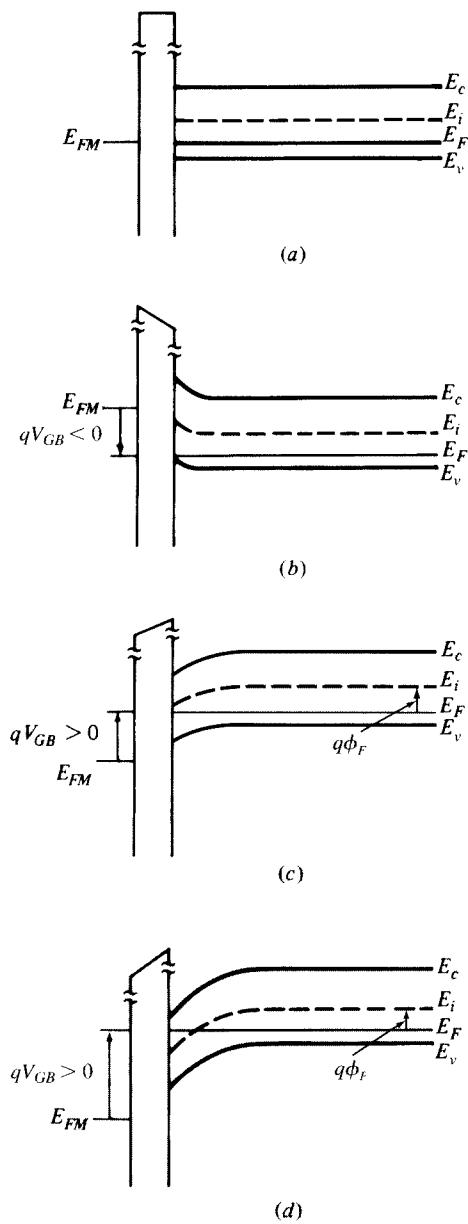


FIGURE 2.7
Energy band diagrams for a two-terminal MOS structure with a *p*-type body, assuming $\phi_{MS} = 0$ and $Q'_o = 0$, for various values of V_{GB} . (a) Flatband condition; (b) accumulation; (c) depletion; (d) inversion.³

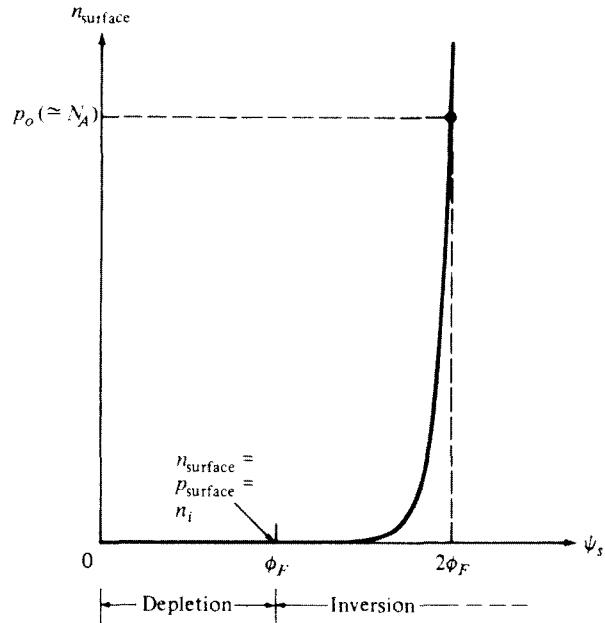


FIGURE 2.8
Electron concentration at the surface vs. surface potential. Linear axes are used.

The surface electron concentration is plotted vs. the surface potential in Fig. 2.8. Some interesting points are shown in the figure. At $\psi_s = \phi_F$, n_{surface} becomes equal to the intrinsic concentration as seen from (2.4.11a); from (1.2.16) then, $n_{\text{surface}} = p_{\text{surface}}$. The total band bending in this case is $q\phi_F$, i.e., E_i bends just enough to touch E_F . This is defined as the boundary between the depletion and inversion regions, as indicated in Fig. 2.8. As seen from (2.4.11a), n_{surface} will be nonzero even in depletion but will be much smaller than n_i , even for ψ_s smaller than ϕ_F by only a few ϕ_F . With increasing ψ_s above ϕ_F , n_{surface} increases drastically, and at $\psi_s = 2\phi_F$ we have $n_{\text{surface}} = p_o \approx N_A$, as seen from (2.4.11b) and (2.4.12).

If the body is made of *n*-type material, the above description should be modified in a rather obvious manner. The inversion layer in this case will consist of holes, which will be attracted to the surface if V_{GB} is sufficiently negative. The immobile charge in the depletion region will consist of positively charged ionized donor atoms. With V_{GB} sufficiently positive, electrons will pile up at the surface and we will have accumulation. An example of an energy band diagram for an *n*-type substrate in inversion is shown in Fig. 2.9. Since the case of *n*-type substrates is complementary to that of *p*-type substrates, it will not be discussed separately.

2.4.4 General Analysis

Of the three operation regions defined previously, inversion is by far the most important for the purposes of this book. This region is treated in detail in the following section. It should be noted at this point, though, that a general analysis is possible through which one can determine ψ_s and Q'_C for any value of V_{GB} , be it in accumulation,

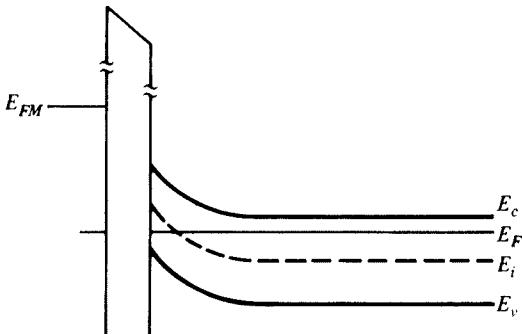


FIGURE 2.9
Energy band diagram for a two-terminal MOS structure with an *n*-type body in inversion.

depletion, or inversion. Although such general analysis will not be of much use in this book, we will summarize the principles behind it for completeness. The mathematical details will be left for Appendix C. Consider a point of ordinate y in the body (Fig. 2.5), and let $\psi(y)$ be the potential there referenced to a point deep in the bulk. From (1.2.18) we will have, for the electron concentration at y ,

$$n(y) = n_o e^{\psi(y)/\phi_t} \quad (2.4.13)$$

where n_o is the electron concentration in the bulk. This equation reduces to (2.4.10) at the surface. Similarly, for the hole concentration at y , we will have, according to (1.2.19)

$$p(y) = p_o e^{-\psi(y)/\phi_t} \quad (2.4.14)$$

where p_o is the hole concentration deep in the bulk.

In the general case, one must allow for the presence of both electrons and holes below the insulator and for the presence of both acceptor and donor atoms. Thus, we use (1.2.21):[†]

$$\rho(y) = q[p(y) - N_A + N_D - n(y)] \quad (2.4.15)$$

Substituting in this (2.4.13) and (2.4.14), we can write Poisson's equation (1.2.29) as follows:

$$\frac{d^2\psi}{dy^2} = -\frac{q}{\epsilon_s} [p_o e^{-\psi(y)/\phi_t} - N_A + N_D - n_o e^{\psi(y)/\phi_t}] \quad (2.4.16)$$

With the help of this equation and the three before it, and for a given ψ_s , one can determine $\psi(y)$, $n(y)$, $p(y)$, and $\rho(y)$ and the corresponding total charge per unit

[†] If not all ions are ionized (e.g., at very low temperatures), N_A and N_D should be replaced by the densities of ionized acceptor and donor atoms, respectively; the determination of these densities is discussed in the literature.¹⁻⁴

area in the silicon, Q'_C (Appendix C). Unfortunately, only Q'_C can be derived in closed form; the other quantities require a numerical evaluation. Solving (2.4.16) as explained in Appendix C, we obtain

$$Q'_C = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} \sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (N_A - N_D) \psi_s + n_o \phi_t (e^{\psi_s/\phi_t} - 1)}$$

Hole contribution	Dopant ion contribution	Electron contribution
----------------------	----------------------------	--------------------------

(2.4.17)

where $\text{sgn}(\psi_s)$ is the signum function, defined as +1 for $\psi_s > 0$, 0 for $\psi_s = 0$, and -1 for $\psi_s < 0$.

The contributions noted come from tracing the contributions of the terms in (2.4.16) in the course of integrating this equation (Appendix C). The quantity $\sqrt{2q\epsilon_s}$ is the same quantity encountered in our discussion of the *pn* junction in Sec. 1.5 and has the value of $5.8 \times 10^{-16} \text{ C} \cdot \text{V}^{-1/2} \cdot \text{cm}^{-1/2}$. This equation exhibits symmetry and is very general; it is valid for both *p*-type and *n*-type substrates of any doping type and density. This equation leads to other relations, e.g., for the surface potential and for capacitances, that are likewise general and with easily recognizable terms; such relations are valuable and are given in Appendix C. However, in most of this text we will be using alternative forms of such equations, which involve the Fermi potential of the substrate; these lead to widely used models. To develop such equations, note that when $\psi = 0$ (flatband), the body is neutral everywhere; this means, from (2.4.13) to (2.4.15), that

$$p_o - N_A + N_D - n_o = 0 \quad (2.4.18)$$

Using this equation, $N_A - N_D$ in (2.4.17) can be eliminated, giving

$$Q'_C = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} \sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (p_o - n_o) \psi_s + n_o \phi_t (e^{\psi_s/\phi_t} - 1)} \quad (2.4.19)$$

We can express n_o in terms of p_o using (1.2.11) as follows:

$$n_o = p_o e^{-2\phi_F/\phi_t} \quad (2.4.20a)$$

and from now on we will focus on the case of *p* substrates with an effective acceptor concentration N_A , assuming that all acceptor atoms are ionized; thus, from (1.2.6)

$$p_o \approx N_A \quad (2.4.20b)$$

For simplicity, in what follows we will replace \approx with an equal sign in this equation. Using (2.4.20) in (2.4.19), we obtain

$$Q'_C = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s N_A} \sqrt{\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-2\phi_F/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t)} \quad (2.4.21)$$

A plot of this equation is shown in Fig. 2.10. In accumulation ($\psi_s < 0$), the charge is dominated by holes and is positive. In depletion ($0 < \psi_s < \phi_F$), the charge is dominated by acceptor atoms and is negative. In inversion ($\psi_s > \phi_F$), in the beginning the electron charge is small; Q'_C is still dominated by acceptor atoms (flat part). As ψ_s increases further, the electron density becomes dominant and the charge becomes strongly negative, as shown. The broken straight line will be discussed shortly.

To complete our set of basic equations, note that the charge per unit area above the oxide, Q'_G , can be simply related to the potential across the oxide, ψ_{ox} , and the oxide capacitance per unit area, C'_{ox} , by (Prob. 2.11):

$$Q'_G = C'_{ox} \psi_{ox} \quad (2.4.22)$$

At this point, we have four equations that completely characterize the MOS structure under our assumptions. These equations are:

- *Potential balance* (2.3.1): $V_{GB} = \psi_{ox} + \psi_s + \phi_{MS}$
- *Charge balance* (2.3.4): $Q'_G + Q'_o + Q'_C = 0$
- *The semiconductor charge-potential relation* (2.4.21), which is of the form $Q'_C = Q'_C(\psi_s)$
- *The gate charge-potential relation* (2.4.22), $Q'_G = C'_{ox} \psi_{ox}$

In these equations, the quantities ϕ_{MS} , Q'_o , C'_{ox} , and N_A are assumed known for a given fabrication process. For a given V_{GB} , we can, in principle, completely solve these four equations to determine ψ_{ox} , ψ_s , Q'_G , and Q'_C . Unfortunately, this is not possible to do analytically, and thus numerical techniques are used. It is instructive, though, to look at a graphical solution, as is now discussed.

Let us assume that V_{GB} is given, and we need to determine Q'_C and ψ_s . Note that these quantities must satisfy (2.4.21) because of semiconductor properties in the body; however, the values of these two quantities must also be consistent with the other relations that characterize the structure, summarized in the preceding bulleted list. To determine what constraints these impose, we solve the charge balance equation for Q'_C , and in the result we use the gate charge-potential relation, the potential balance relation, and the definition of V_{FB} from (2.2.6); this gives

$$Q'_C = -C'_{ox}(V_{GB} - V_{FB} - \psi_s) \quad (2.4.23)$$

Thus, in principle Q'_C and ψ_s are the solution of the system consisting of (2.4.21) and (2.4.23). The plot of (2.4.23) is indicated in Fig. 2.10 using a broken line for the given V_{GB} . The slope of this line is C'_{ox} ; the abscissa intercept is $V_{GB} - V_{FB}$. Because both (2.4.21) and (2.4.23) must be satisfied, the solution is at the intersection of the two lines, shown by a dot in the figure; this gives the values of ψ_s and Q'_C for the given V_{GB} . This graphical construction is useful for obtaining a qualitative picture, as we will see.

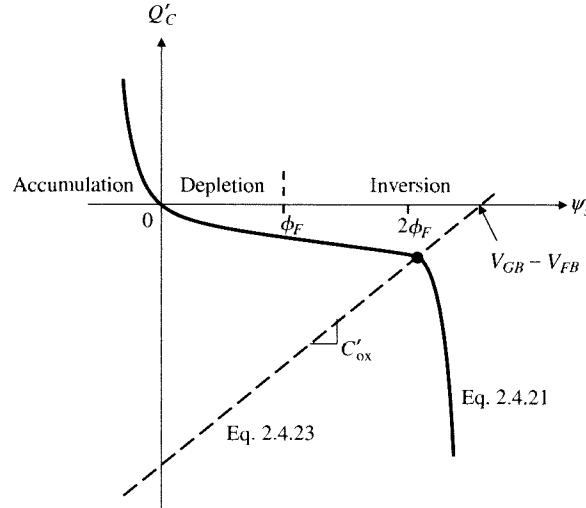


FIGURE 2.10
Total semiconductor charge vs. surface potential (ψ_s) (solid line) and constraint imposed by the rest of the system (broken line).

Let us now consider in more detail the solution for the surface potential ψ_s (the effect) for a given value of the externally applied voltage V_{GB} (the cause). Solving (2.4.23) for V_{GB} we get

$$V_{GB} = V_{FB} + \psi_s - \frac{Q'_C(\psi_s)}{C'_{\text{ox}}} \quad (2.4.24)$$

Using (2.4.21) for $Q'_C(\psi_s)$ we get

$$V_{GB} = V_{FB} + \psi_s + \text{sgn}(\psi_s) \gamma \sqrt{\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-2\phi_F/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t)} \quad (2.4.25)$$

where

$$\boxed{\gamma = \frac{\sqrt{2q\epsilon_s N_A}}{C'_{\text{ox}}}} \quad (2.4.26a)$$

with $\sqrt{2q\epsilon_s} = 5.8 \times 10^{-16} \text{ F} \cdot \text{cm}^{-1/2} \cdot \text{V}^{1/2}$ (the units chosen here are equivalent to those given earlier and are convenient for use in this equation). This quantity is called the *body effect coefficient*, for reasons that will become apparent in Chap. 3. The above formula can be transformed into the following form:

$$\gamma = (0.53 \text{ V}^{1/2}) \left(\frac{t_{\text{ox}}}{10 \text{ nm}} \right) \sqrt{\frac{N_A}{10^{17} \text{ cm}^{-3}}} \quad (2.4.26b)$$

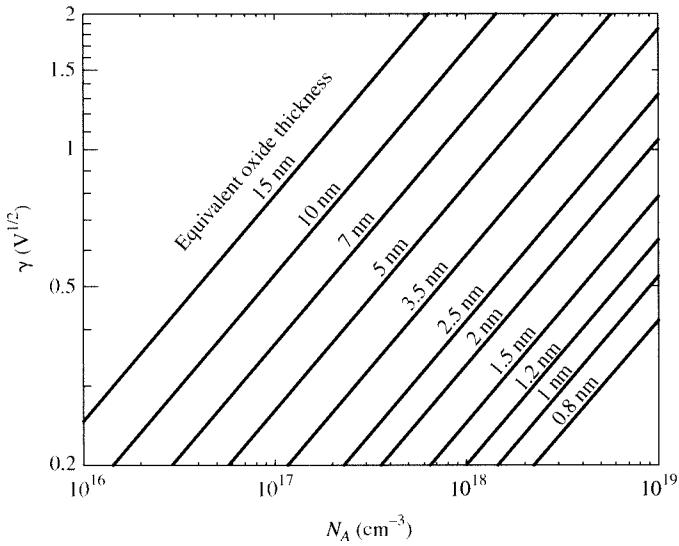


FIGURE 2.11
Body effect coefficient vs. substrate doping, for several values of equivalent oxide thickness.

which is convenient and easy to remember. The values of γ are shown in Fig. 2.11 for a variety of process parameters.

Equations (2.4.21), (2.4.23), and (2.4.25) are not independent; each of them can be obtained from the other two. We show all three, because each has its uses depending on the situation, as we will see.

Equation (2.4.25) has no analytical solution in terms of ψ_s . Numerical techniques must be used to find ψ_s for a given V_{GB} (these will be discussed in Chap. 4). Nevertheless, we can plot ψ_s vs. V_{GB} without using such techniques by simply giving values to ψ_s and finding the corresponding V_{GB} from this equation. Doing this gives the solid-line plot of Fig. 2.12. The broken-line plot will be discussed in Sec. 2.5.

To conclude the analysis, we need the behavior of the charge as a function of the applied voltage. This can be obtained by viewing (2.4.21) and (2.4.25) as a parametric representation of Q'_C vs. V_{GB} . We can give values to ψ_s , calculate Q'_C and V_{GB} from these two equations, and then plot Q'_C vs. V_{GB} . This results in the plot of Fig. 2.13. We see that although in the depletion region V_{GB} does not have a strong effect on Q'_C , in accumulation and deep in inversion it does. In other words, once high concentrations of holes or electrons are induced at the surface, their density is effectively controlled by V_{GB} .

Although it takes a lot of number crunching to produce the last two plots, one could have *expected* the general behavior exhibited by them without any calculations. Thus, return to the graphical construction in Fig. 2.10 and follow the coordinates of the point of intersection between the two lines, shown by a thick dot. It can be seen that if $V_{GB} - V_{FB}$ is zero and is increased, initially ψ_s changes significantly, but Q'_C does not. Eventually, when $V_{GB} - V_{FB}$ is about as large as assumed in the figure, further increases in V_{GB} have a significant effect on Q'_C , whereas ψ_s is not affected much.

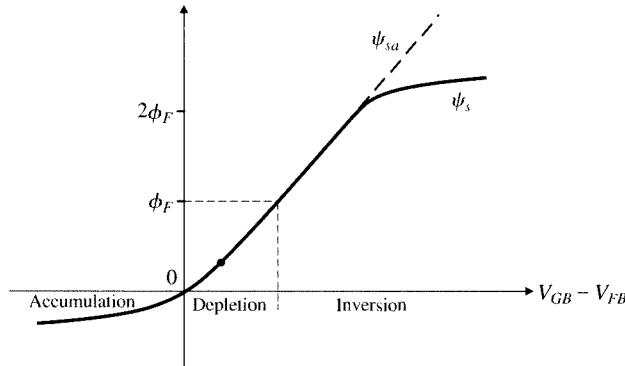


FIGURE 2.12
Surface potential vs. $V_{GB} - V_{FB}$.

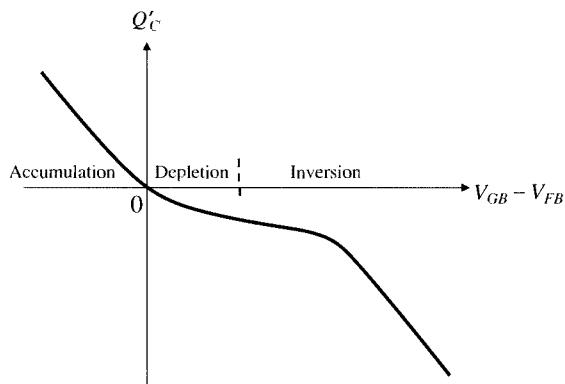


FIGURE 2.13
Total semiconductor charge vs. $V_{GB} - V_{FB}$.

This agrees with what is observed in Figs. 2.12 and 2.13; the reasons for this behavior will be considered in detail in Sec. 2.5. The qualitative behavior in accumulation can be similarly predicted using the preceding graphical technique.

As already mentioned, the quantitative analysis of the MOS structure is complicated. The complexity is, in part, due to the fact that complete generality was sought (i.e., validity in accumulation, depletion, and inversion). By focusing on particular regions, certain approximations become possible that simplify the analysis, as will be seen in the following sections. We now seek such simplifications.

2.5 Accumulation and Depletion

In accumulation and depletion (Figs. 2.6a and b), the contribution of electrons can be neglected. Dropping the corresponding term from (2.4.17) and using (2.4.20) and $N_D = 0$ gives

$$Q'_C = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s N_A} \sqrt{\phi_t e^{-\psi_s/\phi_t} - \phi_t + \psi_s} \quad (2.5.1)$$

This equation can be used as long as $\psi_s < \phi_F$ and it gives results essentially identical to the left part of the plot in Fig. 2.10. All terms in (2.5.1) are important in maintaining a smooth transition from accumulation to depletion, as ψ_s changes from negative to positive.

Deep in Accumulation Deep in accumulation (for ψ_s below zero by several ϕ_t), the exponential term in (2.5.1) dominates and the other terms can be neglected; thus, we have

$$Q'_C \approx \sqrt{2q\epsilon_s N_A \phi_t} e^{-\psi_s/(2\phi_t)} \quad (2.5.2)$$

This exponential behavior for negative ψ_s is apparent in Fig. 2.10.

Deep in Depletion Deep in depletion (for ψ_s above zero by several ϕ_t , but less than ϕ_F as indicated after (2.5.1)), the exponential in (2.5.1) becomes negligible; if we also neglect ϕ_t ,[†] we have

$$Q'_C \approx -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_s} \quad (2.5.3)$$

If this equation is substituted in (2.4.24), we obtain

$$V_{GB} = V_{FB} + \psi_s + \gamma \sqrt{\psi_s} \quad (2.5.4)$$

where γ is given by (2.4.26). The preceding equation could also have been obtained from (2.4.25) by dropping negligible terms in it.

It is interesting that, in contrast to the general case, (2.5.4) can be solved explicitly for ψ_s . Denoting the solution by ψ_{sa} , we have

$$\psi_{sa} = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 \quad (2.5.5)$$

This particular result will be helpful in a subsequent development. To summarize, ψ_{sa} is the surface potential obtained deep in depletion (where both hole and electron concentrations are negligible) for a given externally applied voltage, V_{GB} . The quantity ψ_{sa} is plotted by a broken line in Fig. 2.12, where it is compared with the general result for the surface potential ψ_s (2.4.25). It can be seen that it predicts the surface potential accurately in deep depletion (it is only plotted above the point marked by a dot) and even in weak inversion (surface potential between ϕ_F and $2\phi_F$). However, in moderate and strong inversion ψ_s can be very different from ψ_{sa} because in those regions there exists a considerable inversion layer charge that has not been taken into account in developing (2.5.5).

[†] In other treatments, ϕ_t is retained, and the quantity $\sqrt{\psi_s}$ in the equations that follow is replaced by $\sqrt{\psi_s - \phi_t}$. The origin of the ϕ_t term under the square root can be traced to the contribution of the holes in the substrate, which are not expelled from the depletion region if ψ_s is not too large [see (2.4.17)]. This term is included in advanced surface potential-based models for CAD, discussed in Chap. 10.

2.6 Inversion

2.6.1 General Relations and Regions of Inversion

In inversion (Fig. 2.6c), the hole contribution is negligible. We only need to consider the contributions from the acceptor atoms and the electrons (we assume always $N_D = 0$). Also, in the last term in (2.4.17), the quantity 1 is negligible in comparison to the exponential (recall that, as is marked in Fig. 2.8, in inversion the surface potential is larger than ϕ_F , which is typically several tenths of a volt, whereas ϕ_i is about 26 mV at room temperature). Thus, we have[†]

$$Q'_C = -\sqrt{2q\epsilon_s} \sqrt{N_A\psi_s + n_0\phi_i e^{\psi_s/\phi_i}} \quad (2.6.1a)$$

Using in this (1.2.18) and (2.4.20) gives an alternative form:

$$Q'_C = -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_s + \phi_i e^{(\psi_s - 2\phi_i)/\phi_i}} \quad (2.6.1b)$$

The total charge (per unit area) below the oxide is the sum of the charge due to the electrons in the inversion layer Q'_I and the charge due to the ionized acceptor atoms in the depletion region Q'_B :

$$Q'_C = Q'_I + Q'_B \quad (2.6.2)$$

We now undertake the evaluation of Q'_I and Q'_B . Consider first the electrons in the inversion layer. At any point of ordinate y in Fig. 2.6c, the electron concentration $n(y)$ will be given by (2.4.13). As one goes away from the surface, $\psi(y)$ decreases from ψ_s toward zero, and $n(y)$ decreases rapidly owing to its exponential dependence on $\psi(y)$. Hence, one can choose a point $y = y_c$ below which the electron concentration will be negligible. Practically all of the free electrons are then contained in a layer between $y = y_{\text{surface}}$ and $y = y_c$ (Fig. 2.6c). The number of electrons contained in a thin layer of thickness Δy , parallel to the surface and centered around y , will be $n(y)(A \Delta y)$, where A is the cross-sectional area as seen from the top. The total charge due to these electrons will be $(-q)n(A \Delta y)$. We can then express the charge due to all electrons in the inversion layer, denoted by Q_I , as follows:

$$Q_I = \int_{y_{\text{surface}}}^{y_c} (-q)n(y)(A dy) \quad (2.6.3)$$

Therefore, the inversion layer charge per unit area Q'_I will be given by

$$Q'_I = -q \int_{y_{\text{surface}}}^{y_c} n(y) dy \quad (2.6.4)$$

Evaluating Q'_I in this manner is a lengthy process and can be done through numerical integration^{1,3} (Appendix C). We will instead follow a widely used

[†] Equation (2.6.1a) is also valid in the upper part of the depletion region (i.e., for ψ_s larger than a few ϕ_i), because the exponential there is negligible and the equation simplifies to (2.5.3).

simplifying approach: We will determine a sufficiently accurate expression for Q'_B and then return to evaluate Q'_I from (2.6.2) and (2.6.1).

As in the case of the *pn* junction, we will adopt the *depletion approximation*: We will assume that the depletion region contains only acceptor atoms and is defined by a sharp boundary at a depth d_B below the surface, with the semiconductor being neutral below it. This is shown in Fig. 2.14a. This approximation gives a constant volume charge density in the bulk, as shown in Fig. 2.14b. Let A be the area as seen from above and d_B the depth of the depletion region (we will discuss how to determine this depth shortly). The product of the volume charge density times the volume $d_B A$ of the depletion region is the total charge Q_B ; thus the product of the volume charge density and d_B is Q_B/A , i.e., it is the area charge density, Q'_B . This is indicated in Fig. 2.14b. We are also assuming that the charges on the gate and in the oxide are contained in very thin layers, as shown in Fig. 2.14b. The areas under the corresponding portions of the volume charge density plot are Q'_G and Q'_O , respectively, as shown. Finally, in drawing the figure, we have assumed that the inversion layer is much thinner than the depletion region, and that the volume charge density is constant in the inversion layer.[†] The inversion layer charge per unit area is denoted by Q'_I in Fig. 2.14b.

From the volume charge density plot of Fig. 2.14b, the electric field plot can be obtained from (1.2.26), using as y_0 a point on the gate above the charge region there, at which the field is zero. As y increases in the “running integral” in (1.2.26), it picks up more and more of the charges and produces the field plot shown in Fig. 2.14c. At boundaries between regions of different permittivity, we observe jumps in the electric field, by factors equal to the ratio of the two permittivities involved (Appendix A), as shown. From the field plot, we can obtain the potential plot using (1.2.28), as shown in Fig. 2.14d. These plots are considered in more detail in Prob. 2.15.

Charge Sheet Approximation A convenient approximation that leads to manageable models is that the inversion layer is a sheet of negligible thickness.^{8–13} This has been called the *charge sheet* approximation.¹³ We will thus let the inversion layer thickness approach zero in Fig. 2.14.[‡] The potential drop across the inversion layer, which is the area of the corresponding part in the field plot, approaches zero, too (Prob. 2.13), and we can assume that all of the surface potential ψ_s is dropped across the depletion region. The depth of the depletion region now becomes all of d_B in the figure.

The field at the top of the depletion region, \mathcal{E}_B in Fig. 2.14c, can be easily found from the laws of electrostatics in Sec. 1.2.6 and Appendix A. Thus, this field is the area under the volume charge density, $qN_A d_B$, divided by the permittivity ϵ_s :

$$\mathcal{E}_B = \frac{qN_A d_B}{\epsilon_s} \quad (2.6.5)$$

[†] This assumption is just a drawing convenience and is not valid in practice. However, it can be seen from (1.2.26) that, as the thickness of the inversion layer is allowed to approach zero (which is what we will be assuming shortly), the details of the shape of the charge density in it become irrelevant anyway. All that counts is the corresponding integral of that density, which in Fig. 2.14b has been denoted by Q'_I .

[‡] Finite inversion layer thickness effects are discussed in Sec. 5.11.

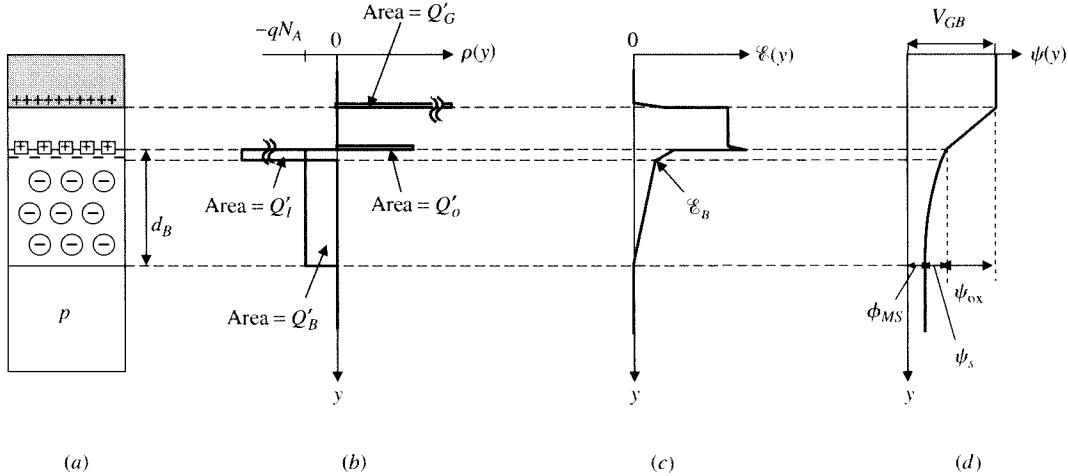


FIGURE 2.14
 (a) Two-terminal MOS structure with p -type substrate; (b) charge density; (c) electric field, all vs. depth.

The surface potential drop ψ_s is then the area under the corresponding triangle in the field plot, or $\psi_s = E_B d_B / 2$. Using in this (2.6.5) and solving for d_B , we obtain

$$d_B = \sqrt{\frac{2\epsilon_s}{qN_A}} \sqrt{\psi_s} \quad (2.6.6)$$

which is of the same form as the equation we obtained for the depletion region depth on the p side of a n^+p junction, (1.5.16). Indeed, the system consisting of the inversion layer and the p bulk behaves, not surprisingly, like a n^+p junction in some respects; we will make use of this fact to increase our intuition later on.

We have already seen that for a pure depletion region, the charge is related to the surface potential as in (2.5.3):

$$Q'_B = -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_s} \quad (2.6.7)$$

This equation could have been obtained directly under the depletion approximation, using the same approach as for the n^+p junction; indeed, it has the same form as (1.5.18).

Using the preceding equation and (2.6.1b) in (2.6.2) and solving for Q'_I , we get[†]

$$Q'_I = -\sqrt{2q\epsilon_s N_A} \left(\sqrt{\psi_s + \phi_i e^{(\psi_s - 2\phi_f)/\phi_i}} - \sqrt{\psi_s} \right) \quad (2.6.8)$$

[†] Even though (2.6.8) is in principle correct, it can have numerical problems for the following reason. The quantity in parentheses is of the form $\sqrt{\psi_s + \xi} - \sqrt{\psi_s}$, where $\xi = \phi_i \exp[(\psi_s - 2\phi_f)/\phi_i]$. When ψ_s is below $2\phi_f$,

We have plotted $|Q'_I|$ and $|Q'_B|$ vs. ψ_s in Fig. 2.15 by using the preceding two equations; their sum is $|Q'_C|$ as shown (note that in this figure and in other figures to follow we are plotting charge magnitudes, in contrast to plotting algebraic values as we had done in Fig. 2.10).[†]

It is convenient to divide the inversion region into three subregions: These are marked *weak*, *moderate*, and *strong* inversion in Fig. 2.15. To be consistent with the definition of inversion in Fig. 2.8, one can define the *onset* of weak inversion in Fig. 2.15 at $\psi_s = \phi_F$. The *upper* limit of weak inversion is defined in much of the literature at $\psi_s = 2\phi_F$. It is seen that for surface potentials less than about this value, practically all the surface charge is due to the charge in the depletion region. The corresponding inversion layer charge is too small to be shown in the scale of the figure but can nevertheless cause nonnegligible conduction when the MOS structure is part of a transistor. As ψ_s is raised above $2\phi_F$, $|Q'_I|$ starts to become significant because of the exponential in (2.6.8). For ψ_s exceeding $2\phi_F$ by a few ϕ_t , Q'_I becomes a very strong function of ψ_s . This should come as no surprise. From Fig. 2.8 we see that at $\psi_s = 2\phi_F$ the surface electron concentration is already as high as the concentration of acceptor atoms. An increase of ψ_s above $2\phi_F$ by several ϕ_t would be enough to provide very large n_{surface} because of the exponential dependence of the latter on ψ_s (2.4.12). The concentration $n(y)$ at points very close to the surface would also increase drastically, and Q'_I as seen from (2.6.4) would begin "taking off." A point around which this happens can be defined as the onset of strong inversion. This point will be above $2\phi_F$ by a quantity ϕ_{z0} , which is several ϕ_t , and will be discussed further later on.

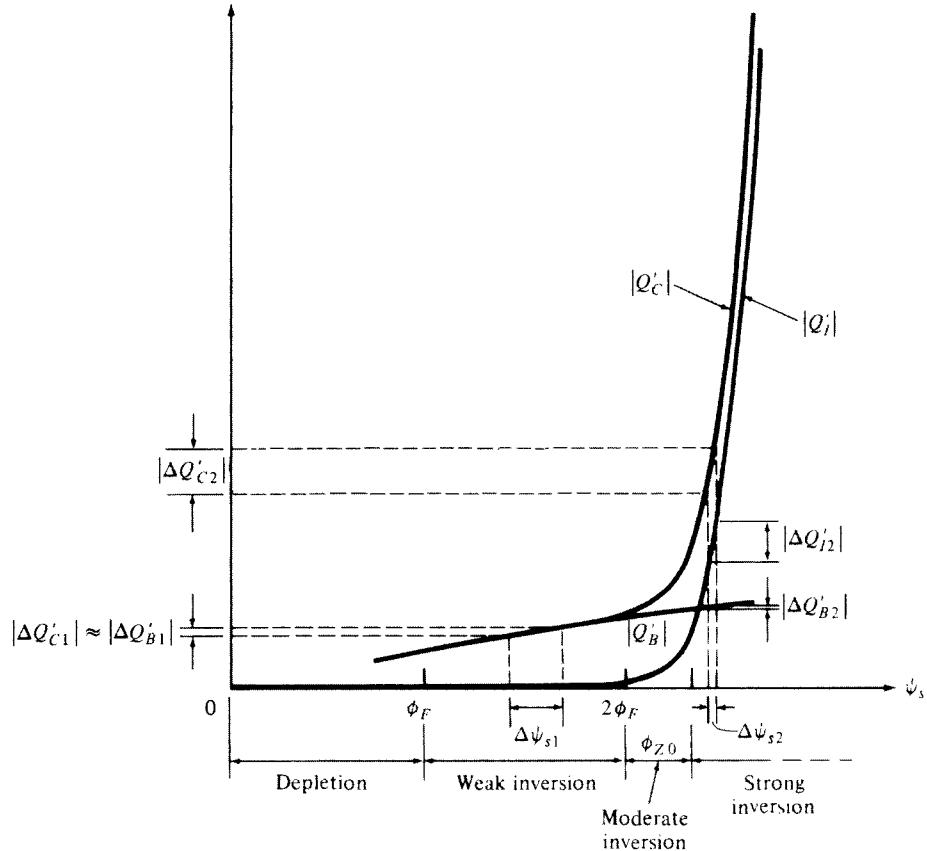
In addition to the relative contributions of Q'_I and Q'_B to Q'_C , it is very important to consider the relative contributions to *changes* $\Delta Q'_C$, associated with changes $\Delta \psi_s$ in the surface potential. Such charge changes are provided partly by a change in the inversion layer charge and partly by a change in the depletion region charge; that is, as seen from (2.6.2), we have

$$\Delta Q'_C = \Delta Q'_I + \Delta Q'_B \quad (2.6.9)$$

Let us first consider the case where V_{GB} is low, so that ψ_s and $|Q'_C|$ are also low. Assume a change ΔV_{GB1} of V_{GB} results in a change $\Delta Q'_{C1}$ as shown in Fig. 2.15. From the figure it is clear that practically all of this change is provided by a change $\Delta Q'_{B1}$ in the depletion region charge. Furthermore, the change $\Delta \psi_{s1}$ of the surface potential, required to accommodate these charge changes, is quite significant. It is

[†] ξ becomes very small, and the two square roots are of similar magnitude; this can cause significant computational error in the difference. There is an interesting trick that can be used to avoid this problem. Multiplying $\sqrt{\psi_s + \xi} - \sqrt{\psi_s}$ by $(\sqrt{\psi_s + \xi} + \sqrt{\psi_s}) / (\sqrt{\psi_s + \xi} + \sqrt{\psi_s})$ gives $\xi / (\sqrt{\psi_s + \xi} + \sqrt{\psi_s})$, which, although equal to the original quantity, no longer relies on the difference between two quantities of similar magnitude.

[‡] In computer-aided analysis, it is desirable to have, for each charge, a single equation covering the entire operation from inversion to accumulation. In that case, (2.6.7) and (2.6.8), derived assuming inversion, cannot be used; they will give wrong values in accumulation and slightly above flatband. In such cases, one can use for Q'_B the right-hand side of (2.5.1), and for Q'_I the difference between the complete Q'_C in (2.4.21) and this new Q'_B . The Q'_I obtained this way is practically identical to that obtained in inversion and decreases to negligible values in a continuous manner in depletion and accumulation.

**FIGURE 2.15**

Magnitude of inversion layer charge, depletion region charge, and their sum (all per unit area) vs. surface potential.

clear that what has just been described is characteristic of most of what has been marked “weak inversion” in Fig. 2.15, with the possible exception of points very close to its upper limit.

Assume now that V_{GB} is larger, causing a large ψ_s and a large $|Q'_C|$. Assume a change ΔV_{GB2} causes a change $\Delta Q'_{C2}$, as shown in Fig. 2.15. Things are now different. Practically all $\Delta Q'_C$ is provided by a change $\Delta Q'_{I2}$ of the inversion layer charge, $\Delta Q'_{B2}$ being negligible, and the surface potential change $\Delta\psi_{s2}$ required to accommodate these changes is very small. This behavior is evident throughout what has been marked “strong inversion” in Fig. 2.15.

We now develop an alternative expression for Q'_I , which will be found useful in subsequent developments. Using (2.4.23) and (2.6.2), we obtain

$$Q'_I = -C'_{ox}(V_{GB} - V_{FB} - \psi_s) - Q'_B \quad (2.6.10)$$

If, in this relation, we use Q'_B from (2.6.7) and the definition of γ from (2.4.26a), we obtain

$$Q'_I = -C'_{\text{ox}}(V_{GB} - V_{FB} - \psi_s - \gamma\sqrt{\psi_s}) \quad (2.6.11)$$

To predict Q'_I and Q'_B as a function of the externally applied voltage, V_{GB} , we need a relation between this voltage and the surface potential. We proceed as for the general case. To obtain this, we use (2.4.24) with Q'_C from (2.6.1b) and with γ from (2.4.26a) to obtain

$$V_{GB} = V_{FB} + \psi_s + \gamma\sqrt{\psi_s + \phi_F e^{(\psi_s - 2\phi_F)/\phi_F}} \quad (2.6.12)$$

We note that (2.6.12) is consistent with (2.6.11) and (2.6.8); in fact, any one of these three equations can be derived from the other two.

Equation (2.6.12) is plotted toward the right in Fig. 2.16, where it is shown as ψ_s vs. V_{GB} (although the equation cannot be solved explicitly for ψ_s). The plot of charges vs. ψ_s from Fig. 2.15 is repeated in the left part of Fig. 2.16, with the same ψ_s axis so that the two plots can be related. Notice that for low points on the $\psi_s(V_{GB})$ curve, significant changes $\Delta\psi_s$ are required to accommodate the charge changes demanded by ΔV_{GB} . For points high on the $\psi_s(V_{GB})$ curve, a slight $\Delta\psi_s$ is sufficient to accommodate these changes; this is because of the steepness of the $Q'_I(\psi_s)$ curve at such points. The weak inversion region is characterized by significant and nearly constant slope $d\psi_s/dV_{GB}$; in strong inversion, this slope drops to small values. The strong inversion region can be defined where ψ_s is practically constant (see below).

In terms of V_{GB} , the onsets of weak, moderate, and strong inversion will be denoted by V_{L0} , V_{M0} , and V_{H0} , as shown in Fig. 2.16. V_{L0} and V_{M0} can be found from

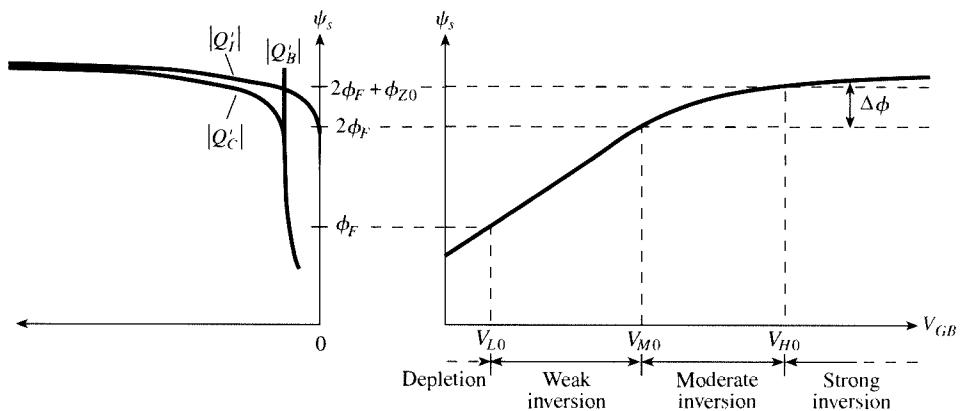


FIGURE 2.16
Surface potential vs. gate-substrate voltage (right) and charges vs. surface potential (left).

(2.6.12) by using in it, respectively, $\psi_s = \phi_F$ and $\psi_s = 2\phi_F$. Neglecting the resulting small exponential terms gives[†]

$$V_{L0} = V_{FB} + \phi_F + \gamma\sqrt{\phi_F} \quad (2.6.13a)$$

$$V_{M0} = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F} \quad (2.6.13b)$$

Strong inversion begins at a surface potential of $2\phi_F + \phi_{Z0}$, where ϕ_{Z0} is several ϕ_F , its value depending on substrate doping and oxide thickness. The corresponding value of V_{GB} at the onset of strong inversion, denoted by V_{H0} , is several tenths of 1 V above V_{M0} and will be discussed further in Sec. 2.6.2. More elaborate definitions¹⁴ for V_{M0} and V_{H0} are discussed in Appendix D.

It is desirable in the present development to find a relation of the form

$$Q'_I = Q'_I(V_{GB}) \quad (2.6.14)$$

since this would give directly the influence of the externally applied voltage (the “cause”) on the resulting inversion layer charge (the final “effect” we are interested in, since Q'_I will result in current flow in the MOS transistor). Unfortunately, this has to be done numerically if a value for V_{GB} is given and the value for Q'_I is desired. For our purposes, instead of (2.6.14), we can consider its parametric representation, which consists of (2.6.8) and (2.6.12), i.e., the relations that were plotted in Fig. 2.16. If values are assumed for ψ_s , the corresponding Q'_I and V_{GB} can be found from these equations. Q'_I can be plotted vs. V_{GB} as in Fig. 2.17. The broken line in the figure will be discussed in the following subsection.

We will now consider each region of inversion (weak, moderate, and strong) separately. We will identify the dominant phenomena in each region and show certain approximations that can be made in each case. We will seek to develop explicit expressions of the form of (2.6.14), based on such approximations. This will be desirable not only for ease of computation but also because the functional form of the resulting approximate expressions will help make evident the dominant phenomena in the regions under consideration. Also, such expressions will be needed for developing simple models for the MOS transistor in subsequent chapters, which are used widely in circuit design.

2.6.2 Strong Inversion

As seen in Fig. 2.16, in strong inversion large changes in V_{GB} result in very small changes in ψ_s . A very common assumption is that in strong inversion ψ_s is practically “pinned” to a constant value:

[†] Readers of past editions of this book have wondered how the subscripts in these symbols were chosen. M was chosen to stand for the onset of moderate inversion; for the onset of strong inversion, though, we could not choose S , as this is reserved for the source of the MOS transistor later on. Thus, H was chosen, standing for “high” or “heavy.” The opposite of these terms is “low” or “light,” so for consistency L was chosen to denote the onset of weak inversion. The reason for the zero subscript will become obvious in Chap. 3.

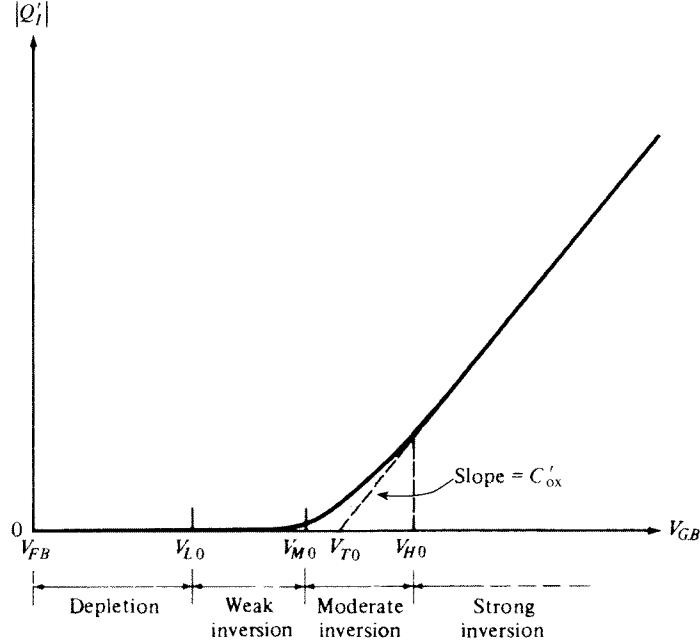


FIGURE 2.17
Magnitude of inversion layer charge per unit area vs. gate-substrate voltage.

$$\psi_s \approx \phi_0 \quad (2.6.15)$$

The value of ϕ_0 is close to $2\phi_F + \phi_{Z0}$ in Fig. 2.16 but is difficult to define exactly. Obviously, if one is interested in operating points limited only to high V_{GB} , the value chosen for ϕ_0 would be somewhat larger than that needed for low-voltage operation.[†] Usually a single value is chosen for ϕ_0 as a compromise.

Historically, the most commonly used value is

$$\phi_0 = 2\phi_F \text{ (a bad choice)} \quad (2.6.16)$$

It is difficult to justify the use of this value, looking at Fig. 2.16. From that figure, it is clear that at $\psi_s = 2\phi_F$ the plot of $\psi_s(V_{GB})$ has not flattened out yet, and that in strong inversion the value of ψ_s is considerably larger than $2\phi_F$. Thus, a more appropriate value for ϕ_0 is

$$\phi_0 = 2\phi_F + \Delta\phi \quad (2.6.17)$$

[†] The situation is analogous to assuming a constant value for the base-emitter voltage for bipolar transistors. The value chosen depends on the context and the application and can be expected to be higher in power circuit work than in micropower circuit work.

where $\Delta\phi$ will be considered constant for simplicity and, for the case of uniform substrates considered so far, is several ϕ_t (for uniform substrates about $5\phi_t$ to $6\phi_t$ can be used, which is an average for a variety of practical process parameters and V_{GB} values). For nonuniform substrates, the analysis presented in this chapter is, in principle, not valid. Nevertheless, MOS transistor equations based on this analysis (Chap. 4) are often used for devices with nonuniform substrates and in such equations ϕ_0 appears (Chap. 9). In such cases, to achieve reasonable agreement with experimental results, $\Delta\phi$ must be modified from the estimate given previously; it may or may not be close to zero, or it may even be negative. In any case, the historical value of $2\phi_F$ is, in general, not the best choice for ϕ_0 , so (2.6.17) should be used, with an appropriate choice for $\Delta\phi$. An approximation for $\Delta\phi$ in strong inversion is discussed in Prob. 2.5.

Equation (2.6.15) implies that the depletion region width d_B is assumed to reach a maximum value d_{Bm} and not to increase with V_{GB} anymore. From (2.6.6) and (2.6.15),

$$d_{Bm} = \sqrt{\frac{2\epsilon_s}{qN_A}} \sqrt{\phi_0} \quad (2.6.18)$$

Similarly, the depletion region charge is assumed to have reached a maximum value, Q'_{B0} , given by using $\psi_s = \phi_0$ in (2.6.7):

$$Q'_{B0} = -\sqrt{2q\epsilon_s N_A} \sqrt{\phi_0} \quad (2.6.19)$$

The inversion layer charge Q'_I can be found from (2.6.10), by using in it $\psi_s = \phi_0$ and the preceding result for Q'_B . Lumping all terms other than V_{GB} in the parentheses together, we can write that equation in the form

$$Q'_I = -C'_{ox}(V_{GB} - V_{T0}) \quad (2.6.20)$$

where, by comparing this equation to (2.6.10) and using the definition of V_{FB} from (2.2.6), we obtain

$$V_{T0} = \phi_{MS} - \frac{Q'_o}{C'_{ox}} + \phi_0 - \frac{Q'_{B0}}{C'_{ox}} \quad (2.6.21)$$

or, by comparing (2.6.20) to (2.6.11) instead and using (2.6.15), we obtain[†]

$$V_{T0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0} \quad (2.6.22)$$

[†] We give more than one way to obtain a result, and more than one way to write it, in an attempt to provide a feel for the relations between the many equations presented in this chapter, and also to present some intermediate results that will be needed in subsequent chapters. This practice will be repeated; we ask for the reader's patience.

$|Q'_I|$ from (2.6.20) is shown by the broken line in Fig. 2.17, where it can be compared with the accurate $|Q'_I(V_{GB})|$ (solid line). The quantity V_{T0} is called the *extrapolated threshold voltage* of the MOS two-terminal structure, for obvious reasons. The fact that the plot of $Q'_I(V_{GB})$ is practically a straight line in strong inversion will prove to be crucial in shaping the properties of the structure in this region.

Example 2.3. Estimate V_{T0} for the process of Example 2.2.

In Example 2.2, we found $V_{FB} = -1.043$ V. Using the data of that example in (2.4.26), we find $\gamma = 0.337 \text{ V}^{1/2}$. Let us use $\phi_0 = 2\phi_F + 6\phi_t = 1.107$ V. Then (2.6.22) gives

$$V_{T0} = -1.042 \text{ V} + 1.107 \text{ V} + 0.337\sqrt{1.107} \text{ V} = 0.419 \text{ V}$$

It should be emphasized that, whereas V_{T0} is a quantity that appears in the strong inversion expression (2.6.20), the MOS structure is *not* in strong inversion at $V_{GB} = V_{T0}$; this crucial observation becomes apparent by examining Fig. 2.17.

We can now return to the discussion of the onset of strong inversion. As seen in Fig. 2.17, the transition of the exact Q'_I plot toward the straight-line behavior described by (2.6.20) is *very* gradual. No critical point can be identified that could conveniently be taken as the onset of strong inversion. Thus, a possible definition of this onset can be the minimum V_{GB} value for which (2.6.20) predicts Q'_I within an acceptable error. The value one ends up with depends, of course, on what is meant by “acceptable error.” For some applications, a 10 percent error in the value of Q'_I may be tolerated. Then the onset of strong inversion can be taken relatively close to V_{T0} . If only a 2 percent error can be tolerated, one should take the onset at a somewhat larger V_{GB} value. Finally, for some applications (e.g., the evaluation of small-signal capacitances and conductances in Chap. 7), accuracy is desirable not only for the plot of Q'_I vs. V_{GB} but also for the *slope* of this plot. Then the onset of strong inversion should be defined accordingly. As is evident from the figure, at points where the slope is accurately predicted by the straight line, Q'_I itself is accurately predicted by it. Thus, definitions in terms of the slope are conservative. If a slope error of about 10 percent can be tolerated, it turns out that the onset of strong inversion V_{H0} should be taken about 0.6 V above V_{M0} at room temperature and for uniform substrates; as already discussed in relation to (2.6.17), different considerations apply for nonuniform substrates (Chap. 9). There are ways to define and calculate V_{M0} and V_{H0} precisely, if desired.¹⁴ These are somewhat elaborate and are given for the interested reader in Appendix D.

The choice $\phi_0 = 2\phi_F$ (which, as we have already discussed, is in general not the best) makes V_{T0} in (2.6.22) identical to the upper limit of weak inversion, V_{M0} in (2.6.13b):

$$V_{T0} = V_{M0} \quad \text{if } \phi_0 = 2\phi_F \text{ is used (a bad choice)} \quad (2.6.23)$$

In fact, in some of the literature, the onset of strong inversion is taken as V_{M0} , and no distinction is made between the quantities V_{M0} , V_{T0} , and V_{H0} . In such treatments, all three quantities are taken to be the same point, called “threshold,” and assumed

to be given by (2.6.13b). No moderate inversion region is defined in such cases, and the strong-inversion region is taken to be adjacent to weak inversion. This practice originates in the early days of MOS work, when large V_{GB} values were common. If, for example, the actual values of V_{M0} , V_{T0} , and V_{H0} happened to be 1.0, 1.2, and 1.6 V, respectively, and the V_{GB} values of interest were 20 to 30 V, then using 1.0 or 1.6 V in lieu of V_{T0} in (2.6.20) would not result in excessive error. However, with V_{GB} values limited to about 1.2 V in today's leading technologies, and with a constant drive to even lower voltages, a careful distinction between the three voltages is necessary.

2.6.3 Weak Inversion

In weak inversion $Q'_I(V_{GB})$ will be shown to assume a simple form, albeit very different from that in strong inversion. To start, consider (2.6.8) and define

$$\xi = \phi_t e^{(\psi_s - 2\phi_F)/\phi_t} \quad (2.6.24)$$

Consider the term $\sqrt{\psi_s + \xi}$ in (2.6.8). In weak inversion, ψ_s is smaller than $2\phi_F$, as seen in Fig. 2.16. Thus $\xi \ll \psi_s$. The function $\sqrt{\psi_s + \xi}$ can be approximated by the first two terms of its Taylor expansion around $\xi = 0$:[†]

$$\sqrt{\psi_s + \xi} = \sqrt{\psi_s} + \frac{1}{2\sqrt{\psi_s}} \xi \quad (2.6.25)$$

Using the previous two equations in (2.6.8), we obtain

$$Q'_I \approx -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_s}} \phi_t e^{(\psi_s - 2\phi_F)/\phi_t} \quad (2.6.26)$$

To obtain a relation between Q'_I and the external bias V_{GB} , we need to relate ψ_s in the preceding equation to V_{GB} . This turns out to be easy. Because in weak inversion the inversion layer charge is extremely small compared with the depletion region charge, it does not influence the surface potential. The total semiconductor charge is practically equal to the depletion region charge, and the equations developed earlier for depletion apply here; in particular, ψ_{sa} in (2.5.5) was seen to be the surface potential obtained deep in depletion, for a given V_{GB} ; this is precisely what we need. Thus, let us assume

$$\psi_s \approx \psi_{sa} = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2, \quad \text{weak inversion} \quad (2.6.27)$$

[†] Equation (2.6.25) can also be obtained without a series expansion, by using the alternative form of (2.6.8) discussed in the footnote to the discussion of that equation, by letting ξ become small.

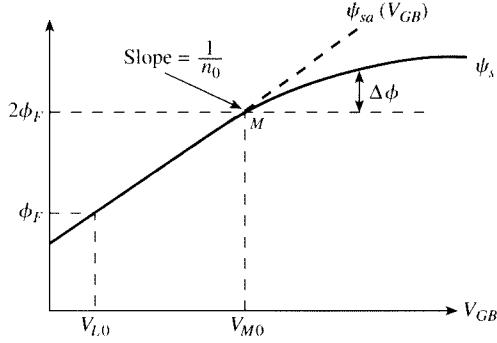


FIGURE 2.18

Surface potential and potential ψ_{sa} vs. gate-substrate voltage.

In Fig. 2.18, we repeat the $\psi_s(V_{GB})$ plot from Fig. 2.16, and we show $\psi_{sa}(V_{GB})$ for comparison. (A similar comparison has already been made in Fig. 2.12, but we have marked the plots in Fig. 2.18 with some quantities relevant to our discussion of weak inversion.)

It is seen that the agreement in weak inversion is very good, except perhaps at points very close to the upper limit of the region. We can thus use $\psi_s \approx \psi_{sa}$ in (2.6.26) to obtain:

$$Q'_I = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_{sa}(V_{GB})}} \phi_t e^{[\psi_{sa}(V_{GB}) - 2\phi_t]/\phi_t} \quad (2.6.28)$$

This equation, with $\psi_{sa}(V_{GB})$ given by (2.6.27), is an *explicit* expression of Q'_I in terms of V_{GB} .

It can be seen in Fig. 2.18 that the slope of ψ_{sa} (and thus of ψ_s in weak inversion) vs. V_{GB} is almost constant. The *inverse* of this slope is often denoted by n . Thus

$$n \equiv \left(\frac{d\psi_{sa}}{dV_{GB}} \right)^{-1} \quad (2.6.29)$$

The value of n can be found in terms of V_{GB} by using (2.6.27) in the preceding definition, or, easier, by using $\psi_s \approx \psi_{sa}$ in (2.5.4) and differentiating the resulting equation. We obtain

$$n = 1 + \frac{\gamma}{2\sqrt{\psi_{sa}(V_{GB})}} \quad (2.6.30)$$

As seen, n is larger than unity.[†] This fact could have been anticipated intuitively, since changes in V_{GB} appear partly as changes in ψ_{ox} and partly as changes in ψ_s , and thus $\Delta\psi_s < \Delta V_{GB}$. Since $\psi_s \approx \psi_{sa}$, this means $\Delta\psi_{sa}/\Delta V_{GB} < 1$, which implies $n > 1$ from (2.6.29). Typically, n is between 1 and 1.5.

[†] If the interface trap charge (Sec. 2.2) varies with the surface potential, an extra term will be needed in this equation. This is shown at the end of Sec. 2.7.

The relation of Q'_I to V_{GB} , given by (2.6.28) and (2.6.27), is rather complicated. A simpler approximate form can be obtained.¹⁵⁻¹⁷ First, note that when V_{GB} is varied, the variation of $\sqrt{\psi_{sa}}$ in (2.6.28) is negligible when compared with the drastic variation of the exponential term in that equation; thus we assume that $\sqrt{\psi_{sa}}$ is practically fixed and we can replace that term by $\sqrt{\phi_x}$, where ϕ_x is the value of the surface potential at any convenient point in weak inversion. It is common to choose^{15,16} $\phi_x = 2\phi_F$:[†]

$$Q'_I \approx -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{2\phi_F}} \phi_I e^{[\psi_{sa}(V_{GB}) - 2\phi_F]/\phi_I} \quad (2.6.31)$$

This equation can be further simplified by taking advantage of the almost constant slope of ψ_{sa} with V_{GB} , a fact already discussed. At $\psi_{sa} = 2\phi_F$, the value of n [which is the inverse of the slope $d\psi_{sa}/dV_{GB}$, as seen from (2.6.29)] is, from (2.6.30),

$$n_0 = n \Big|_{\psi_{sa} = 2\phi_F} = 1 + \frac{\gamma}{2\sqrt{2\phi_F}} \quad (2.6.32)$$

Thus, taking changes *below* point M in Fig. 2.18, we can write

$$\psi_{sa} - 2\phi_F = \frac{1}{n_0} (V_{GB} - V_{M0}) \quad (2.6.33)$$

Using this in (2.6.31), we obtain

$$Q'_I \approx Q'_{M0} e^{(V_{GB} - V_{M0})/(n_0\phi_I)} \quad (2.6.34)$$

with

$$Q'_{M0} = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{2\phi_F}} \phi_I \quad (2.6.35)$$

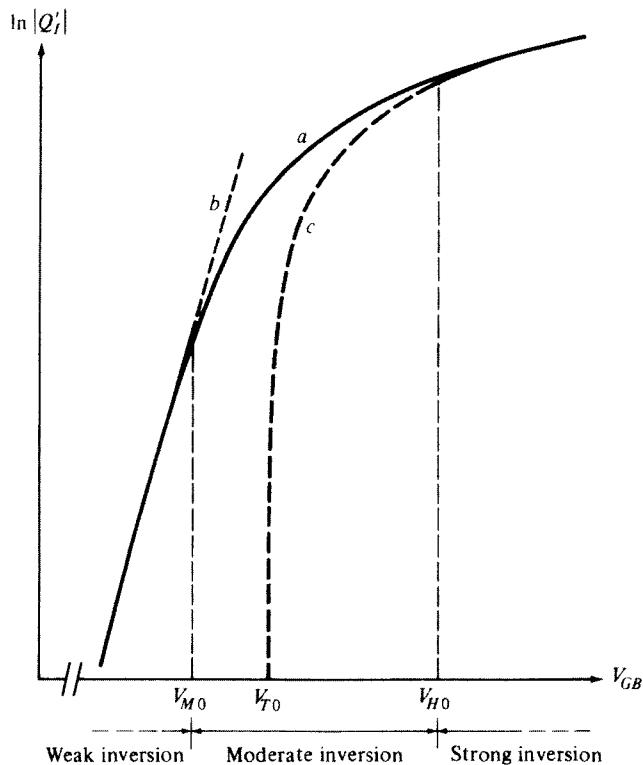
Q'_{M0} represents the value of Q'_I at the upper limit of weak inversion ($V_{GB} = V_{M0}$).[‡] Here one may worry that, at that point, the slope of $\psi_s(V_{GB})$ is not representative of the slope lower in the weak inversion region (Fig. 2.18). However, note that n_0 in the preceding development was *not* defined in terms of $\psi_s(V_{GB})$, but rather in terms of $\psi_{sa}(V_{GB})$, the slope of which does *not* change drastically at $V_{GB} = V_{M0}$, as seen in Fig. 2.18.

In Fig. 2.19, we have plotted $\ln |Q'_I|$ vs. V_{GB} , using (2.6.34) (line *b*).[§] The solid line represents the accurate equations (2.6.8) and (2.6.12) and describes the structure correctly for all regions of inversion. Over the weak-inversion region, the agreement

[†] Obviously, a more appropriate point would be $\phi_x = 1.5\phi_F$, which lies at the middle of weak inversion.¹⁷ However, choosing $\phi_x = 2\phi_F$ is consistent with the literature and with the level of approximation we are seeking at this point.

[‡] As indicated in (2.6.23), V_{T0} coincides with this limit if ϕ_0 is taken equal to $2\phi_F$. The appropriateness of this choice has been questioned in Sec. 2.6.2.

[§] In showing the logarithm of various quantities in this book, we assume that these quantities have been normalized to their units, so that they are rendered dimensionless before taking their logarithm. For simplicity, we do not define new symbols for these dimensionless quantities.

**FIGURE 2.19**

Logarithm of inversion layer charge magnitude per unit area vs. gate-substrate voltage. (a) Equations (2.6.8) and (2.6.12); (b) (2.6.34), (c) (2.6.20).

is seen to be very good except at points very close to the upper limit of the region. In weak inversion, the plot of $\ln |Q'_l|$ is seen to be practically a straight line, verifying that $|Q'_l|$ is nearly exponential with respect to V_{GB} .

As has been seen, the development of (2.6.34) involved a number of approximations. Thus, although this equation and Fig. 2.19 suggest that $Q'_l(V_{GB})$ is exponential in weak inversion, it should be kept in mind that this is not *exactly* so. Indeed, more accurate equations, such as (2.6.8) with (2.6.12), or even (2.6.28), predict minute deviations from "exponentiality." Although such deviations are too small to be visible in the plot of $\ln |Q'_l|$ vs. V_{GB} in Fig. 2.19, they can make a difference in a plot of the *slope* of $\ln |Q'_l|$ vs. V_{GB} . This affects the evaluation of small-signal quantities (Sec. 2.7 and Chap. 7). In order to take this effect into account, the quantity n_0 is sometimes allowed to be a weak function of V_{GB} as in (2.6.30).

2.6.4 Moderate Inversion

In moderate inversion, none of the simplifications discussed previously is valid. The plot of $Q'_l(V_{GB})$ here is neither a straight line nor an exponential. If accurate results

are needed in this region, the complete equations (2.6.8) and (2.6.12) can be used. However, this requires a numerical solution for a given V_{GB} value, as (2.6.12) is an implicit equation in ψ_s . Efforts have been made to produce semiempirically explicit expressions for ψ_s and for the charges in terms of V_{GB} directly.^{7,18-23} Such efforts involve several approximations.

As has already been noted, in many treatments no moderate-inversion region is defined. Sometimes this region is considered the bottom part of strong inversion. In some treatments, a point is defined, falling somewhere in our moderate-inversion region, and $Q'_I(V_{GB})$ is assumed to be exponential directly below and a straight line directly above that point. Such models can lead to large errors. This can be seen in Fig. 2.19, where Q'_I as calculated from (2.6.8) and (2.6.12) is compared with the approximate equations (2.6.34) and (2.6.20). *Neither* of these two equations provides satisfactory accuracy in moderate inversion.

2.7 Small-Signal Capacitance

If V_{GB} is increased by a small amount ΔV_{GB} in Fig. 2.5a, a positive charge[†] $\Delta Q'_G$ will flow into the gate terminal. For overall charge neutrality, a charge of equal value must flow out of the body terminal or, equivalently, a charge of value $-\Delta Q'_G$ must flow *into* the body terminal. An incremental (small-signal) capacitance per unit area, C'_{gb} , can thus be defined to relate charge changes to voltage changes. This is illustrated in Fig. 2.20. We define

$$C'_{gb} \equiv \frac{dQ'_G}{dV_{GB}} \quad (2.7.1)$$

The charge $-\Delta Q'_G$ flowing into the body goes to change the semiconductor charge Q'_C by an amount $\Delta Q'_C$:

$$\Delta Q'_C = -\Delta Q'_G \quad (2.7.2)$$

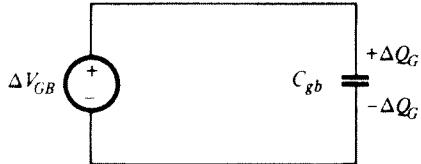
The gate-to-body voltage change will be distributed partly across the oxide (as $\Delta\psi_{ox}$) and partly across the semiconductor (as a change $\Delta\psi_s$ in the surface potential). Equation (2.3.2) is repeated here for convenience:

$$\Delta V_{GB} = \Delta\psi_{ox} + \Delta\psi_s \quad (2.7.3)$$

Taking $1/C'_{gb}$ from (2.7.1) and using (2.7.3) in the result (with small changes replaced by differentials), we have

$$\frac{1}{C'_{gb}} = \frac{d\psi_{ox}}{dQ'_G} + \frac{d\psi_s}{dQ'_G} \quad (2.7.4)$$

[†] We use the term “charge” rather than “charge per unit area” for simplicity. The fact that we are talking about charges *per unit area* will be clear from the primes used in the symbols for these quantities. The same applies to capacitances.

**FIGURE 2.20**

Small-signal equivalent circuit for a two-terminal MOS structure driven by a voltage source.

The preceding equation can be written, by using (2.7.2), as

$$\frac{1}{C'_{gb}} = \frac{1}{dQ'_G/d\psi_{ox}} + \frac{1}{-dQ'_C/d\psi_s} \quad (2.7.5)$$

To interpret this equation, note first that from (2.4.22) we have

$$\frac{dQ'_G}{d\psi_{ox}} = C'_{ox} \quad (2.7.6)$$

The quantity $-dQ'_C/d\psi_s$ in (2.7.5) can be interpreted with the help of Fig. 2.5a as follows. If the potential across the semiconductor is changed by $\Delta\psi_s$, the charge in that region will change by $\Delta Q'_C$. This additional charge must enter the region through its *bottom*, coming from the body terminal. If now a capacitor had across it a voltage $\Delta\psi_s$, and the value of its capacitance were adjusted so that a charge $\Delta Q'_C$ had to flow into its *bottom* plate, the situation illustrated in Fig. 2.21 would result. From this figure, we clearly have

$$C'_c \equiv -\frac{dQ'_C}{d\psi_s} \quad (2.7.7)$$

The denominator of the last fraction in (2.7.5), therefore, can be interpreted as a small-signal capacitance corresponding to the semiconductor charge region. It relates the changes of the potential across that region to the corresponding changes in its charge.

Using now (2.7.6) and (2.7.7) in (2.7.5), we get

$$\frac{1}{C'_{gb}} = \frac{1}{C'_{ox}} + \frac{1}{C'_c} \quad (2.7.8)$$

Therefore, the small-signal capacitance C_{gb} is the same as that exhibited by two capacitors of values C_{ox} and C_c connected in series as in Fig. 2.22.

Before using the preceding equation, we need to determine C'_c . Using (2.4.21) in (2.7.7) gives

$$C'_c = \text{sgn}(\psi_s) \sqrt{2q\epsilon_s N_A} \frac{1 - e^{-\psi_s/\phi_t} + e^{-2\phi_t/\phi_t} (e^{\psi_s/\phi_t} - 1)}{2\sqrt{\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-2\phi_t/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t)}}, \quad \psi_s \neq 0$$

$$= \sqrt{q\epsilon_s N_A (1 + e^{-2\phi_t/\phi_t})/\phi_t}, \quad \psi_s = 0 \quad (2.7.9)$$

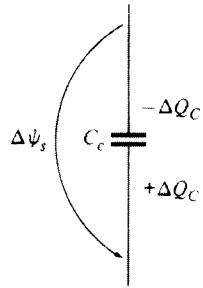


FIGURE 2.21
Illustrating the concept of semiconductor charge region small-signal capacitance.

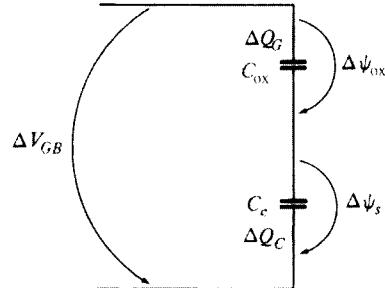


FIGURE 2.22
Circuit representation of (2.7.8).

At $\psi_s = 0$, both the numerator and the denominator in the first branch of this expression become zero; we have thus used l'Hôpital's rule to find C'_c at that point, as given by the second branch (Prob. 2.18). An alternative form of this equation corresponding to (2.4.17) is given in Appendix C.

C'_c is plotted vs. ψ_s in Fig. 2.23. We can easily relate the shape of this curve to the negative of the slope of Q'_c in Fig. 2.10, as expected from (2.7.7). The rest of the plots in the figure will be discussed shortly.

C'_c Deep in Accumulation Deep in accumulation (for ψ_s below zero by several ϕ_t), (2.7.9) can be simplified by dropping negligible terms, as we have already done for the charge in Sec. 2.5. We get

$$C'_c = \sqrt{\frac{q\epsilon_s N_A}{2\phi_t}} e^{-\psi_s/\phi_t} \quad (2.7.10)$$

This result could also have been obtained using the accumulation charge expression, (2.5.2), directly in the definition (2.7.7). Although here differentiating a simplified charge expression gives an appropriate result under our simplifying assumptions,[†] doing so in general is risky; small inaccuracies in a charge can lead to large inaccuracies in the derivative of that charge. We will see a case of this shortly.

C'_c Deep in Depletion and in Inversion Assume now that ψ_s is above zero by several ϕ_t . Dropping negligible terms in (2.7.9), we get

$$C'_c = \sqrt{2q\epsilon_s N_A} \frac{1 + e^{(\psi_s - 2\phi_t)/\phi_t}}{2\sqrt{\psi_s + \phi_t} e^{(\psi_s - 2\phi_t)/\phi_t}}, \quad \psi_s > 5\phi_t \quad (2.7.11)$$

[†] These have included, explicitly or implicitly, the assumption that the surface of the semiconductor is not degenerate. Surface degeneracy and other effects can make (2.7.10) inaccurate. Nevertheless, this expression is well behaved and provides qualitatively the correct behavior, including approaching infinity as ψ_s becomes increasingly negative.

It is interesting to consider the individual contributions of the depletion region and inversion layer charges to C'_c . Equation (2.6.9) is repeated below

$$\Delta Q'_C = \Delta Q'_B + \Delta Q'_I \quad (2.7.12)$$

Using this equation in (2.7.7), we obtain

$$C'_c = \frac{-dQ'_B}{d\psi_s} + \frac{-dQ'_I}{d\psi_s} \quad (2.7.13)$$

We have then separated the total semiconductor capacitance C'_c into two components, one related to the depletion region charge and one related to the inversion layer charge. With a reasoning analogous to that preceding (2.7.7), we can define a depletion region incremental capacitance per unit area:

$$C'_b \equiv -\frac{dQ'_B}{d\psi_s} \quad (2.7.14)$$

This capacitance relates changes of the potential across the depletion region to the associated changes of the depletion charge.

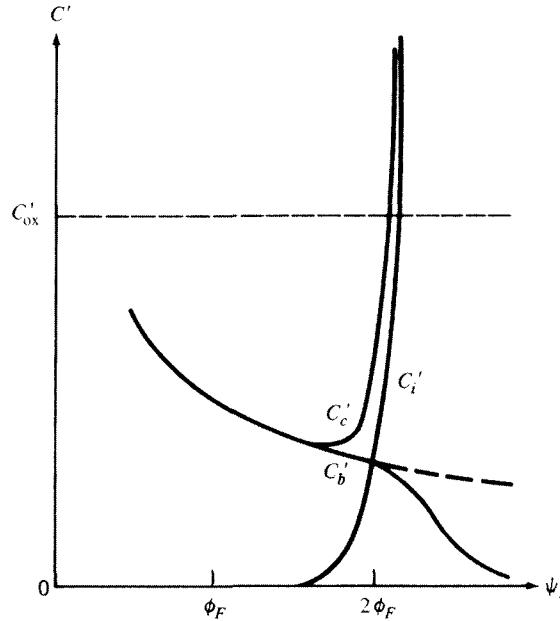
Let us also define a capacitance per unit area associated with the inversion layer. This capacitance should relate changes in the charge of that layer to the associated potential changes. In analogy with (2.7.14),

$$C'_i \equiv -\frac{dQ'_I}{d\psi_s} \quad (2.7.15)$$

From these definitions it is seen that C'_b and C'_i are the slopes of plots $|Q'_B|$ vs. ψ_s and $|Q'_I|$ vs. ψ_s , respectively. However, to evaluate C'_b and C'_i , one would have to use very accurate expressions for Q'_B and Q'_I , since even a small error in predicting a certain function can cause a severe error in predicting its derivative. For example, consider $|Q'_B|$ as predicted by the charge sheet model in (2.6.7) and plotted in Fig. 2.15. The corresponding C'_b is shown by the broken line in Fig. 2.23. A more exact model would be based on the general analysis outlined in Sec. 2.4.4, which allows for the spreading of the inversion layer into the depletion region and for the presence of holes there. Such a model gives a $|Q'_B|$ plot that, in moderate and strong inversion, is flatter than that shown in Fig. 2.15, becoming practically horizontal in very strong inversion. The corresponding error in the approximate $|Q'_B|$ is very small; yet, the slopes of the exact and approximate plots are very different (see solid line for C'_b in Fig. 2.23). The resulting exact C'_b and C'_i are given by (Appendix C)

$$C'_b = \sqrt{2q\epsilon_s N_A} \frac{1}{2\sqrt{\psi_s + \phi_i e^{(\psi_s - 2\phi_i)/\phi_i}}} \quad (2.7.16)$$

$$C'_i = \sqrt{2q\epsilon_s N_A} \frac{e^{(\psi_s - 2\phi_i)/\phi_i}}{2\sqrt{\psi_s + \phi_i e^{(\psi_s - 2\phi_i)/\phi_i}}} \quad (2.7.17)$$

**FIGURE 2.23**

Small-signal capacitances per unit area vs. surface potential. C'_i is the inversion layer capacitance; C'_b is the depletion region capacitance (solid line is exact; broken line is as predicted by the charge sheet model); C'_c is the semiconductor charge region capacitance; C'_ox is the oxide capacitance.

It is easy to see that, using Q'_B as predicted by the charge sheet model in (2.6.7), we would have obtained (2.7.16) without the exponential term; that term, however, becomes dominant in strong inversion. Note also that using the accurate formulas just discussed we obtain $C'_i = C'_b$ at $\psi_s = 2\phi_F$.

Using (2.7.14) and (2.7.15) in (2.7.13), we obtain

$$C'_c = C'_b + C'_i \quad (2.7.18)$$

which is verified by the expressions we have just discussed. Equation (2.7.8) becomes

$$\frac{1}{C'_{gb}} = \frac{1}{C'_\text{ox}} + \frac{1}{C'_b + C'_i}$$

(2.7.19)

which can be represented by the circuit of Fig. 2.24. It should be kept in mind that this is a *small-signal equivalent* circuit, relating small *changes* of potentials and charges around a bias point. It does *not* relate total values of potentials and charges.

Our ultimate goal in this development is to plot the total capacitance seen externally (C'_{gb}) vs. the total externally applied bias (V_{GB}). This can be done by giving values to ψ_s , calculating V_{GB} from (2.4.25) and C'_c from (2.7.9), and using the result

in (2.7.8). The resulting plot is shown by the solid line in Fig. 2.25. Let us consider some important features of this plot.

Deep in accumulation, C'_{ox} is seen to approach C'_{ox} . The reason for this is as follows. As ψ_s becomes more and more negative, C'_c in (2.7.10) becomes very large, corresponding to the steep slope of Q'_C in Fig. 2.10; once there are plenty of holes near the surface, a small change in surface potential is enough to drastically change their density. Thus the last term in the sum in (2.7.8) becomes negligible, and the total capacitance C'_{gb} reduces to approximately C'_{ox} . Intuitively, this should make sense. Deep in accumulation, there is an abundance of holes that can provide a conducting path from the bottom of the body, through the semiconductor, to the surface; the negative surface potential attracts a large number of holes immediately below the oxide, in a sense forming the bottom “plate” of the oxide capacitor. As a result, deep in accumulation, the total incremental capacitance seen between the two terminals of the MOS junction is basically that of the oxide, C'_{ox} . To the right of that region, C'_{ox} is seen to vary considerably with V_{GB} ; this property makes possible the use of the two-terminal MOS structure as a “varactor,” i.e., a bias-controlled small-signal capacitor, that is used in radio-frequency circuits. Such operation requires careful modeling.²⁴

In the weak inversion region, the plot in Fig. 2.25 is seen to reach a minimum. This is because in weak inversion (except for points close to the upper limit of the region), the inversion layer capacitance is negligible, as we can see from Fig. 2.23. From (2.7.19), then, C'_{gb} is basically the series combination of C'_{ox} and C'_b . As V_{GB} is increased, C'_b becomes smaller, as seen from Fig. 2.23. Therefore, the series combination of C'_{ox} and C'_b also decreases, as seen in Fig. 2.25. Above weak inversion, C'_i becomes significant and drastically increases if V_{GB} is raised further (Fig. 2.23). This capacitance is in parallel with C'_b . The last fraction in (2.7.19) decreases drastically, and C'_{gb} approaches C'_{ox} . Physically, an abundance of electrons exists at high V_{GB} immediately below the oxide and provides the bottom “plate” of the oxide capacitor, just as an abundance of holes provided that plate in the case of accumulation.

The previous discussion has been in terms of “static” changes, i.e., it was assumed that after V_{GB} is changed by a small amount ΔV_{GB} , it remains fixed at its new value. We then wait long enough for a new equilibrium to be reached, and we record the changes in the various potentials and charges. If ΔV_{GB} is a small-signal sinusoidal voltage, the steady-state charge changes will also be sinusoidal. They will correspond to equilibrium values if the frequency is low enough (e.g., 1 Hz). However, if the frequency is high (e.g., 100 kHz), things will be different, as shown by the broken line in Fig. 2.25. The inversion layer charge now cannot keep up with the fast-changing ΔV_{GB} (see discussion in the following paragraph), and the required charge changes must be provided by covering or uncovering acceptor atoms at the bottom of the depletion region, just as in the case of depletion operation. Thus, as in the depletion case, C'_{gb} is effectively the series combination of C'_{ox} and C'_b . In strong inversion, the width of the depletion region reaches a maximum value, and C'_b attains a limit value, which can be found approximately by using (2.6.7) in (2.7.14) and using $\psi_s \approx \phi_0$ in the result (see Sec. 2.6.2). Calculating then

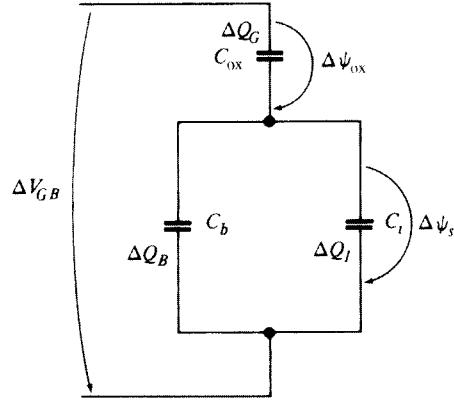


FIGURE 2.24

Small-signal equivalent circuit for the two-terminal MOS structure, showing individual capacitances, and assuming zero interface trap charge and very low frequency operation.

the corresponding minimum value of C'_{gb} , denoted by $C'_{gb,l}$ in Fig. 2.25, from (2.7.8) gives (Prob. 2.18)

$$C'_{gb,l} \approx \frac{C'_\text{ox}}{1 + (2/\gamma)\sqrt{\phi_0}} \quad (2.7.20)$$

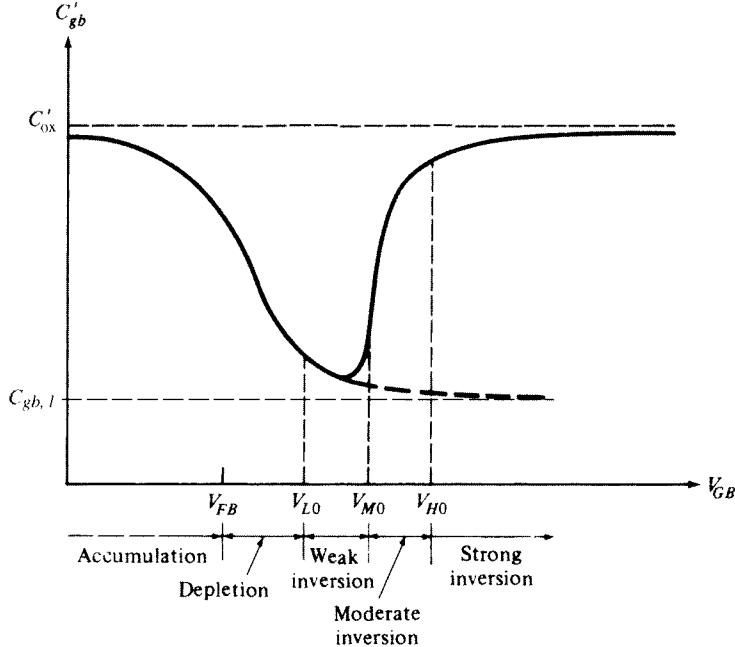
This limit value is not quite reached by the low-frequency capacitance, as seen in Fig. 2.25 (solid line). A far more detailed description of high-frequency capacitance behavior, which can be highly complex, can be found in Ref. 1.

The reason the inversion layer charge cannot follow fast enough if the frequency is high is that it is in a sense isolated from the outside world by the oxide on top and the depletion region below. Therefore, the electron concentration there can be changed only by the mechanisms of thermal generation and recombination, which in this case are very slow (no external irradiation is assumed). If, instead, communication with the outside world were possible, in the sense that inversion layer charge could be provided or removed externally, then the behavior exhibited by the solid curve in strong inversion would persist up to much higher frequencies. This communication with the outside world is provided by the source and drain regions in a MOS transistor, as discussed in subsequent chapters.

The incremental capacitances defined in this section can be used to provide convenient expressions for the slopes of various plots in Sec. 2.6. Thus, using (2.4.24), (2.6.2), (2.7.14) and (2.7.15), we have

$$\frac{d\psi_s}{dV_{GB}} = \frac{C'_\text{ox}}{C'_\text{ox} + C'_b + C'_i} \quad (2.7.21)$$

which could also have been obtained from the equivalent circuit in Fig. 2.24. This relation gives the slope of the ψ_s vs. V_{GB} plot of Fig. 2.16. In strong inversion $|C'_i|$

**FIGURE 2.25**

Total gate-body capacitance per unit area vs. gate-body bias. Solid line: "static" behavior; broken line: high-frequency behavior.

becomes very large and the slope diminishes in agreement with that plot. Similarly, the slope of the $|Q'_I|$ vs. V_{GB} plot in Fig. 2.17 can be found as follows:

$$\frac{d|Q'_I|}{dV_{GB}} = \frac{d|Q'_I|}{d\psi_s} \frac{d\psi_s}{dV_{GB}} \quad (2.7.22)$$

and, using (2.7.15) and (2.7.21),

$$\frac{d|Q'_I|}{dV_{GB}} = \frac{C'_{ox} C'_i}{C'_{ox} + C'_b + C'_i} \quad (2.7.23)$$

Deep in strong inversion, C'_i becomes dominant and the slope predicted by this equation approaches C'_{ox} , as expected and as indicated in Fig. 2.17. Below strong inversion, C'_i in the preceding equation diminishes, and so does the slope.

Finally, the slope of the $\ln |Q'_I|$ vs. V_{GB} plot in Fig. 2.19 can be found by using (2.7.23):

$$\frac{d \ln |Q'_I|}{dV_{GB}} = \frac{C'_{ox}}{C'_{ox} + C'_b + C'_i} \frac{C'_i}{|Q'_I|} \quad (2.7.24)$$

TABLE 2.1
Regions of inversion and properties

	Weak inversion	Moderate inversion	Strong inversion
Definition in terms of surface potential ψ_s	$\phi_F \leq \psi_s < 2\phi_F$	$2\phi_F \leq \psi_s < 2\phi_F + \phi_{Z0}$	$2\phi_F + \phi_{Z0} \leq \psi_s$
Definition in terms of gate-body voltage V_{GB}	$V_{L0} \leq V_{GB} < V_{M0}$	$V_{M0} \leq V_{GB} < V_{H0}$	$V_{H0} \leq V_{GB}$
$ Q'_I $	$\ll 1$	Varies	$\gg 1$ deep in strong inversion; not necessarily so near the bottom of the region
$\frac{C'_i}{C'_b}$	$\ll 1$ deep in weak inversion; not necessarily so near the top of the region	Varies	$\gg 1$
$\frac{d\psi_s}{dV_{GB}}$	Approximately constant; attains its maximum value in this region	Varies	Small
Dependence of Q'_I on V_{GB}	Approximately exponential	—	Approximately first-degree polynomial
$\frac{d \ln Q'_I }{d \psi_s}$	$\frac{1}{\phi_i}$	Varies	$\frac{1}{2\phi_i}$

Deep in weak inversion C'_i can be neglected in (2.7.21) (Fig. 2.23). The inverse of the quantity $d\psi_s/dV_{GB}$ in this region is practically equal to the quantity denoted by n in Sec. 2.6 [see (2.6.29)]; then, from (2.7.21):

$$n = 1 + \frac{C'_b}{C'_{ox}} \quad (2.7.25)$$

which can be easily shown to reduce to (2.6.30).

It should be noted here that throughout our discussion so far we have assumed that the equivalent interface charge Q'_o is fixed and independent of voltage. This may not be accurate in some devices (especially those fabricated by older techniques) in which a significant density of interface traps may exist (Sec. 2.2). These traps, located at the oxide-silicon interface, can exchange carriers with the silicon. The charge trapped in them depends on the value of the surface potential ψ_s . Thus, let Q'_{it} represent the fraction of Q'_o associated with interface traps. Then, in going from (2.3.4) to (2.3.5), one should include the term $\Delta Q'_{it}$ in the latter. If this term is also included in (2.7.2) and in the development following that equation, it can be easily seen that we can take the variation of Q'_{it} with ψ_s into account in the same way that

we took into account the variation of Q'_B and Q'_I with ψ_s ; that is, we can define an incremental capacitance corresponding to the interface traps in analogy with (2.7.14) and (2.7.15). This capacitance will then be

$$C'_{it} \equiv -\frac{dQ'_{it}}{d\psi_s} \quad (2.7.26)$$

C'_{it} will appear in parallel with C_b and C_i in Fig. 2.24. If it is significant, the various formulas should be modified to include it. For example, (2.7.25) would be modified as follows:¹⁵

$$n = 1 + \frac{C'_b + C'_{it}}{C'_{ox}} \quad (2.7.27)$$

Measurements performed on MOS transistors have shown that, with well-developed fabrication processes, C'_{it} is often much smaller than C'_b and may be neglected.¹⁶

2.8 Summary of Properties of the Regions of Inversion

From the relations given in this chapter, and from the associated discussions, we can summarize certain important properties for each region of inversion as shown in Table 2.1. Most of these properties have already been discussed; the last one is considered in Prob. 2.17.

References

1. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, John Wiley, New York, 1982.
2. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley, New York, 1967.
3. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, New York, 2007.
4. R. F. Pierret, *Field Effect Devices*, 2nd ed., Prentice Hall, New York, 1990.
5. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
6. N. Arora, *MOSFET Models for VLSI Circuit Simulation—Theory and Practice*, World Scientific, Singapore, 2007.
7. K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, N.J., 1993.
8. I. R. M. Mansour, "On the modeling of MOS devices," *Proceedings of the Third International Symposium on Network Theory*, Yugoslavia, 1975, pp. 705–713.
9. I. R. M. Mansour, "Improved modeling of MOS devices," *Proceedings of the European Conference on Circuit Theory and Design*, Italy, 1976.
10. G. Baccarani, M. Rudan, and G. Spadini, "Analytical i.g.f.e.t. model including drift and diffusion currents," *IEEE Journal of Solid-State and Electron Devices*, vol. 2, pp. 62–68, 1978.
11. J. R. Brews, "A charge-sheet model of the MOSFET," *Solid-State Electronics*, vol. 21, pp. 345–355, 1978.
12. F. Van de Wiele, "A long-channel MOSFET model," *Solid-State Electronics*, vol. 22, pp. 991–997, 1979.
13. J. R. Brews, "Physics of the MOS transistor," chap. 1, *Silicon Integrated Circuits, Part A*, D. Kahng (editor), Applied Solid State Science Series, Academic Press, New York, 1981.

14. Y. Tsividis, "Moderate inversion in MOS devices," *Solid-State Electronics*, vol. 25, pp. 1099–1104, 1982; see also Erratum, *Solid State Electronics*, vol. 26, p. 823, 1983.
15. R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-7, pp. 146–153, 1972.
16. E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation," *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 224–231, 1977.
17. R. J. Overstraeten, G. J. Declerk, and P. A. Nuls, "Theory of the MOS transistor in weak inversion—New method to determine the number of surface states," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 282–288, 1975.
18. L. L. Lewyn and J. D. Meindl, "An IGFET inversion charge model for VLSI systems," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 434–440, 1985.
19. M. A. Maher and C. A. Mead, "A physical charge-controlled model for MOS transistors," pp. 211–229, P. Lesleben (editor), *Advanced Research VLSI*, The MIT Press, Cambridge, Mass., 1987.
20. C.-K. Park, C.-Y. Lee, K. Lee, B.-J. Moon, Y. H. Byun, and M. Shur, "A unified current-voltage model for long-channel nMOSFETs," *IEEE Transactions on Electron Devices*, vol. 38, pp. 399–406, 1991.
21. B. Iñiguez and E. G. Moreno, "Explicit C_x -continuous and general model for nMOSFETs," *Electronics Letters*, vol. 29, pp. 1036–1037, 1993.
22. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An explicit physical model for the long-channel MOS transistor including small-signal parameters," *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
23. V. Altschul and Y. Schacham-Diamond, "Modeling of the MOSFET inversion charge and drain current in moderate inversion," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1909–1915, 1990.
24. S. Pavan, Y. Tsividis, and K. Nagaraj, "Modeling of accumulation MOS capacitors for analog design in digital VLSI processes," *Proceedings of the International Symposium on Circuits and Systems*, vol. VI, pp. 202–205, Orlando, 1999.

Problems

- 2.1. Show that, independently of how many different materials are used in the external path in Fig. 2.2c, the value of V_{GB} needed to make the charges disappear is given by (2.2.2).
- 2.2. Calculate the flatband voltage for an *n*-type substrate with $N_D = 5 \times 10^{17} \text{ cm}^{-3}$, an SiO_2 insulator with $t_{\text{ox}} = 2 \text{ nm}$, and a polysilicon gate doped *n*-type with $N_D = 10^{20} \text{ cm}^{-3}$. Assume $Q'_o = 2 \times 10^{-9} \text{ C/cm}^2$.
- 2.3. Produce plots like those in Figs. 2.10, 2.12, and 2.13 for $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $t_{\text{ox}} = 2 \text{ nm}$, and $V_{FB} = -0.8 \text{ V}$.
- 2.4. In Sec. 2.6.2, we have shown that, if ψ_s is assumed to be "pinned" to a constant, the plot of Q'_I vs. V_{GB} will be a straight line. Show that a pinned ψ_s is not a necessary condition for a straight-line plot (although it is sufficient); find the correct necessary condition.
- 2.5. Find an approximation for $\Delta\phi$ in (2.6.17) as follows. Show that (2.6.12) can be written as $\psi_s = 2\phi_F + \phi_t \ln\{[(V_{GB} - V_{FB} - \psi_s)^2/\gamma^2 - \psi_s]/\phi_t\}$. In the argument of the logarithm, replace ψ_s by the approximation $\psi_s = 2\phi_F$ from (2.6.16). This gives an expression in the form of (2.6.17), with $\Delta\phi$ weakly dependent on V_{GB} . Show that, for practical cases, where γ is between 0.2 and 2 $\text{V}^{1/2}$, and $V_{GB} - V_{FB} - 2\phi_F$ is up to several volts, $\Delta\phi$ is several ϕ_t ; an average value is $6\phi_t$. Thus, a better estimate can be obtained if, in the preceding equation, ψ_s in the argument of the logarithm is replaced by $2\phi_F + 6\phi_t$. *Caution:* This expression totally fails in moderate and weak inversion.
- 2.6. For a device with $N_A = 1 \times 10^{18} \text{ cm}^{-3}$, $t_{\text{ox}} = 2.5 \text{ nm}$, and $V_{FB} = -0.752 \text{ V}$, plot $\ln |Q'_I|$ vs. V_{GB} in weak inversion, using (a) (2.6.8) with (2.6.12); (b) (2.6.26) with (2.6.27); and (c) (2.6.34) with (2.6.35). Comment on the accuracy of the last two approaches.

- 2.7. For $N_A = 3 \times 10^{17} \text{ cm}^{-3}$, $t_{\text{ox}} = 6 \text{ nm}$, and $V_{FB} = -0.983 \text{ V}$, plot $Q'_C(V_{GB})$ using (a) (2.6.8) with (2.6.12); (b) (2.6.20), from $V_{GB} = V_{M0} + 0.6 \text{ V}$ to $V_{GB} = 3 \text{ V}$. Comment on the accuracy of (2.6.20). Assume initially $\phi_0 = 2\phi_F + 6\phi_i$ and modify this value if necessary to decrease the error.
- 2.8. Rewrite the equations of this chapter for the case of *n*-type substrates.
- 2.9. Show that in depletion (with ψ_s larger than a few ϕ_i) and in weak inversion C'_{gb} is approximately given by

$$C'_{gb} = \frac{C'_{\text{ox}}}{\sqrt{1 + \frac{4}{\gamma^2}(V_{GB} - V_{FB})}}$$

- 2.10. Determine the value of the surface potential and the gate-body voltage at the onset of moderate inversion (ϕ_{M0}, V_{M0}), as well as the extrapolated threshold (V_{T0}) for the process of Example 2.2. Also, give rough estimates for the onset of strong inversion (ϕ_{H0}, V_{H0}).
- 2.11. Use equations in Appendix A to prove (2.4.22). Note that the result is independent of the thickness of the shaded area containing Q_C in Fig. 2.5.
- 2.12. Equation (2.6.1b) was derived as a special case of the general $Q'_C(\psi_s)$ relation in (2.4.21). Show that (2.6.1b) can also be proved directly,¹³ using the several hints that follow. For $\psi(y)$ larger than a few ϕ_i , the hole charge is negligible, and $\rho(y) = -q[N_A + n(y)]$, with $n(y)$ related to $\psi(y)$ by $n(y) = N_A e^{(\psi - 2\phi_F)/\phi_i}$ in analogy with (2.4.12). Thus Poisson's equation (1.2.29) will be

$$\frac{d^2\psi}{dy^2} = \frac{qN_A}{\epsilon_s} (1 + e^{(\psi - 2\phi_F)/\phi_i})$$

Multiply this equation by $2(d\psi/dy)$ and recognize the resulting left-hand side as $(d/dy)(d\psi/dy)^2$. Integrate both sides from a point y_{bulk} outside the depletion region to a point y_{surface} at the surface; note that in the bulk, $\psi = 0$ and $d\psi/dy = 0$. Solve for $d\psi/dy$ at the surface, and then relate that quantity to Q'_C (see Appendix A). Solve for Q'_C and show that the result is practically the same as (2.6.1b) for ψ_s larger than a few ϕ_i .¹³

- 2.13. Show that, independently of the charge in the inversion layer, if the thickness of this layer is assumed to be zero, the potential across the layer will also be zero (use the material in Appendix A).
- 2.14. Use (A.7) in Appendix A to prove (2.6.6), assuming the “depletion approximation” (Sec. 1.5).
- 2.15. Plots of charge density, electric field, and potential vs. distance from the two-terminal MOS structure are given in Fig. 2.14. Verify these plots by using the basic laws of electrostatics of Appendix A and give expressions for each part of the plots.
- 2.16. (a) Show that, if a fixed charge sheet of charge per unit area Q' is located in the oxide at a distance d from the gate, its contribution to the flatband voltage will be $-(1/C'_{\text{ox}})[(dt_{\text{ox}})Q']$ (use material from Appendix A).
- (b) Show that, if a charge is distributed within the oxide uniformly along the horizontal dimension in Fig. 2.2, as described by a charge density per unit volume $\rho(y)$, its contribution to the flatband voltage will be $-(1/C'_{\text{ox}})\int_{y_g}^{y_s} (y/t_{\text{ox}})qp(y)dy$, where y_g is y at the gate-oxide interface and y_s is y at the oxide-substrate interface.

114 OPERATION AND MODELING OF THE MOS TRANSISTOR

- 2.17. Plot the logarithm of the magnitude of the inversion layer charge, $\ln |Q'_I|$, vs. the surface potential ψ_s and verify that this plot approaches straight-line behavior in the weak and strong inversion regions. Use the process parameters of Prob. 2.6. Determine and compare the slopes of the plot in these two regions. (*Hint:* Show first that $d \ln |Q'_I|/d\psi_s = C'_i/|Q'_I|$.)
- 2.18. Prove (2.7.9). Express the denominator of the first branch in terms of $Q'_C(\psi_s)$, and verify that this branch gives 0/0 for $\psi_s = 0$. To obtain the second branch of the equation, apply l'Hôpital's rule to the first branch (as expressed above) by differentiating both numerator and denominator, leaving the denominator in the form $dQ'_C/d\psi_s$. Relate the denominator to C'_c . Use $\psi_s = 0$ in the result and solve for C'_c . Also, verify (2.7.10).
- 2.19. Prove (2.7.20).
- 2.20. Prove (2.7.24).

3

The Three-Terminal MOS Structure

3.1 Introduction

A complete MOS transistor is formed by adding two more terminals to the basic MOS structure of Chap. 2 to contact two opposite ends of the inversion layer. Through these terminals a potential difference can be applied across the inversion layer and a current can be caused to flow in it. A number of phenomena can then be observed, some directly associated with the current flow, some not. It is not convenient to introduce all these phenomena when they are present simultaneously. Some of these, specifically the ones not directly associated with current flow, can best be isolated and studied by themselves by means of a structure simpler than the MOS transistor, which we call a *three-terminal MOS structure*. This structure is formed by contacting the inversion layer of the basic MOS structure at only one end. We will study the changes that take place in the charges and the potential distribution of the three-terminal structure, caused by the application of an external voltage between this new terminal and the substrate. By the end of this chapter, we will have all the facts necessary for a careful and convenient look at the MOS transistor, which will be the subject of all succeeding chapters.

3.2 Contacting the Inversion Layer

Assume that an n^+ region is added to the basic MOS two-terminal structure, so that the structure shown in Fig. 3.1a is obtained. This region can be contacted through a terminal C. We will now study the concentration of electrons at the surface of the p -type body at points away from the n^+ region. (Points very close to the n^+ region will be affected by the two-dimensional field distribution around it; we will assume for the

present that the gate is long and wide, so that such edge effects can be neglected over practically all of the gate's length and width.) Consider the n^+p junction formed by the n^+ region and the body. As explained in Sec. 1.5, one can short-circuit the junction, as shown in Fig. 3.1b, without altering the potential on the n^+ region. Thus the structure to the right of that region does not sense any difference when C is shorted to B; *for this connection, the part of the structure away from the n^+ region is still governed by the basic equations we have developed for the two-terminal structure.*

We will now study the three-terminal structure in several different situations, focusing on how the potential, and the corresponding potential energy, vary *horizontally at the surface*, as we go from the n^+ region to the p region. Recall that electron potential energy changes in the opposite direction from electrostatic potential, due to the negative charge of electrons; and that electron potential energy is represented by E_c , the bottom of the conduction energy band (Sec. 1.2). The energy level E_v (top of valence energy band) and the intrinsic energy level E_i follow the variation of E_c as has been discussed in previous chapters (see, for example, Fig. 1.22); in the figures we are about to show, only E_c will be shown, to avoid clutter.

Assume that initially V_{GB} is equal to the flatband voltage, V_{FB} , so that the p body is neutral away from the n^+ region, as shown in Fig. 3.2a. The depletion region on the p side contains ionized acceptor atoms as shown. The narrow depletion region part in the n^+ material containing ionized donor atoms (Fig. 1.21) is not shown for simplicity. The n^+p junction is at zero bias. The n^+ region is at a potential ϕ_{bi} with respect to the p region deep in the body, where ϕ_{bi} is the built-in potential of the

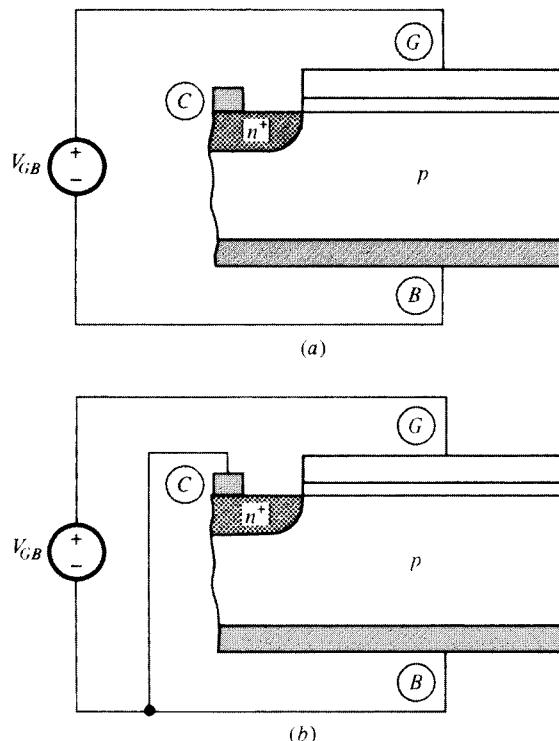
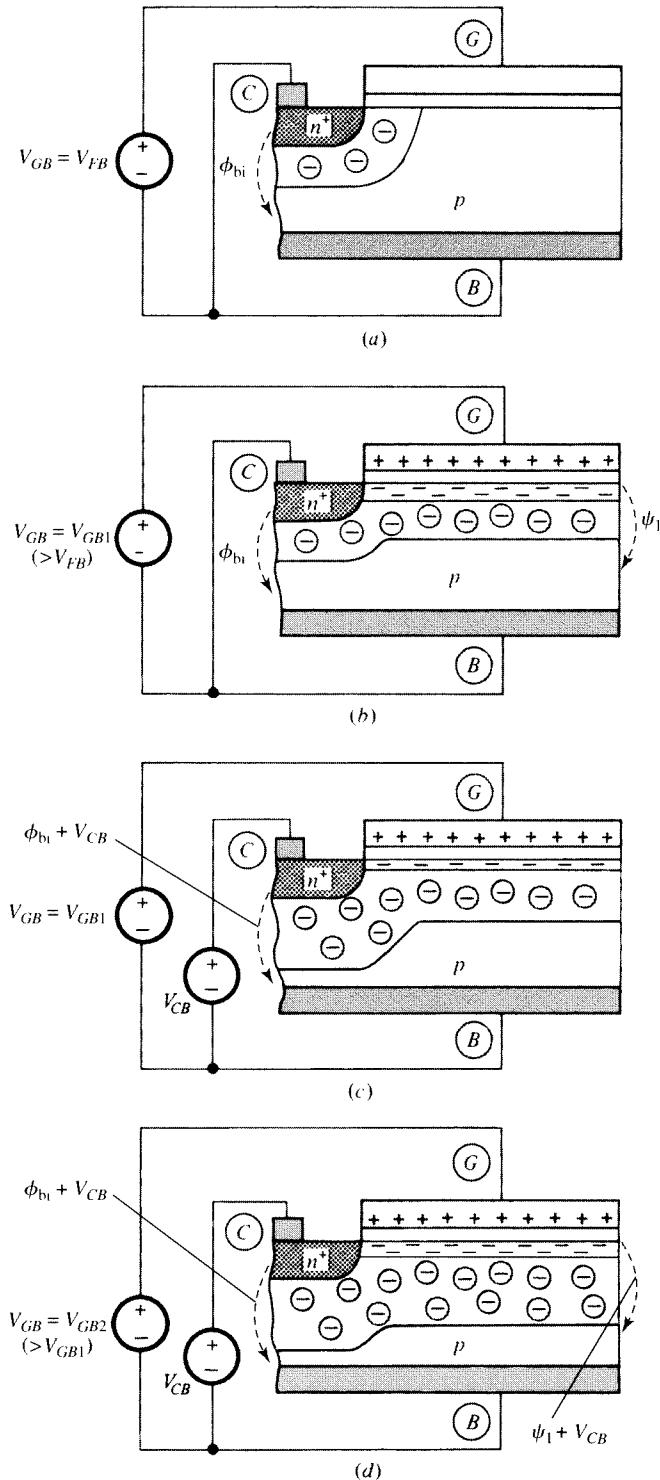


FIGURE 3.1
Three-terminal MOS structure, with n^+ region (a) left open; (b) short-circuited to p -type body.

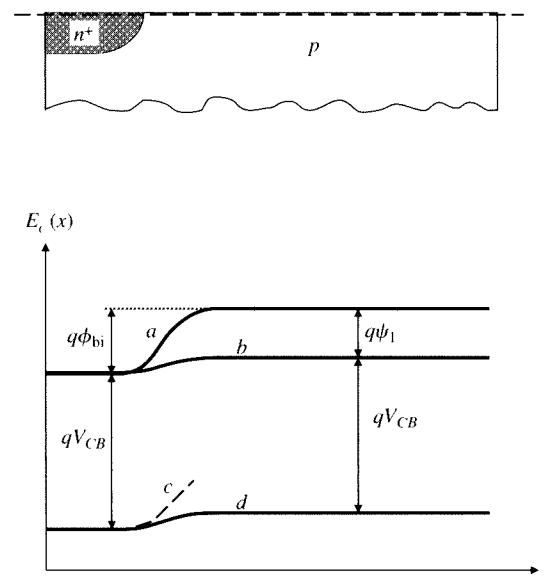
**FIGURE 3.2**

(a) Three-terminal MOS structure in flat-band condition, with $V_{CB} = 0$. (b) The structure of (a) with a larger V_{GB} , causing surface inversion. (c) The structure of (b) with a positive V_{CB} , which reduces electron concentration at the surface. (d) The structure of (c) with further increased V_{GB} , which restores electron concentration at the surface to the level in (b).

junction (Sec. 1.5). The potential at the surface of the p region with respect to deep in the body is zero; thus as we go from the n^+ region, along the surface, to the p region, the total potential drop we see is ϕ_{bi} . This means that the electron potential energy rises as we go from n^+ horizontally to p , by the amount $q\phi_{bi}$, as it did in Fig. 1.22. The plot of E_c at the surface vs. horizontal position x for this case is shown by line *a* in Fig. 3.3. We say that a potential energy *barrier* exists from the n^+ region to the p region. The term “barrier” suggests that it is difficult for electrons to cross. In the n^+ region, electrons have statistically distributed energies; higher energy levels are less probable than lower ones, and in fact it can be shown that the number of electrons with energies above a given level decreases exponentially with that level;¹ the density of electrons that have energies exceeding that of the energy barrier is proportional to $\exp(-\Delta E_c/kT)$, where ΔE_c is the magnitude of this barrier. Due to the large magnitude of this barrier for curve *a*, practically no electrons cross the barrier.

We can lower the barrier by making the potential on the p side (i.e., the surface potential, ψ_1), more positive, as has been shown in Fig. 1.23. As we know from Chap. 2, this can be achieved by increasing the gate-body voltage, V_{GB} , to some value above V_{FB} , denoted by V_{GB1} in Fig. 3.2*b*. This causes the surface potential ψ_1 to attain a positive value, denoted by ψ_1 in the figure. The new potential across the source junction, from the n to the p side, becomes $\phi_{bi} - \psi_1$, which means that the energy barrier is now $q(\phi_{bi} - \psi_1)$. Thus the energy barrier has been lowered, as shown by curve *b* in Fig. 3.3. A significant number of electrons now have energies sufficient to cross the reduced barrier, and they enter the channel region by diffusion. Their density will be proportional to $\exp(-\Delta E_c/kT) = \exp[-q(\phi_{bi} - \psi_1)/kT] = \exp[(\psi_1 - \phi_{bi})/\phi_t]$, i.e., it will be proportional to $\exp(\psi_1/\phi_t)$, in agreement with (2.4.10) at the surface.

Let us now place a voltage source of value V_{CB} between the n^+ region and the substrate terminal, as shown in Fig. 3.1*c*. The value of V_{CB} will be assumed *nonnegative* to ensure that the n^+p junction is *not* forward-biased. This will correspond to

**FIGURE 3.3**

Plots show conduction energy band edge vs. horizontal position along a line adjacent to the surface (broken line at the top of the figure). Plots *a*–*d* correspond to the situations in Figs. 3.2*a*–*d*.

the practical cases of interest when the three-terminal structure will become part of a MOS transistor in Chap. 4. If V_{CB} is not zero, communication of the inversion layer with the external voltage source (through the n^+ region) can drastically change the situation in comparison with the equilibrium case discussed in Chap. 2, which was valid in the previous discussion. The structure is now in *nonequilibrium*, and extra care should be exercised when attempting to apply relations from that chapter. As we will see, most such relations will have to be modified before they can apply to the present case. With V_{CB} applied, the total potential vertically across the n^+p junction will increase from ϕ_{bi} to $\phi_{bi} + V_{CB}$, as expected from Fig. 1.24 (with the role of the reverse bias being played here by V_{CB}). The electron potential energy on the n^+ region will thus be lowered by qV_{CB} ; this causes once again a large energy barrier to exist against the movement of electrons from the n^+ region to the p region, as suggested by broken line *c* in Fig. 3.3 (we will calculate the details of this behavior later on in this chapter). Thus the electron density at the surface of the body will decrease; in fact, if V_{CB} is large enough, the inversion layer can disappear all together. To restore the barrier height to its previous value, we must increase the surface potential by V_{CB} also, which can be achieved by increasing the gate-body voltage to a new value V_{GB2} , as shown in Fig. 3.2d. We now have a barrier height of $q(\phi_{bi} + V_{CB}) - q(\psi_s + V_{CB}) = q\phi_{bi} - q\psi_s$, as shown by line *d* in Fig. 3.3, which is the same as for line *b*. The electron density at the surface of the body is thus restored to the value it had in Fig. 3.2b.

We thus see that what determines the “attractiveness” of the surface for the electrons is not how large the surface potential ψ_s is, but rather how large ψ_s is in comparison to V_{CB} . What counts is the *difference* $\psi_s - V_{CB}$. As long as that difference is fixed, the electron concentration at the surface is also fixed. The role of ψ_s in (2.4.10) is then played here by $\psi_s - V_{CB}$, and we have^{1,2}

$$n_{\text{surface}} = n_o e^{(\psi_s - V_{CB})/\phi_i} \quad (3.2.1)$$

A proof of this result using energy band concepts will be given shortly. Proceeding in the same way as after (2.4.10), we obtain

$$n_{\text{surface}} = n_i e^{(\psi_s - V_{CB} - \phi_F)/\phi_i} \quad (3.2.2a)$$

$$= p_o e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_i} \quad (3.2.2b)$$

and, using $p_o \approx N_A$ from (1.2.6) in the last equation, we obtain

$$n_{\text{surface}} \approx N_A e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_i} \quad (3.2.3)$$

Comparing this last equation to (2.4.12), we see that, just as the two-terminal structure ψ_s was “fighting against” $2\phi_F$ to increase the electron density at the surface, here ψ_s is instead fighting against $2\phi_F + V_{CB}$. This observation will make it easy to go from results for the two-terminal structure to results for the three-terminal structure. For example, as depicted in Fig. 2.8, for the two-terminal structure, the surface electron density becomes equal to the hole concentration in the neutral bulk when $\psi_s = 2\phi_F$; in comparison, for the three-terminal structure, this happens when $\psi_s = 2\phi_F + V_{CB}$, as seen from (3.2.2b). Similarly, for the two-terminal structure, the surface electron density becomes equal to the intrinsic concentration when the surface

potential is equal to ϕ_F ; instead, for the three-terminal structure, this happens when the surface potential is equal to $\phi_F + V_{CB}$, as seen from (3.2.2a).[†]

Proof Using Energy Bands We now consider the energy bands in the body in more detail, and prove (3.2.1) with the help of Fig. 3.4a; we have rotated the structure, so that we can conveniently draw the bands in their common orientation. First, let us recall what happens in an n^+p junction (Sec. 1.5), by going from n^+ to p along the *lower* broken line in Fig. 3.4a. When $V_{CB} = 0$, we have the situation shown in Fig. 1.22. Assume further that the gate-body voltage is such that we are in inversion, with the surface potential value being ψ_1 . Then, going along the *upper* broken line in Fig. 3.4a, we have the situation shown in Fig. 3.4b, which is basically the same as Fig. 2.7d. The total band bending from the bulk to the surface is $q\psi_1$ as shown. Note that the band diagram on the p side for this case (Fig. 3.4b) and the one for the corresponding n^+p junction (Fig. 1.22) are rather similar. This is not surprising, given that the inversion layer is in contact with the n^+ region. In fact, if the surface is strongly inverted, it is sometimes helpful to think of the inversion layer as an extension of the n^+ region. The electron density at the surface is in general given, from (1.2.20b), by

$$n_{\text{surface}} = n_i e^{(E_{Fn} - E_{i,\text{surface}})/kT} \quad (3.2.4)$$

where E_{Fn} is the electron quasi-Fermi level, which, for the zero- V_{CB} case, is the same as the Fermi level, and E_F is a quantity characterizing the entire structure when it is in equilibrium.

Assume now that $V_{CB} > 0$. The structure is no longer in equilibrium. Now, along the *lower* broken line in Fig. 3.4a, we have the situation shown in Fig. 1.24 (with $V_R = V_{CB}$). E_{Fn} in the n^+ region is lowered in comparison with E_{Fp} on the p side, by an amount qV_{CB} . Because the n^+ region is in contact with the inversion layer, this change is communicated to the inversion layer electrons, and E_{Fn} is lowered there as well by the same amount. Thus, as dictated by (3.2.4), their surface density will decrease. If we want to restore this density to its previous level, we will have to raise V_{GB} in order to decrease E_i at the surface (with respect to its position deep in the p side) by as much as E_{Fn} has decreased, i.e., by qV_{CB} . This means that, at the surface, we will need an extra band bending of qV_{CB} . Then, along the upper broken line, we will have the situation shown in Fig. 3.4c. If $E_{Fn} - E_i$ at the surface for Figs. 3.4b and c is the same, n_{surface} will also be the same.

The electron concentration deep in the neutral bulk is given by the equilibrium concentration n_o , which, according to (1.2.8b), is $n_o = n_i e^{(E_{Fn} - E_{i,\text{bulk}})/kT}$. From this equation and (3.2.4), we get $n_{\text{surface}} = n_o e^{(E_{i,\text{bulk}} - E_{i,\text{surface}})/kT} e^{(E_{Fn} - E_{Fp})/kT}$. The numerator in the first exponent of this can be seen to correspond to the total band bending in Fig. 3.4c, which is $q\psi_s$ as shown; the numerator in the second exponent corresponds to

[†] However, in contrast to the case of the two-terminal MOS structure, at this point the *hole* surface concentration is *not* equal to n_p , but is smaller than this quantity (assuming $V_{CB} > 0$). This can be deduced from the general analysis that follows.

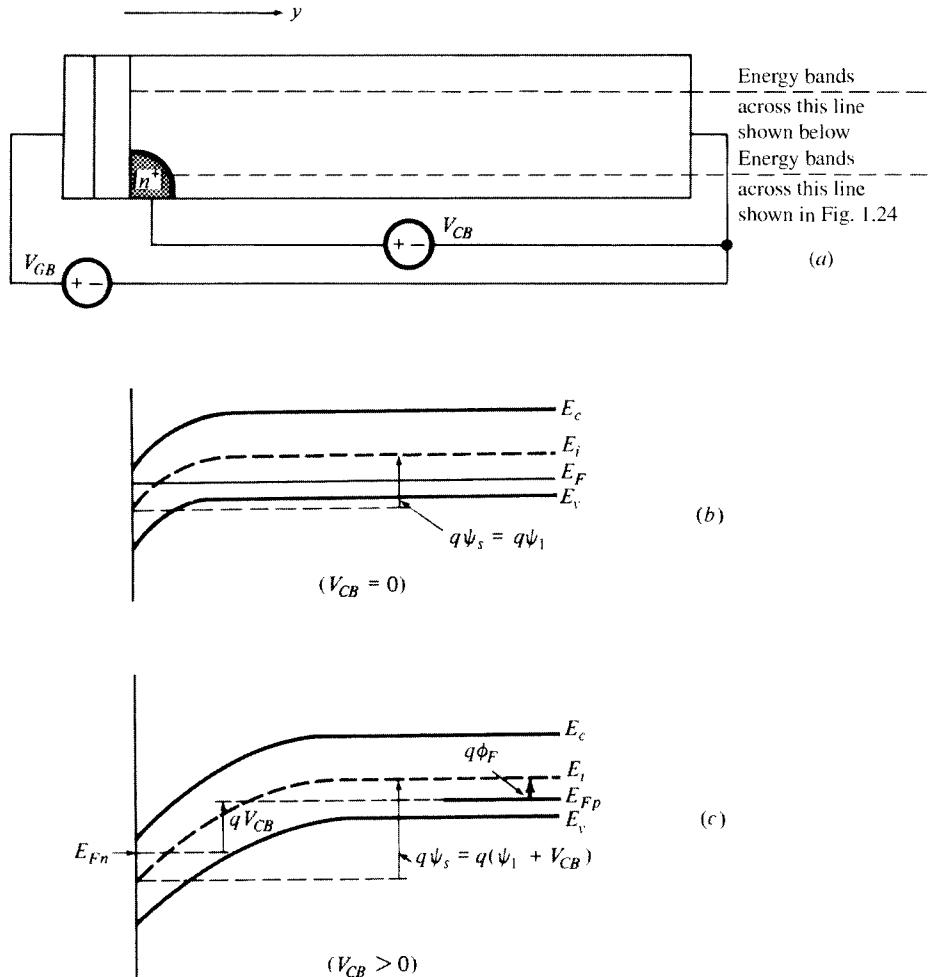


FIGURE 3.4
(a) Three-terminal MOS structure; energy band diagram along *upper* broken line in (a), for (b) $V_{CB} = 0$, and (c) $V_{CB} > 0$.

the quasi-Fermi level split; as we have already discussed and as marked on the figure: this is $-qV_{CB}$. Using these facts and $kT/q = \phi_f$ from (1.2.12), we obtain again (3.2.1).

It should be noted that, with $V_{CB} > 0$, a small current will flow through the battery. This current is due partly to the junction reverse-bias current flowing from the n^+ region to the substrate (Sec. 1.5) and partly to a similar current flowing from the inversion layer to the substrate. The magnitude of the current is very small except at very high temperatures, and we will neglect both its electron and hole components in much of this book. It is important, though, to acknowledge its presence, as this

current is a manifestation of the communication between the inversion layer and the external battery, and the resulting nonequilibrium condition.

The arguments leading to (3.2.1) cannot be extended to holes, as will be explained below. Thus the hole concentration will still be given by the same relations used for the two-terminal structure in Sec. 2.4. A general analysis including the effect of holes and electrons, and valid in accumulation, depletion, and inversion, is now undertaken. Because this analysis parallels the one for the two-terminal structure in Sec. 2.4, we will be rather brief.

General Analysis The derivation of the expression relating the total semiconductor charge to the surface potential can proceed as in Sec. 2.4.4. Following the reasoning given earlier (Sec. 1.5), we assume that the quasi-Fermi level is constant as one goes away from the surface and into the bulk. Thus, the previous arguments remain valid for a point at distance y from the surface, where the potential is $\psi(y)$. In analogy with (3.2.1), we will have

$$n(y) = n_o e^{[\psi(y) - V_{CB}]/\phi_t} = n_o e^{-V_{CB}/\phi_t} e^{\psi(y)/\phi_t} \quad (3.2.5a)$$

Unlike in the case of electrons discussed earlier, the hole quasi-Fermi level, E_{Fp} , remains at its bulk value throughout the semiconductor. This is because the holes do not “communicate” with the external source V_{CB} in the sense previously discussed for electrons; the n^+ region does not attract holes to it, as it does electrons. Thus, we have, just as for the two-terminal structure in (2.4.14),

$$p(y) = p_o e^{-\psi(y)/\phi_t} \quad (3.2.5b)$$

Using these equations in the charge density equation (2.4.15), Poisson's equation (1.2.29) becomes

$$\frac{d^2\psi}{dy^2} = -\frac{q}{\epsilon_s} (p_o e^{-\psi(y)/\phi_t} - N_A + N_D - n_o e^{-V_{CB}/\phi_t} e^{\psi(y)/\phi_t}) \quad (3.2.6)$$

Compare this equation with (2.4.16). It is identical to it, except that n_o has been replaced by $n_o e^{-V_{CB}/\phi_t}$, where V_{CB} is a constant independent of $\psi(y)$. Thus, proceeding exactly as for the two-terminal structure, we obtain general equations that have the same form as in that section, except that n_o is replaced by $n_o e^{-V_{CB}/\phi_t}$ (this is discussed in more detail in Appendix E, where some fine points are taken up). Thus, for example, for the total semiconductor charge we get, instead of (2.4.17):

$Q'_C = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} \sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (N_A - N_D) \psi_s + n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi_s/\phi_t} - 1)}$
Hole contribution Dopant ion contribution Electron contribution

(3.2.7)

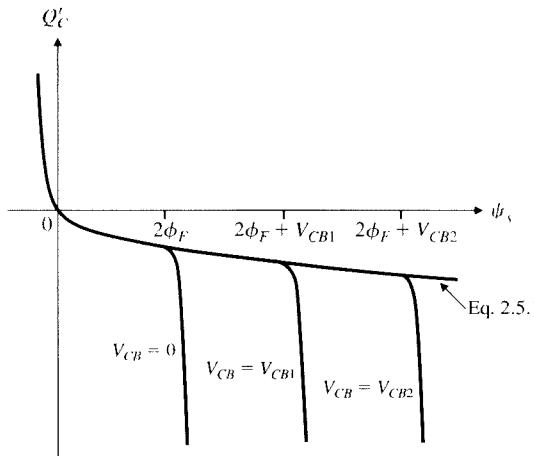


FIGURE 3.5
Total semiconductor charge vs. surface potential for three different values of V_{CB} , with $0 < V_{CB1} < V_{CB2}$. The accumulation-depletion equation (2.5.1) is shown for comparison.

and for the case of a *p*-type substrate we get, instead of (2.4.21):

$$Q'_C = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s N_A} \sqrt{\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-(2\phi_F + V_{CB})/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t)} \quad (3.2.8)$$

Figure 3.5 shows the plot of Q'_C vs. ψ_s , for three values of V_{CB} . For $V_{CB} = 0$, the plot is the same as that in Fig. 2.10 (although a different scale is used). It “takes off” downward when the inversion layer charge becomes significant, which happens at about $\psi_s = 2\phi_F$, as expected. For a positive V_{CB} value V_{CB1} , the plot is qualitatively the same, but its “taking off” is postponed until about $= 2\phi_F + V_{CB1}$, as expected from the preceding discussion. Similarly, for an even larger V_{CB} value V_{CB2} , the plot takes off at about $\psi_s = 2\phi_F + V_{CB2}$. This type of behavior also holds for other quantities; we will see several examples in the next sections. The role of (2.5.1), marked on the figure, will be discussed shortly.

We now proceed in analogy to our treatment of the two-terminal structure. We have four equations that completely characterize the three-terminal MOS structure under our assumptions. These equations are:

- *Potential balance*, (2.3.1): $V_{GB} = \psi_{ox} + \psi_s + \phi_{MS}$
- *Charge balance*, (2.3.4): $Q'_G + Q'_o + Q'_C = 0$
- *The bulk charge-potential relation* (3.2.8), which is of the form $Q'_C = Q'_C(\psi_s, V_{CB})$
- *The gate charge-potential relation* (2.4.22), $Q'_G = C'_{ox}\psi_{ox}$

If the external voltages (V_{GB} and V_{CB}) and the fabrication process parameters are known, the preceding system can in principle be solved to determine ψ_{ox} , ψ_s , Q'_G , and Q'_C . At this point, we should make an important observation. As seen, V_{CB} enters this system of equations *only* through Q'_C . Thus, V_{CB} can affect the results *only* if it can affect Q'_C . However, as we have seen, the only way that V_{CB} can affect Q'_C is through the electron density. In accumulation and depletion, where this density is negligible, this control through V_{CB} is lost, and thus the behavior of the structure becomes the same as that of the two-terminal structure in Chap. 2. This is why for small values of the surface potential, the plots in Fig. 3.5 merge into (2.5.1).

Proceeding as in Sec. 2.4, we find that the structure, due to the requirements of charge balance, potential balance, and the gate-charge relation, must satisfy (2.4.23):

$$Q'_C = -C'_{\text{ox}}(V_{GB} - V_{FB} - \psi_s) \quad (3.2.9)$$

or

$$V_{GB} = V_{FB} + \psi_s - \frac{Q'_C(\psi_s)}{C'_{\text{ox}}} \quad (3.2.10)$$

where V_{FB} is given by (2.2.6).

For the relation between the surface potential and the externally applied gate-body voltage, we use (3.2.8) in (3.2.10) to obtain

$$V_{GB} = V_{FB} + \psi_s + \text{sgn}(\psi_s) \gamma \sqrt{\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-(2\phi_F + V_{CB})/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t)}$$

(3.2.11)

which corresponds to (2.4.25). This equation is one of the foundations of charge sheet models, but cannot be explicitly solved for ψ_s , given V_{GB} and V_{CB} . Highly efficient numerical solutions have been proposed, and extremely accurate explicit approximate solutions have been developed, as will be discussed in Chap. 4.

Refer now to Fig. 3.2c, and assume that V_{CB} is held constant. We can then define a per-unit-area small-signal capacitance of the gate to the rest of the structure, $C'_g = dQ'_G/dV_{GB}$; reasoning as in Sec. 2.7, we will have

$$\frac{1}{C'_g} = \frac{1}{C'_{\text{ox}}} + \frac{1}{C'_c} \quad (3.2.12)$$

where the per-unit-area capacitance associated with the total semiconductor charge is defined as in (2.7.7). For this capacitance we have, instead of (2.7.9),

$$\begin{aligned} C'_c &= \text{sgn}(\psi_s) \sqrt{2q\epsilon_s N_A} \frac{1 - e^{-\psi_s/\phi_t} + e^{-(2\phi_F + V_{CB})/\phi_t} (e^{\psi_s/\phi_t} - 1)}{2\sqrt{\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-(2\phi_F + V_{CB})/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t)}}, \quad \psi_s \neq 0 \\ &= \sqrt{q\epsilon_s N_A (1 + e^{-2\phi_F/\phi_t})/\phi_t}, \quad \psi_s = 0 \end{aligned} \quad (3.2.13)$$

These results are valid in all regions, from accumulation to inversion. As in Chap. 2, we now derive simplified expressions valid in particular regions.

Accumulation and Depletion As already mentioned, all differences discussed between the three- and two-terminal cases involve the population of electrons, which is drastically affected by V_{CB} . If the electron density is negligible, such differences are irrelevant and the corresponding equations derived in Chap. 2 are valid for the three-terminal structure as well. This is precisely the case in accumulation and in depletion. All results derived for those regions in Sec. 2.5 are valid for the three-terminal structure and will not be repeated here. If desired, those results can be derived in a straightforward manner directly from the equations given in the previous section, by neglecting appropriate terms in these regions (Prob. 3.15). An example can be seen

in Fig. 3.5: (2.5.1) remains valid as long as we are in accumulation or depletion for a given V_{CB} .

Inversion The development of equations in inversion parallels that for the two-terminal structure in Sec. 2.6.1. As expected from the previous discussion, all we need to do is replace n_o by $n_o e^{-V_{CB}/\phi_i}$, or $2\phi_F$ by $2\phi_F + V_{CB}$. Alternatively, we can obtain the required equations simply by ignoring negligible terms in the general equations above. Either way we obtain, for the total semiconductor charge,

$$Q'_C = -\sqrt{2q\epsilon_s} \sqrt{N_A \psi_s + n_o \phi_i e^{(\psi_i - V_{CB})/\phi_i}} \quad (3.2.14)$$

or

$$Q'_C = -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_s + \phi_i e^{[\psi_i - (2\phi_F + V_{CB})]/\phi_i}} \quad (3.2.15)$$

This charge consists of inversion layer and depletion region components:

$$Q'_C = Q'_I + Q'_B \quad (3.2.16)$$

Under the charge sheet approximation for the inversion layer, we have a pure depletion region for which (2.6.7) applies:

$$Q'_B = -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_s} \quad (3.2.17a)$$

$$= -\gamma C'_{ox} \sqrt{\psi_s} \quad (3.2.17b)$$

Thus, from the above three equations we obtain^{†‡}

$$Q'_I = -\sqrt{2q\epsilon_s N_A} \left(\sqrt{\psi_s + \phi_i e^{[\psi_i - (2\phi_F + V_{CB})]/\phi_i}} - \sqrt{\psi_s} \right) \quad (3.2.18)$$

For future use, we also develop an alternative expression for Q'_I ; from (3.2.16) and (3.2.9), we obtain

$$Q'_I = -C'_{ox} (V_{GB} - V_{FB} - \psi_s) - Q'_B \quad (3.2.19)$$

which, using (3.2.17b), gives

$$Q'_I = -C'_{ox} (V_{GB} - V_{FB} - \psi_s - \gamma \sqrt{\psi_s}) \quad (3.2.20)$$

[†] We emphasize that (3.2.17) and (3.2.18) are valid only in inversion and in the upper part of depletion. To ensure accuracy for the total semiconductor charge in accumulation and in the accumulation-depletion transition region, in CAD models Q'_B is replaced by the right-hand side of (2.5.1), which represents acceptor charge in depletion and inversion, hole charge in accumulation, and a combination in between. One can then express Q'_I by subtracting (2.5.1) from (3.2.8). The values of Q'_I predicted in this way in the inversion region are practically identical to those from (3.2.18). In other regions, the accuracy of this approach is immaterial, as Q'_I is totally negligible there anyway. This approach guarantees proper behavior for the bulk charge in all regions and continuity of this quantity as the bounds between different regions are crossed.

[‡] In moderate and weak inversion the two square roots of (3.2.18) are of similar magnitude, and this can be a cause of numerical problems. A trick that can be used to avoid such problems has been mentioned in the footnote to (2.6.8).

126 OPERATION AND MODELING OF THE MOS TRANSISTOR

Finally, we can develop a relation between the externally applied gate-body voltage and the surface potential by using (3.2.15) in (3.2.10):

$$V_{GB} = V_{FB} + \psi_s + \gamma \sqrt{\psi_s + \phi_t e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_t}} \quad (3.2.21)$$

where γ is given by (2.4.26). This equation cannot be solved explicitly for ψ_s . If V_{GB} and V_{CB} are given and ψ_s is desired, the equation can be solved numerically. We will revisit the solution of this equation in the context of the complete MOS transistor in Chap. 4.

Equations (3.2.18), (3.2.20), and (3.2.21) are not independent; any one of them can be obtained from the other two. So we will be using two of these equations at a time, but the choice of which two will be determined by convenience in developing specific results.

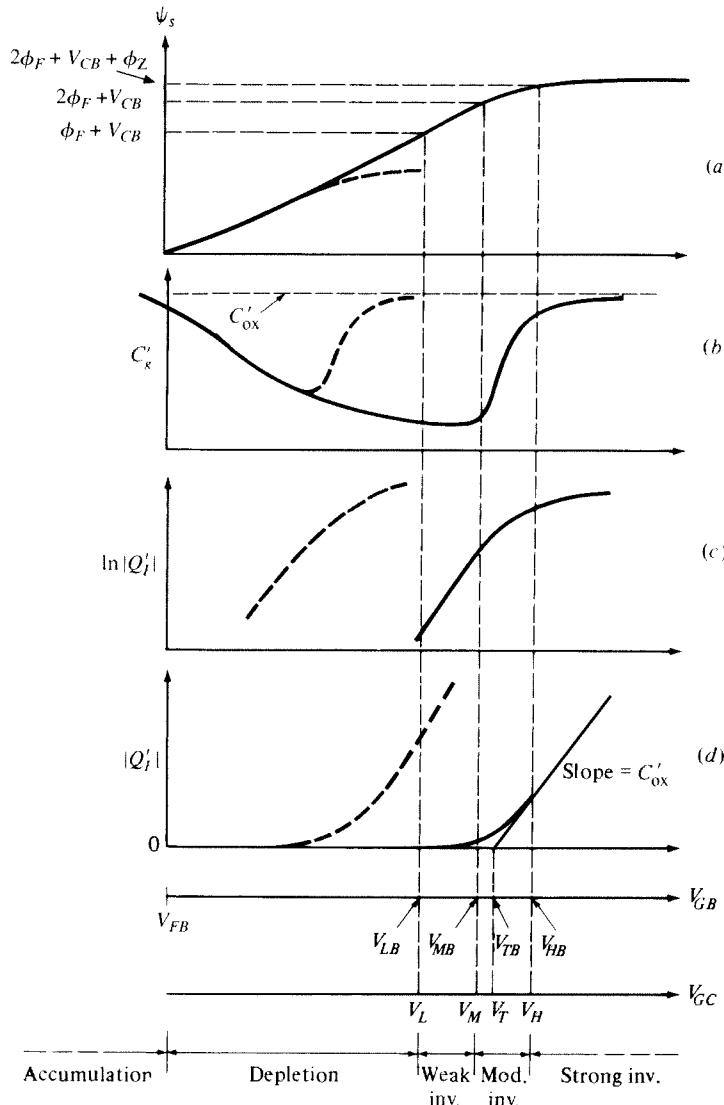
Depletion region and inversion layer capacitances can be defined as in (2.7.14) and (2.7.15). Their expressions can be obtained from the corresponding ones in Sec. 2.7, with $2\phi_F$ replaced by $2\phi_F + V_{CB}$ (Appendix E):

$$C'_b = \sqrt{2q\epsilon_s N_A} \frac{1}{2\sqrt{\psi_s + \phi_t e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_t}}} \quad (3.2.22)$$

$$C'_i = \sqrt{2q\epsilon_s N_A} \frac{e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_t}}{2\sqrt{\psi_s + \phi_t e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_t}}} \quad (3.2.23)$$

As in Chap. 2, the expressions we give for C'_b and C'_i are more accurate than those that would be obtained by differentiating our expressions for Q'_B and Q'_I . This is because (3.2.17) and (3.2.18), although adequate for calculating Q'_B and Q'_I , are not accurate enough to provide correct derivatives with respect to ψ_s [see discussion preceding (2.7.16) for details]. It is easy to note from (3.2.22) and (3.2.23) that $C'_i = C'_b$ at the point $\psi_s = 2\phi_F + V_{CB}$.

Plots of ψ_s , C'_g , $\ln|Q'_I|$, and $|Q'_I|$ vs. V_{GB} are shown in Fig. 3.6; the V_{GC} axis at the bottom will be considered later. The plots shown by the broken lines are for the case of Fig. 3.2b. These plots are the same as for a corresponding two-terminal structure, with one important exception for the C'_g plot. Let us compare this plot to Fig. 2.25. In the latter, the broken line represents C'_{gb} for the two-terminal structure at high frequencies of operation. This behavior was claimed to be different from that observed at low frequencies (solid line). The reason was traced to the fact that in the two-terminal structure the inversion layer is practically “isolated from the outside world.” However, in the structure of Fig. 3.2b or Fig. 3.2c, communication with the outside world is possible through the n^+ region. Thus, in strong inversion plenty of electrons are available just below the oxide, and their total charge “tracks” variations in V_{GB} even if these variations are of rather high frequency. Low-frequency and high-frequency behaviors then coincide, and only one curve is observed, as shown by the broken line in Fig. 3.6b. However, if an *extremely* high frequency is used, the supply of charge to the inversion layer might once again be unable to keep up; a behavior similar to that shown by the broken line in Fig. 2.25 could then be observed.

**FIGURE 3.6**

Various quantities characterizing the structure of Fig. 3.2c plotted vs. V_{GB} and V_{GC} . (a) Surface potential; (b) gate capacitance to the rest of the structure per unit area; (c) logarithm of inversion layer charge magnitude per unit area; (d) inversion layer charge magnitude per unit area. For the broken lines $V_{CB} = 0$; the solid lines are for a V_{CB} of a given positive value.

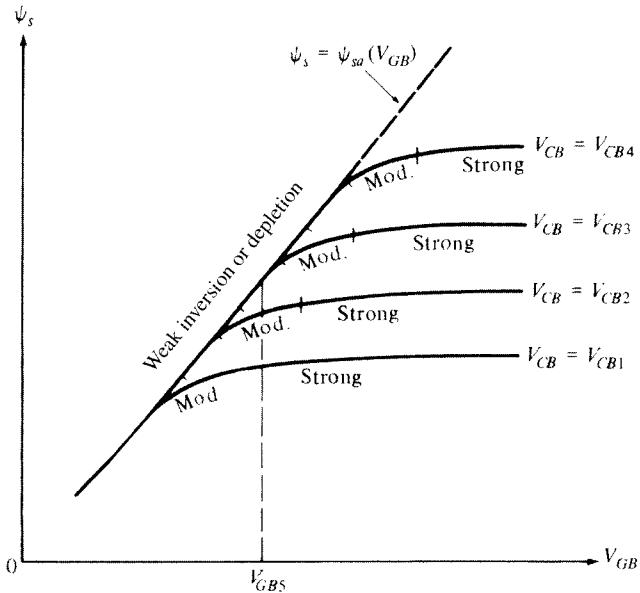
The plots given by the solid lines in Fig. 3.6 are for a given $V_{CB} > 0$. The behavior seen is qualitatively similar to that for $V_{CB} = 0$. Since the latter case is related to the two-terminal structure of Chap. 2, we can carry over several of the concepts and definitions discussed there to the general case of $V_{CB} > 0$. Thus, weak-, moderate-, and strong-inversion regions are again defined, as indicated at the bottom of the figure. The lower limits of these regions in terms of V_{GB} are denoted correspondingly by

V_{LB} , V_{MB} , and V_{HB} as shown. Definitions for each region and values for the preceding quantities will be given later on. For now, it suffices to say that the regions are defined in such a way that in weak inversion the plot of $|Q'_l|$ vs. V_{GB} is essentially exponential; in strong inversion, it is essentially a straight line; and in moderate inversion, it is neither.

Note that the plots in Fig. 3.6 are for a given V_{CB} value. If V_{CB} is increased above that value, the plots will be qualitatively similar but the points where the effects of inversion become apparent will be shifted to the right, as predicted by the corresponding equations. This has already been explained using energy bands, in relation to Fig. 3.3. To provide further intuition, we will now give a similar explanation without using energy bands. Let us assume that in Fig. 3.2c, V_{CB} is fixed at some value. Assume that V_{GB} is so low that not only the inversion layer is absent but also the depletion region width is smaller than shown. Let us now increase V_{GB} . More positive charge will now be placed on the gate, which must be balanced by more negative charge under the oxide. An increase in the depletion region width accomplishes this balance. If V_{GB} is increased far enough, a point will eventually be reached where the depletion region under the oxide will be almost as wide as that under the n^+ region; the value of V_{GB} needed for this to happen depends on the value of V_{CB} . When this happens, the potential across the depletion regions under the oxide and under the n^+ region will be almost the same. Thus, a point at the surface is at almost the same potential (with respect to points deep in the bulk) as is a point in the n^+ region. Now the surface is about as “attractive” for electrons as is the n^+ region. Electrons are attracted most to regions with the most positive potential, and they have now no reason to prefer only the n^+ region as opposed to the “surface” on its right.[†] An inversion layer is thus formed. If now V_{CB} is raised further, the depletion region under the n^+ region will become again wider than under the inversion layer. The electrons will tend again to favor the n^+ region rather than the surface, and the inversion layer will tend to disappear. If it is desired to restore the level of inversion previously achieved, V_{GB} will have to be raised further to make the depletion region width under the gate almost the same as that under the n^+ region once again. Then the potential at the surface will again be almost the same as the potential of the n^+ region (both with respect to the bulk), and electrons can once more be attracted to the surface. From the above picture, the competing roles of V_{CB} and V_{GB} are clear. Increasing V_{CB} tends to make the level of inversion lighter; increasing V_{GB} tends to make that level heavier.

It is interesting to investigate the effect of these phenomena on the surface potential. In Fig. 3.7, we show ψ_s vs. V_{GB} , as it results from (3.2.21), for different values of V_{CB} . The regions of inversion indicated can be interpreted in terms of the corresponding Q'_l plots (which are not shown, but are nevertheless nearly exponential in weak, nearly straight lines in strong, and neither in moderate inversion). Note that

[†] To be more quantitative than this, one must recall that the potential at the surface is ψ_s , whereas that of the n^+ region is $\phi_{bi} + V_{CB}$, where ϕ_{bi} is the built-in potential of the n^+p junction (Sec. 1.5). It can be verified by using (3.2.4) and the material in Sec. 1.5 that, to make the electron concentration at the surface equal to that in the n^+ region, require $\psi_s = \phi_{bi} + V_{CB}$. For other values of ψ_s , the surface concentration will be different, as determined by (3.2.4), and there will be a transition region along the surface around the n^+p junction boundary, over which the potential changes from $\phi_{bi} + V_{CB}$ to ψ_s . As already mentioned, we are focusing on the part of the structure to the right of this transition region. As one goes from weak toward strong inversion, the potential across this horizontal transition region diminishes.

**FIGURE 3.7**

Surface potential vs. gate-substrate voltage with V_{CB} as a parameter. For each V_{CB} the regions of inversion are indicated on the corresponding curve.

increasing V_{CB} “postpones” the tendency of ψ_s to “flatten out” until larger V_{GB} values. This is because it “postpones” the formation of an inversion layer, as already explained. Assume now that, for a given V_{CB} , the value of the gate-body voltage V_{GB} is such that we are in deep depletion. This situation has been covered in Sec. 2.5; the surface potential for this case is denoted by ψ_{sa} and is related to V_{GB} by (2.5.5), repeated here:

$$\boxed{\psi_{sa}(V_{GB}) \approx \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2} \quad (3.2.24)$$

This function is shown by the broken curve in Fig. 3.7. It can be seen that for any V_{CB} , as long as $\psi_s \approx \psi_{sa}(V_{GB})$, the effect of V_{CB} on ψ_s is negligible. Physically, this is because in such a case there is practically no inversion layer charge. Most electric field lines coming from the gate “pass right through” the practically absent inversion layer and terminate on ionized acceptor atoms on the substrate. Thus, as has already been mentioned, Q'_l has “no handle” on the overall balance of the system and, although varying V_{CB} will vary drastically the *negligible* $|Q'_l|$ in a relative sense, the rest of the system simply “does not notice.” Note that $|Q'_l|$ will be negligible as long as the exponential in (3.2.18) is negligible. Things will begin changing significantly when the exponential starts becoming important, which will happen at about $\psi_s \approx 2\phi_F + V_{CB}$. Then the complete (3.2.21) must be used to find ψ_s , and the plot for the latter will begin to flatten out, as seen in Fig. 3.7.

It can be seen from the previous discussion and Fig. 3.7 that $\psi_{sa}(V_{GB})$ is the surface potential that develops, for a given V_{GB} , if the inversion layer is prevented

from forming (by maintaining V_{CB} sufficiently high). As seen in Fig. 3.7, the slope of $\psi_{sa}(V_{GB})$ is a weak function of V_{GB} ; this fact will be seen to lead to certain simplifications later on. Consistent with Sec. 2.6.3, we will denote the *inverse* of this slope by n :

$$n \equiv \left(\frac{d\psi_{sa}}{dV_{GB}} \right)^{-1} \quad (3.2.25)$$

Using (3.2.24) in the preceding definition, we obtain, as in Sec. 2.6.3,

$$n = 1 + \frac{\gamma}{2\sqrt{\psi_{sa}(V_{GB})}} \quad (3.2.26)$$

This quantity is a weak function of V_{GB} . Typical values for n are between 1 and 1.5.

Now let V_{GB} have a given value, say V_{GB5} as shown in Fig. 3.7 (the subscript 5 is used in connection with a discussion later in this chapter). It is clear that one cannot state the level of inversion, knowing only this value of V_{GB} ; rather, the structure can be in strong, moderate, or weak inversion, depending on the value of V_{CB} , as seen in the figure.

Before closing this section, we should note that in some of our discussions we will find it convenient to use the connection shown in Fig. 3.8b rather than the one in Fig. 3.8a. Because

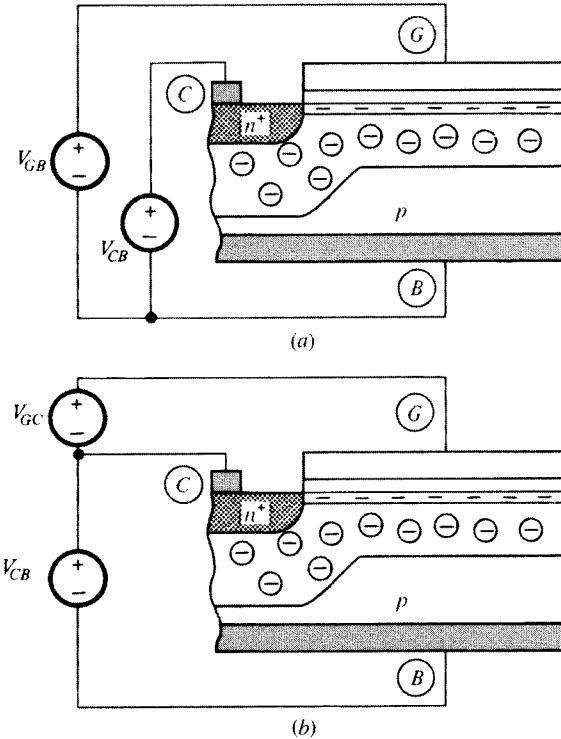


FIGURE 3.8

Three-terminal structure with (a) common B -terminal connection, and (b) common C -terminal connection.

$$\boxed{V_{GB} = V_{GC} + V_{CB}} \quad (3.2.27)$$

the plots of Fig. 3.6 can be viewed as being vs. V_{GC} . This is shown by using a second horizontal axis at the bottom of the figure. The symbols V_L , V_M , and V_H denote the onset of weak, moderate, and strong inversion, respectively, in terms of V_{GC} .

3.3 The Body Effect

From the qualitative discussion in the previous section, it is clear that increasing V_{CB} decreases the level of inversion, unless we also increase V_{GB} by an appropriate amount. It turns out that, if we want to keep the level of inversion the same, the increase in V_{GB} must be *larger* than the increase in V_{CB} . This is referred to as the *body effect* or *substrate effect*. In this section, we will attempt to provide intuition about it. Quantitative results will follow in subsequent sections, where we consider the limits of regions of inversion.

It is best to discuss the effect mentioned with the help of Fig. 3.8b. Body effect really refers to the fact that, if V_{CB} in the connection shown is raised while V_{GC} is kept constant, then the level of inversion will decrease, although $V_{GB} = V_{GC} + V_{CB}$ is increased, and that, if we want to restore the original level of inversion, we will have to *increase* V_{GC} . To see the reason why, let us assume first that V_{GC} is large enough to cause *strong* inversion. The strongly inverted surface, containing an abundance of electrons, is sometimes likened to an n^+p region. In fact, the inversion layer–body combination of this case is sometimes called a *field-induced n^+p junction*. In several respects this junction has a similar behavior as the regular n^+p junction discussed in Sec. 1.5. The role of the reverse bias V_R there is played here by V_{CB} . Thus, V_{CB} will be called the *effective reverse bias* of the field-induced junction in the strong-inversion region. Just like increasing V_R widens the depletion region in a regular junction, increasing V_{CB} will widen the depletion region under the strongly inverted surface. Assume now that we are well into strong inversion, and that V_{CB} in Fig. 3.8b is increased by a small amount. In the connection shown in this figure, this will not affect V_{GC} . The strong-inversion layer, with its abundance of electrons, can be viewed as the bottom plate of a capacitor, with the gate being the top plate. Since V_{GC} has not changed, the voltage across this “capacitor” will remain essentially unchanged, and there will be practically no change in the gate charge.

Since V_{CB} has increased, the depletion region will widen, more acceptor atoms will be uncovered, and the total depletion region charge will become more negative. The *total* negative charge under the oxide, however, must balance the positive gate charge, which has remained practically unchanged, as already explained. Since more ionized acceptor atoms are now available in the depletion region to contribute to this “balancing” function, fewer free electrons are needed in the inversion layer. Thus, the level of inversion will decrease. If it is desired to restore the previous level of inversion, one must then increase V_{GC} by a sufficient amount.

The larger the body doping, the larger the change in the depletion region charge for a given change in V_{CB} [this can be easily deduced by analogy to (1.5.18) and (1.5.3)]. Continuing the preceding argument, this means that a larger increase in the gate voltage will be needed then to restore the original level of inversion. Thus, the

body effect is more pronounced for heavily doped substrates. This effect will also be more pronounced for thicker oxides. The thicker the oxide, the weaker the influence of the gate on the inversion layer charge, and the larger the gate voltage needed to influence that charge to a given degree. The preceding dependence on substrate doping and oxide thickness is manifest in the expression for the *body effect coefficient* γ in (2.4.26a), which will be seen to enter directly in quantitative results for the body effect in Sec. 3.4.1.

These arguments (likening the inversion layer–substrate structure to a n^+p junction, viewing V_{CB} as its “effective reverse bias,” and considering the inversion layer as the bottom plate of a capacitor) are *only valid in strong inversion*. In moderate and, especially, in weak inversion, the electron concentration in the inversion layer is light, and the layer *cannot* be viewed as an n^+ region or as a capacitor plate. A significant portion of the electric field lines from the gate pass through the inversion layer and terminate directly on ionized acceptor atoms in the depletion region. In fact, as already explained in Sec. 3.2, in weak inversion, varying V_{CB} will leave the surface potential practically unaffected (see Fig. 3.7), and so cannot affect appreciably the depletion region width. Concerning, however, the effect of V_{CB} on the *level of inversion*, the same qualitative conclusions as for strong inversion can be reached. Thus, assume that we are in weak inversion and that V_{CB} in Fig. 3.8b is increased. To keep the level of inversion the same, ψ_s must increase by about the same amount, as has already been argued in Sec. 3.2. To increase ψ_s by this amount, V_{GB} must be increased by a *larger* amount, as seen from (3.2.11); thus, $V_{GC} = V_{GB} - V_{CB}$ must be increased.

As a result of the preceding effects, an increase in V_{CB} will result in increases in the values of V_L , V_M , and V_H marked on the V_{GC} axis in Fig. 3.6. The quantity V_T shown there, which is simply the “extrapolated threshold voltage” (in analogy to V_{T0} for the two-terminal structure), is also found to increase (Sec. 3.4.2). In fact, the term *body effect* is commonly taken to mean just the increase of V_T with V_{CB} .

3.4 Regions of Inversion

3.4.1 Approximate Limits

For the two-terminal MOS structure we have defined the onset of weak, moderate, and strong inversion by the surface potential values ϕ_F , $2\phi_F$, and $2\phi_F + \phi_{Z0}$ (see Fig. 2.16). For the three-terminal structure, on the basis of the discussion in Sec. 3.2, we define these onsets as indicated in the top row of Table 3.1. In the right-most entry, ϕ_Z is the width of the moderate inversion region in terms of surface potential; its value is several ϕ_F . The onset of strong inversion is chosen so that in this region $Q_f(V_{GB})$ is essentially a straight line (see Sec. 3.4.2). With $V_{GC} = 0$, the surface potential bounds shown reduce to their corresponding values for the two-terminal MOS structure.

The onsets of weak inversion and of moderate inversion in terms of V_{GB} can be found by using the corresponding surface potentials in (3.2.21) and neglecting the exponential term. Thus, for example, we have

$$V_{MB} = V_{FB} + (2\phi_F + V_{CB}) + \gamma\sqrt{2\phi_F + V_{CB}} \quad (3.4.1)$$

TABLE 3.1
Approximate bounds between regions[†]

	Bound between depletion and weak inversion	Bound between weak and moderate inversion	Bound between moderate and strong inversion
In terms of surface potential ψ_s	$\phi_F + V_{CB}$	$2\phi_F + V_{CB}$	$2\phi_F + V_{CB} + \phi_Z^{\ddagger}$
In terms of V_{GB} , for a given V_{CB}	$V_{LB} = V_L + V_{CB}$	$V_{MB} = V_M + V_{CB}$	$V_{HB} = V_H + V_{CB}$
In terms of V_{GC} , for a given V_{CB}	$V_L = V_{FB} + \phi_F + \gamma\sqrt{\phi_F + V_{CB}}$	$V_M = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{CB}}$	$V_H = V_M + V_Z^{\$}$
In terms of V_{CB} , for a given V_{GB} (see Sec. 3.5)	$V_U = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 - \phi_F$	$V_W = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 - 2\phi_F$	$V_Q = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB} - V_Z} \right)^2 - 2\phi_F$

[†] For more accurate results, see Appendix D.

[‡] ϕ_Z is a weak junction of process parameters, temperature, and V_{CB} . Its value is several ϕ_F . One should *not* attempt to use an approximate value from this expression for ψ_s in expressions like (3.2.21), as a small error in ψ_s can cause a large error in the exponential term.

[§] V_Z is typically 0.5 to 0.6 V at room temperature, for practical combinations of substrate doping and oxide thickness, and for values of V_{CB} up to a few volts. It decreases somewhat with increasing V_{CB} .

The corresponding values in terms of V_{GC} can be found by subtracting V_{CB} , as suggested by (3.2.27). Thus, for example:

$$V_M = V_{MB} - V_{CB} \quad (3.4.2a)$$

$$V_M = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{CB}} \quad (3.4.2b)$$

These results are shown in the second and third rows of Table 3.1 along with expressions for the other limit points. An ambiguity is involved in defining the onset of strong inversion, for the same reasons as those discussed in Sec. 2.6.2. Thus, approximate results are shown in Table 3.1. For more accurate results,³ the reader is referred to Appendix D. Discussion of the last row of Table 3.1 will be postponed until Sec. 3.5.

Plots of V_L , V_M , and V_H vs. V_{CB} for a given fabrication process look as shown in Fig. 3.9 (also shown is the plot for V_T , for which an expression will be derived in Sec. 3.4.2). The values of these quantities at $V_{CB} = 0$ are denoted by adding the subscript zero as shown in the figure. These values are the same as for the two-terminal MOS structure, for reasons already discussed. The fact that the quantities in Fig. 3.9 increase with increasing V_{CB} is a consequence of the body effect discussed in Sec. 3.3. As is apparent from the expressions in Table 3.1, how much V_L , V_M , and V_H will increase for a given increase in V_{CB} is determined by the value of the coefficient γ ; hence, the name *body effect coefficient* for this quantity. We remind the reader that γ was defined in (2.4.26), which is repeated below:

$$\boxed{\gamma = \frac{\sqrt{2q\epsilon_s N_A}}{C'_\text{ox}}} \quad (3.4.3)$$

This quantity has been plotted in Fig. 2.11. It is seen that our intuitive predictions of Sec. 3.3 (that the body effect is stronger for heavier substrate dopings and/or thicker oxides) are verified by the entries for V_L , V_M , and V_H in Table 3.1, in conjunction with (3.4.3).

The definitions and properties of the regions of inversion in terms of ψ_s , V_{GB} , V_{GC} , and V_{CB} are summarized in Table 3.2. (The fourth row of the table will be

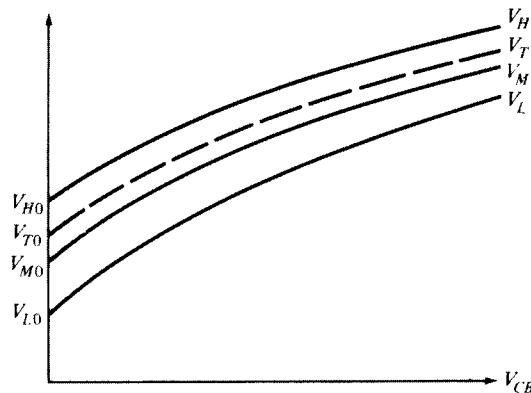


FIGURE 3.9

Onset of strong inversion V_H , extrapolated threshold V_T , onset of moderate inversion V_M , and onset of weak inversion V_L vs. V_{CB} for the three-terminal MOS structure of Fig. 3.8b.

TABLE 3.2
Regions of inversion and properties (three-terminal MOS structure)

	Weak inversion	Moderate inversion	Strong inversion
Definition in terms of surface potential ψ_s (see Fig. 3.6a)	$\phi_F + V_{CB} \leq \psi_s < 2\phi_F + V_{CB}$	$2\phi_F + V_{CB} \leq \psi_s < 2\phi_F + V_{CB} + \phi_Z$	$2\phi_F + V_{CB} + \phi_Z \leq \psi_s$
Definition in terms of $V_{LB} \leq V_{GB} < V_{MB}$ for a given V_{CB} (see Figs. 3.8b and 3.6)		$V_{MB} \leq V_{GB} < V_{HB}$	$V_{HB} \leq V_{GB}$
Definition in terms of $V_L \leq V_{GC} < V_M$ for a given V_{CB} (see Figs. 3.8b and 3.6)	$V_L \leq V_{GC} < V_M$	$V_M \leq V_{GC} < V_H$	$V_H \leq V_{GC}$
Definition in terms of $V_U \geq V_{CB} > V_W$ for a given V_{GB} (see Fig. 3.1c and Sec. 3.5) [†]	$V_U \geq V_{CB} > V_W$	$V_W \geq V_{CB} > V_Q$	$V_Q \geq V_{CB}$
$\frac{ Q'_I }{ Q'_B }$	$\ll 1$	Varies	$\gg 1$ deep in strong inversion; not necessarily so near the bottom of the region
$\frac{C'_I}{C'_B}$	$\ll 1$ deep in weak inversion; not necessarily so near the top of the region	Varies	$\gg 1$
$\frac{d\psi_s}{dV_{GB}}$	Approximately constant; attains its maximum value in this region	Varies	Small
$\frac{d\psi_s}{dV_{CB}}$	Very small	Varies	Close to 1
Dependence of Q'_I on V_{GB} or V_{GC} for V_{CB} constant	Approximately exponential	—	Approximately first-degree polynomial
$\frac{d \ln Q'_I }{d \psi_s}$	$\frac{1}{\phi_I}$	Varies	$\frac{1}{2\phi_I}$

[†] It is assumed here that V_{GB} is sufficiently large that V_Q , V_W , and V_U are nonnegative; see Sec. 3.5.1 for the interpretation of negative values for these quantities.

discussed in Sec. 3.5.) The properties listed correspond to those for the two-terminal MOS structure, which were summarized in Table 2.1. Some of these properties will be discussed in the following sections, in which we look separately at individual regions of inversion. Since the considerations in each region are similar to those for the two-terminal structure of Chap. 2, we will be relatively brief.

3.4.2 Strong Inversion

Exact calculations show that, although deep in strong inversion $|Q'_I| \gg |Q'_B|$, for values of V_{GB} close to V_{HB} we can have Q'_I larger or smaller than Q'_B , depending on substrate doping concentration and oxide thickness. What really makes the structure in strong inversion behave as it does is not the relative magnitude of $|Q'_I|$ and $|Q'_B|$ but rather that of their slopes with respect to ψ_s ; these slopes are simply C'_I and C'_B . One finds that, in strong inversion,

$$C'_I \gg C'_B \quad (3.4.4)$$

As seen in Fig. 3.6a, ψ_s changes only slightly with V_{GB} in strong inversion and can be assumed “pinned” to a fixed value. With $V_{CB} = 0$, that value is the same as that corresponding to the two-terminal structure, which was denoted by ϕ_0 in (2.6.15). Since the pinned value is close to $2\phi_F + V_{CB}$, it roughly follows the dependence of this quantity on V_{CB} . Thus the value of the pinned surface potential in strong inversion can be approximated by

$$\psi_s \approx \phi_0 + V_{CB} \quad (3.4.5)$$

where

$$\phi_0 \approx 2\phi_F + \Delta\phi \quad (3.4.6)$$

with $\Delta\phi$ equal to several ϕ_F (see the related discussion in Sec. 2.6.2).

The depletion region width can then also be assumed pinned at a value d_{Bm} , which, corresponding to (2.6.18), is given by

$$d_{Bm} = \sqrt{\frac{2\epsilon_s}{qN_A}} \sqrt{\phi_0 + V_{CB}} \quad (3.4.7)$$

and the depletion region charge, from (3.2.17) and (3.4.5), is pinned at the value

$$Q'_B = -\sqrt{2q\epsilon_s N_A} \sqrt{\phi_0 + V_{CB}} \quad (3.4.8a)$$

or, using (3.4.3):

$$Q'_B = -\gamma C'_{ox} \sqrt{\phi_0 + V_{CB}}$$

(3.4.8b)

As already explained in Sec. 3.3, *for the case of strong inversion only*, V_{CB} can be interpreted as the effective reverse bias of the field-induced junction formed by the inversion layer and substrate. Changes in the potential between terminals C and B in Fig. 3.2c cause nearly equal changes in the potential ψ_s between the inversion layer and the bulk, as long as strong inversion is maintained. This is why in strong inversion $\partial\psi_s/\partial V_{CB} \approx 1$, as follows from (3.4.5).

The inversion layer charge can be obtained by using (3.4.5) in (3.2.19). The result can be written as follows:

$$Q'_I = -C'_{ox}(V_{GB} - V_{TB})$$

(3.4.9)

with

$$V_{TB} = V_T + V_{CB} \quad (3.4.10)$$

where

$$V_T = V_{FB} + \phi_0 - \frac{Q'_B}{C'_{ox}} \quad (3.4.11)$$

The quantity V_{TB} is the *G-B (gate-body) extrapolated threshold voltage*; its meaning is illustrated in Fig. 3.6d. It is the value of V_{GB} at which the preceding strong inversion equation would predict zero Q'_I , if someone wrongly used it for such low values of V_{GB} . This equation should obviously *not* be used at such low V_{GB} values, as they fall *below* strong inversion! This important point can be seen from Fig. 3.6d.

Note that, using (3.2.27) and (3.4.10), we can write (3.4.9) as

$$Q'_I = -C'_{ox}(V_{GC} - V_T) \quad (3.4.12)$$

Thus, the quantity V_T is the *G-C extrapolated threshold voltage*; it is the value of V_{GC} at which the preceding strong inversion equation would predict zero Q'_I , if one used it for such low V_{GC} values. Again, it is important to note that one should *not* be using (3.4.12) for Q'_I at V_{GC} values near V_T , as such values are *below* the onset of strong inversion. This can be seen in Fig. 3.6d.

Using (3.4.8b) in (3.4.11), we get

$$V_T = V_{FB} + \phi_0 + \gamma\sqrt{V_{CB} + \phi_0} \quad (3.4.13)$$

which can be written as

$$V_T = V_{T0} + \gamma(\sqrt{V_{CB} + \phi_0} - \sqrt{\phi_0}) \quad (3.4.14)$$

where V_{T0} is the value of V_T at $V_{CB} = 0$, given by

$$V_{T0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0} \quad (3.4.15)$$

which is, of course, the same quantity as that encountered for the two-terminal structure in (2.6.22).

In some treatments, as we have seen, $\phi_0 = 2\phi_F$ is assumed. In that case, we have, by comparing (3.4.13) to (3.4.2b),

$$V_T = V_M \quad \text{if } \phi_0 = 2\phi_F \text{ is used (a bad choice)} \quad (3.4.16)$$

However, as has already been remarked (see Sec. 2.6.2), $2\phi_F$ is in general *not* the best choice for ϕ_0 .

The plot of V_T vs. V_{CB} has been included in Fig. 3.9. The threshold increase $V_T - V_{T0}$ due to the body effect is shown vs. V_{CB} for various values of γ in Fig. 3.10.[†]

[†] Plots of $V_L - V_{L0}$, $V_M - V_{M0}$, and $V_H - V_{H0}$ have similar shapes.

We note that, although we have assumed $V_{CB} \geq 0$, our results are approximately valid even if V_{CB} is somewhat below zero (thus forward-biasing the source-body junction), as long as the V_{CB} value is not enough to cause appreciable junction current. This has been confirmed experimentally.⁴ Indeed, our discussion in terms of quasi-Fermi levels in Sec. 3.2 did not explicitly assume that $V_{CB} \geq 0$, although this was assumed in drawing Fig. 3.2.

3.4.3 Weak Inversion

In the weak inversion region, we have

$$Q'_L \ll Q'_B \quad (3.4.17)$$

Consider now (3.2.18). The magnitude of the second term under the first square root is small, because in weak inversion $\psi_s < 2\phi_F + V_{CB}$ (Fig. 3.6). Thus we can use an approach like the one that led to (2.6.26) for the two-terminal structure to obtain

$$Q'_L = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_s}} \phi_t e^{[\psi_s - (2\phi_F + V_{CB})]/\phi_t} \quad (3.4.18)$$

We remind the reader that in weak inversion, V_{CB} *cannot* be interpreted as an effective reverse bias, as explained in Sec. 3.3. In weak inversion, the surface potential is practically independent of V_{CB} and is practically equal to ψ_{sa} (see Fig. 3.6). The expression for ψ_{sa} was given in (3.2.24). Thus

$$\psi_s \approx \psi_{sa}(V_{GB}) = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 \quad (3.4.19)$$

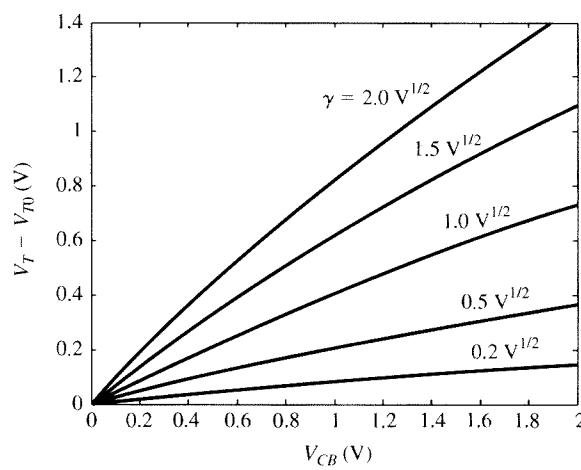


FIGURE 3.10

Increase of the extrapolated threshold voltage V_T above its value at $V_{CB} = 0$, shown vs. V_{CB} for various values of the body effect coefficient γ . A single value of $\phi_0 = 1$ V is assumed.

Therefore, in (3.4.18) the *only* term dependent on V_{CB} is $\exp(-V_{CB}/\phi_t)$. To emphasize this important conclusion, we write that equation as follows:

$$Q'_I = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_{sa}(V_{GB})}} \phi_t e^{[\psi_{sa}(V_{GB}) - 2\phi_F]/\phi_t} \times e^{-V_{CB}/\phi_t} \quad (3.4.20)$$

Dependent only on V_{GB} Dependent only on V_{CB}

For a fixed V_{CB} , Q'_I turns out to be nearly exponentially dependent on V_{GB} . To see this, we now develop an expression for Q'_I that does not contain ψ_{sa} , as we did for the two-terminal structure. In Fig. 3.11, we have repeated one of the curves of Fig. 3.7. Let V'_{CB} denote the constant value of V_{CB} for which this curve is obtained. In the figure we show the weak-inversion limits on the surface potential from Table 3.1. As seen, the width of the region in terms of ψ_{sa} is only ϕ_F . As V_{GB} changes over the region, the corresponding variation of the term $\sqrt{\psi_{sa}}$ in (3.4.20) is very small compared with the large variation of the exponential in that equation. Thus we can assume^{5,6} that $\sqrt{\psi_{sa}}$ is practically fixed at $\sqrt{2\phi_F + V'_{CB}}$ with $2\phi_F + V'_{CB}$ being the value of ψ_{sa} at the top of the weak-inversion region, as shown by point M in Fig. 3.11.[†] Thus (3.4.20) becomes

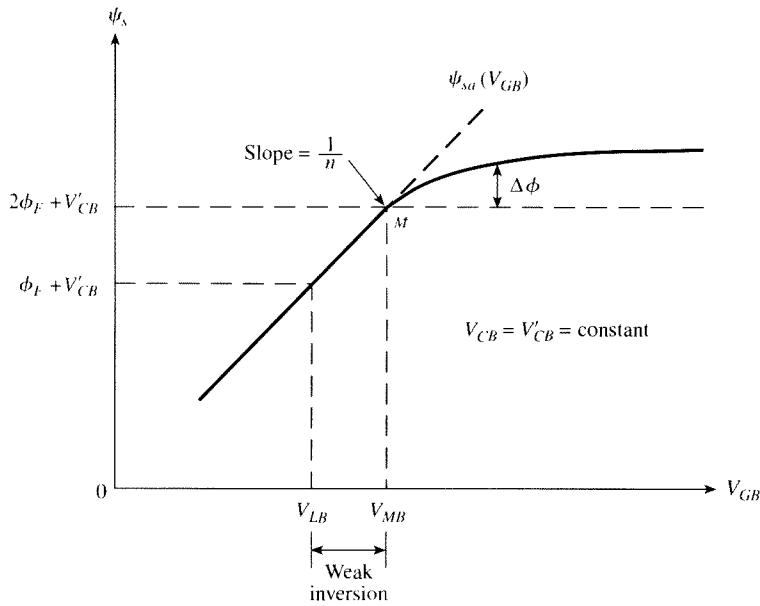


FIGURE 3.11
Surface potential vs. gate-body voltage for a given value V'_{CB} of V_{CB} .

[†] A more appropriate point⁷ would have been $1.5\phi_F + V_{CB}$, but the final results would be somewhat less convenient. At the level of approximation we are seeking, the point $2\phi_F + V_{CB}$ is acceptable.

$$Q'_L \approx -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{2\phi_F + V'_{CB}}} \phi_t e^{[\psi_{sa} - (2\phi_F + V'_{CB})]/\phi_t} \quad (3.4.21)$$

In analogy with the corresponding development for the two-terminal structure, a simplified, approximate expression can be developed based on the observation that the slope of $\psi_{sa}(V_{GB})$ is nearly constant in weak inversion. Recall that the *inverse* of this slope was denoted by n in (3.2.25), and its value was found to be[†]

$$n = 1 + \frac{\gamma}{2\sqrt{\psi_{sa}(V_{GB})}} \quad (3.4.22)$$

where $\psi_{sa}(V_{GB})$ is given by (3.4.19). A typical plot for the quantity n vs. V_{GB} is shown in Fig. 3.12. It is seen that over the width of the weak inversion region (with V_{LB} and V_{MB} determined for a given V_{CB}), it varies little.

Taking changes below point M in Fig. 3.11, we can write

$$\psi_{sa} - (2\phi_F + V'_{CB}) \approx \frac{1}{n} (V_{GB} - V_{MB}) \quad (3.4.23a)$$

$$= \frac{1}{n} (V_{GC} - V_M) \quad (3.4.23b)$$

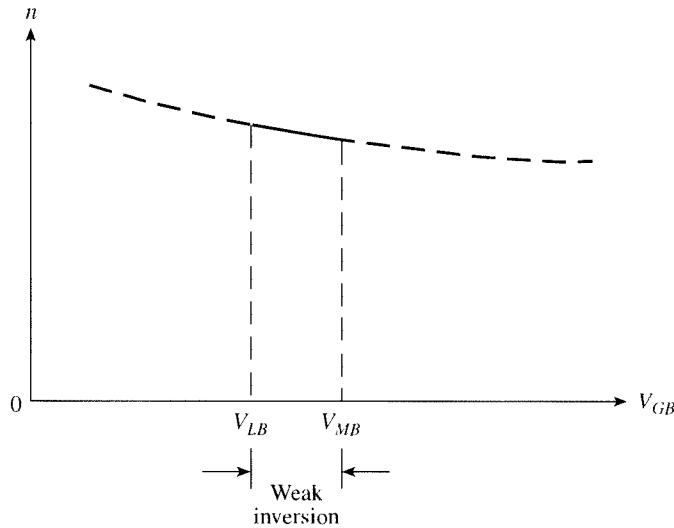


FIGURE 3.12

The parameter n as a function of the gate-body voltage; over the weak inversion region, this parameter does not vary much.

[†] If there is significant variation of the interface state charge (Sec. 2.2) with the surface potential, an extra term should be added to the right-hand side of the equation; see the discussion toward the end of Sec. 2.6.

where V_{MB} and V_M are given in Sec. 3.4.1, and n is evaluated at point M :

$$n = 1 + \frac{\gamma}{2\sqrt{2\phi_F + V'_{CB}}} \quad (3.4.24)$$

Substituting (3.4.23) in (3.4.21), we obtain

$$Q'_l \approx Q'_M e^{(V_{GB} - V_{MB})/(n\phi_t)} \quad (3.4.25a)$$

or

$$Q'_l \approx Q'_M e^{(V_{GC} - V_M)/(n\phi_t)} \quad (3.4.25b)$$

where

$$Q'_M = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{2\phi_F + V'_{CB}}} \phi_t \quad (3.4.26)$$

represents the value of Q'_l at the top of weak inversion ($V_{GB} = V_{MB}$). Equation (3.4.25) predicts exponential behavior for Q'_l in weak inversion. As was the case with the two-terminal structure, this is only approximately true. In cases where the slope of Q'_l vs. V_{GB} is important, this approximation may not be adequate.

If V_{CB} is fixed at a value V'_{CB} and V_{GB} is varied, (3.4.25a) can be a helpful equation since it makes explicit the exponential dependences of Q'_l on V_{GB} . However, if V_{GB} is fixed and V_{CB} is varied instead, (3.4.25a) can be misleading. The reason is that the dependence of Q'_l on V_{CB} is hidden in Q'_M , V_{MB} , and n , each of which depends on V_{CB} in a complicated manner. On the other hand, (3.4.20) is ideal for such cases, since it makes explicit the exponential dependence of Q'_l on V_{CB} in a simple manner.

3.4.4 Moderate Inversion

For a given V_{GB} value in moderate inversion, one can numerically solve the implicit equation (3.2.21) for ψ_s and substitute ψ_s into (3.2.18) to find Q'_l . Approximate explicit equations for ψ_s and the charges in terms of V_{GB} have been proposed.⁸⁻¹⁵ The developments of such expressions are lengthy and involve a number of approximations. Thus, one is better off using the accurate expression (3.2.18), which is valid in all regions of inversion. Using a single expression has the additional advantage that it avoids “if” statements in computer implementation and ensures smooth transitions at the boundaries between regions.

3.5 A “ V_{CB} Control” Point of View

3.5.1 Fundamentals

Our discussion of the three-terminal MOS structure so far was a natural extension of our discussion of the two-terminal structure. In both cases, we described in detail what happens when we increase the gate potential with respect to the body, V_{GB} .

Thus, the gate potential was given the prominent role of the independent variable; the potential of terminal C with respect to the substrate, V_{CB} , was a parameter, as, for example, in Fig. 3.7. [When we developed approximate expressions for the inversion layer charge in particular regions of inversion, often V_{CB} did not appear explicitly, but was rather hidden in the parameters in those equations, such as in V_{TB} in (3.4.9) and in V_{MB} in (3.4.25a).] We will refer to that view as “ V_{GB} control.”

There is another point of view, which we will call “ V_{CB} control,” in which we fix V_{GB} and observe what happens when we vary V_{CB} . Obviously, the two points of view should give equivalent results. However, each point of view adds something to our intuition and leads to a different set of approximations. The V_{CB} control viewpoint leads to a different set of models for the MOS transistor,^{6,12,14} which have found their place alongside the classical V_{GB} control models. It should be clear that “ x control,” as used here, does *not* imply any prominence of x in controlling the mechanisms in the three-terminal MOS structure, but rather refers to the way in which we choose to *describe* those mechanisms.

In order to make clear the connection between the two points of view, let us begin our discussion by returning to Fig. 3.7.

Let V_{GB} be *constant* and equal to the value V_{GB5} indicated on Fig. 3.7 (the subscript 5 is used in connection with a forthcoming discussion). As V_{CB} is raised, the level of inversion becomes lighter. For the values $V_{CB} = V_{CB1}$, V_{CB2} , and V_{CB3} , the structure is correspondingly in strong, moderate, and weak inversion. Further increases in V_{CB} will leave the surface potential value practically unaffected at the value $\psi_s(V_{GB5})$, as seen for example in going from V_{CB3} to V_{CB4} .

These observations can clearly be displayed by plotting ψ_s vs. V_{CB} , with $V_{GB} = V_{GB5}$, using (3.2.21). This is shown in Fig. 3.13. The quantities $V_Q(V_{GB5})$, $V_W(V_{GB5})$, and $V_U(V_{GB5})$ marked on the horizontal axis give the limits between strong and moderate inversion, between moderate and weak inversion, and between weak inversion and depletion, respectively, in terms of V_{CB} ; these quantities will be discussed shortly. The limit between moderate and weak inversion is seen to occur when $\psi_s \approx 2\phi_F + V_{CB}$, as

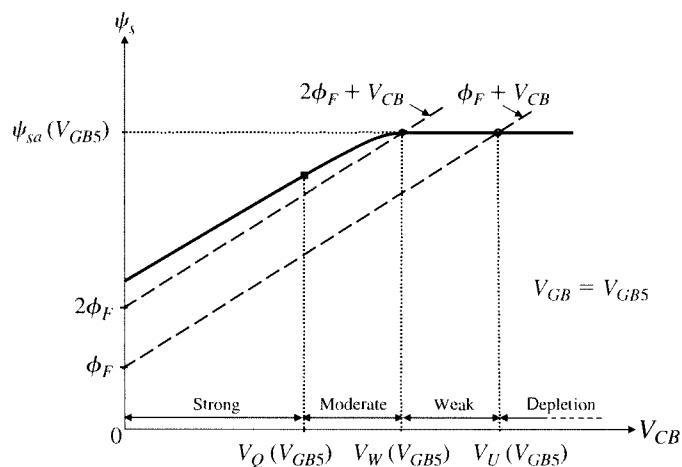


FIGURE 3.13
Surface potential vs. V_{CB} , for a constant value of V_{GB} .

expected from our earlier discussion. Similarly, the limit between weak inversion and depletion occurs when V_{CB} is increased to the point that $\psi_s = \phi_F + V_{CB}$ holds. For large V_{CB} , the inversion layer disappears and ψ_s flattens out at the value ψ_{sa} , which is given by (3.2.24) and depends only on V_{GB} . Note that as V_{CB} is increased, we go from strong, to moderate, to weak inversion. Thus, these regions are encountered in *opposite* order from the case where V_{GB} is increased (Fig. 3.7). This reflects the opposite roles of V_{CB} and V_{GB} in controlling the level of inversion.

Adding curves for other values of V_{GB} produces the plot in Fig. 3.14. The points along the curves show the limits between regions, according to the convention used in Fig. 3.14. The two curves below the top one behave in a similar manner as just discussed, only now V_{GB} is smaller ($V_{GB3} < V_{GB4} < V_{GB5}$ is assumed), and thus smaller values of V_{CB} are needed to reduce the level of inversion to a given point. The value V_{GB2} ($< V_{GB3}$) is low, so even with $V_{CB} = 0$, the surface is only in moderate inversion. Increasing V_{CB} above zero can then only drive the structure into weak inversion and eventually into depletion. Finally, V_{GB1} ($< V_{GB2}$) is so low that the device is only in weak inversion when $V_{CB} = 0$. Increasing V_{CB} above zero then will eventually drive the device into depletion.

Our analysis would seem to predict that ψ_s just to the right of the n^+ region in Fig. 3.2 could be very different from the potential on the n^+ region itself (which, from Sec. 1.5, is $V_{CB} + \phi_{bi}$), depending on V_{CB} (Fig. 3.13). This would seem to indicate a discontinuity in the electrostatic potential as the n^+p boundary is crossed. This apparent problem is caused by our neglecting the “transition” region extending on both sides of the boundary. In this transition region, the electric field is not vertical, and two-dimensional analysis becomes necessary. Such analysis can be carried out numerically and it does, indeed, predict that the potential changes from its value in the n^+ region to the value ψ_s , over a transition region the length of which is roughly equal to the depth of the depletion region under the n^+ region. The external source V_{CB} still “communicates” with the surface through the transition region, and

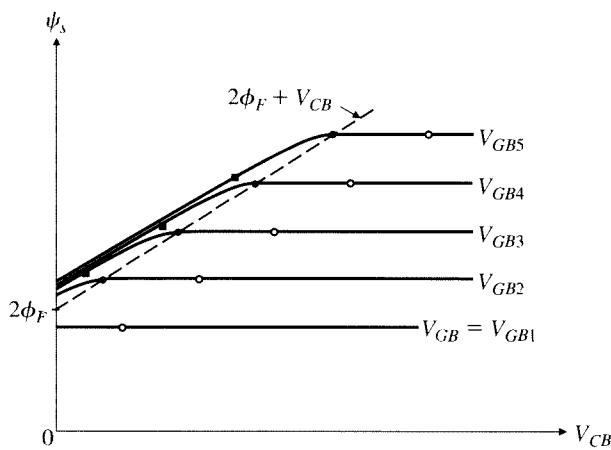


FIGURE 3.14

Surface potential vs. V_{CB} with gate-body voltage as a parameter. For each curve, a full square denotes the limit between strong and moderate inversion; a full circle, the limit between moderate and weak inversion; and an open circle, the limit between weak inversion and depletion.

the effect of V_{CB} to the right of the transition region can be expected to be as has been described.[†] Unfortunately, two-dimensional analysis does not yield manageable analytical results.

The reader is urged to qualitatively produce Fig. 3.14 from Fig. 3.7, and vice versa. This is a good way to establish the connection between the V_{GB} control and V_{CB} control points of view.

Expressions for the bounds V_Q , V_W and V_U between the various regions in terms of V_{CB} (see, for example, Fig. 3.13) can be developed by finding the values of V_{CB} needed to make the corresponding V_{GB} limits (in the second row of Table 3.1) equal to the given value of V_{GB} . Consider, for example, V_W . This quantity is the value of V_{CB} that brings the structure to the boundary between moderate and weak inversion. This means that if V_{CB} in (3.4.1) is set equal to V_W , the resulting value of V_{MB} must be equal to the given V_{GB} . In other words,

$$V_{FB} + 2\phi_F + V_W + \gamma\sqrt{2\phi_F + V_W} = V_{GB} \quad (3.5.1)$$

Solving this for V_W gives

$$V_W = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 - 2\phi_F \quad (3.5.2)$$

This quantity is an increasing function of V_{GB} . This is because, if the structure is at the boundary between weak and moderate inversion and V_{GB} is subsequently increased, inversion will become heavier. To reduce inversion to its previous level, a larger V_{CB} will be needed, and therefore V_W will be larger.

The expressions for V_W and for the other two limit points are summarized in the last row of Table 3.1.

The reader is reminded that V_{CB} is assumed nonnegative, in order for the junction formed by the n^+ region and the substrate not to become forward-biased. In some cases, the expressions in the last row of Table 3.1 can result in negative or even imaginary values. This indicates that there are no positive (or zero) values of V_{CB} , which will bring the structure to the desired point. For example, consider the bottom curve in Fig. 3.14. In this case, V_{GB} is so low that even with $V_{CB} = 0$ the structure is only in weak inversion. Raising V_{CB} above zero can only drive the structure into depletion, as already remarked. Thus, only V_U will come out to have a real, positive value, which indicates that neither strong nor moderate inversion is possible for the given low value of V_{GB} .

In Fig. 3.15, we show how the behavior of ψ_s vs. V_{CB} , which we discussed earlier, affects the charges per unit area, Q'_I and Q'_B . As before, we assume that V_{GB} is fixed. For a given V_{CB} , ψ_s can be found, as already discussed, and is shown on the right. The resulting value of ψ_s determines Q'_I and Q'_B from (3.2.20) and (3.2.17), respectively; this is shown in the left part of the figure. These plots are not shown for

[†] Recall that in a pn junction the electron quasi-Fermi level is taken practically constant from the n region through the depletion region.¹ In our case, the electron quasi-Fermi level (measured from the hole quasi-Fermi level in the bulk) corresponds to the potential V_{CB} (Fig. 3.4).

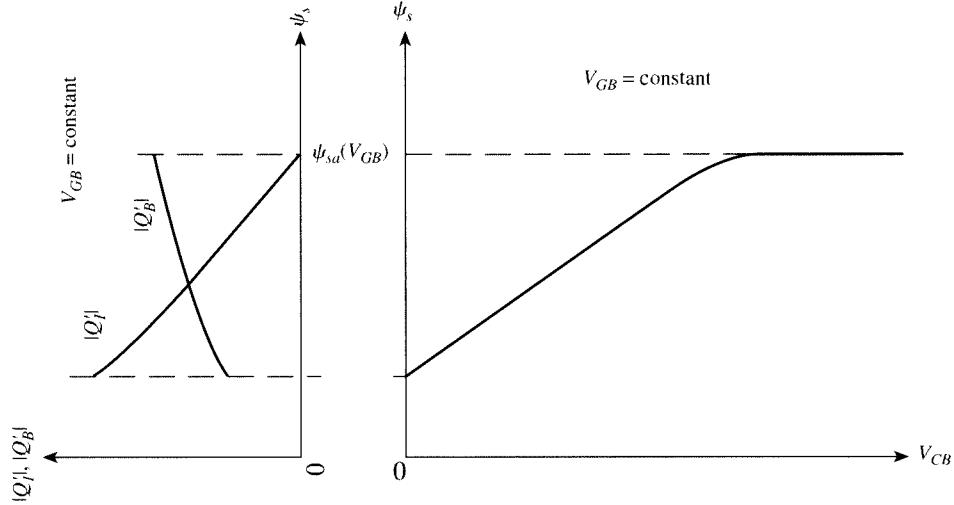


FIGURE 3.15

Surface potential vs. V_{CB} (right plot) and magnitudes of charges per unit area vs. surface potential (left plot).

values of ψ_s larger than $\psi_{sa}(V_{GB})$ because ψ_s cannot attain such values; as seen on the right, ψ_s tends asymptotically to $\psi_{sa}(V_{GB})$, as V_{CB} is increased. Thus, Q'_B is expected to tend to a constant, maximum value and Q'_I is expected to tend to zero, as V_{CB} is raised.

The behavior of $|Q'_I|$ vs. V_{CB} is shown in Fig. 3.16. This plot can be obtained from the parametric representation composed of (3.2.21) and (3.2.20). The various regions of inversion are marked along the horizontal axis. As V_{CB} is raised past V_W , the structure leaves moderate inversion and enters weak inversion; the charge tends to zero as V_{CB} is increased further. The broken line and the quantity V_p are related to approximations and will be discussed in the following subsection. We emphasize that the plot in Fig. 3.16 is for a given V_{GB} ; if V_{GB} changes, the plot will also change, as shown in Fig. 3.17.

3.5.2 The “Pinchoff Voltage”

Consider now the accurate plot of the inversion layer charge magnitude vs. V_{CB} , shown by the solid line in Fig. 3.16. As we have seen in Sec. 3.4.2, in *strong* inversion, this charge can be approximated by

$$Q'_I = -C'_{ox}[V_{GB} - V_{TB}(V_{CB})] \quad (3.5.3)$$

where, from (3.4.10) and (3.4.13),

$$V_{TB}(V_{CB}) = V_{FB} + \phi_0 + V_{CB} + \gamma\sqrt{\phi_0 + V_{CB}} \quad (3.5.4)$$

As seen, as V_{CB} is increased, $V_{TB}(V_{CB})$ rises and $|Q'_I|$ is reduced. This strong inversion approximation for $|Q'_I|$ is shown by the broken line in Fig. 3.16. As seen, it is close to a straight line, although it is *not* exactly a straight line. This plot cuts the

146 OPERATION AND MODELING OF THE MOS TRANSISTOR

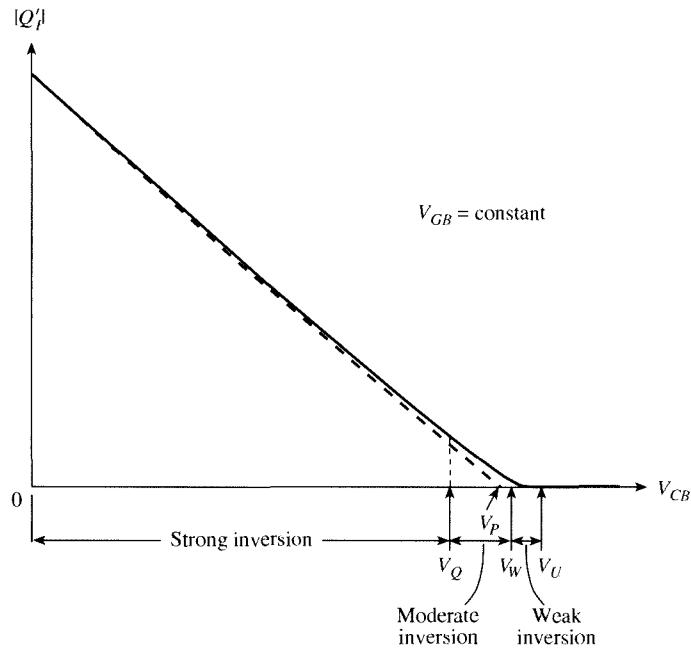


FIGURE 3.16

Inversion layer charge magnitude vs. V_{CB} , for a fixed gate-body voltage. Solid line: general model; broken line: strong inversion approximation of (3.5.3) and (3.5.4).

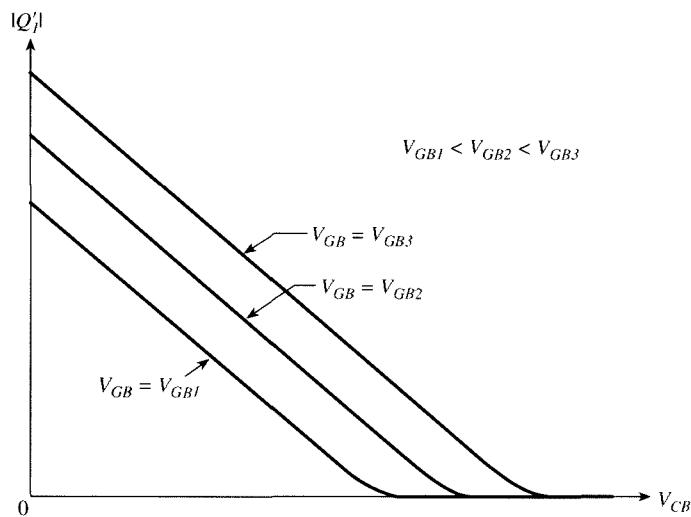


FIGURE 3.17

Inversion layer charge magnitude vs. V_{CB} , for difference gate-body voltages.

horizontal axis at some V_{CB} value, shown as V_p in the figure. This value is known as the “pinchoff voltage.” Needless to say, when the inversion layer charge is close to zero, we are not in strong inversion, so we should not be using the strong inversion expression (3.5.3) if we wanted an accurate value for Q'_i ! But here we are not after an accurate value of Q'_i ; we are only after a classical definition:

The *pinchoff voltage*, V_p , is the value of V_{CB} that makes V_{TB} equal to the externally applied gate-body voltage, and thus causes the strong inversion *approximation* to predict zero inversion layer charge.

The pinchoff voltage finds some use in simplified modeling.¹⁴ We will have use for it in this book, too, in the following chapter. Let us now find the value of V_p . According to the preceding definition

$$V_p = V_{CB} \Big|_{V_{TB} = V_{GB}} \quad (3.5.5)$$

Thus, setting (3.5.4) equal to V_{GB} and solving for $V_{CB} = V_p$, we obtain[†]

$$V_p = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 - \phi_0 \quad (3.5.6)$$

Comparing this to (3.2.24), we find

$$V_p = \psi_{sa}(V_{GB}) - \phi_0 \quad (3.5.7)$$

This result is not surprising. Assume that the surface potential is at the point where the inversion layer charge vanishes, i.e., the surface potential is just enough to cause depletion. Then it will be equal to the depletion surface potential ψ_{sa} discussed in Sec. 2.5. On the other hand, we have assumed that the surface potential in strong inversion is $\phi_0 + V_{CB}$, as in (3.4.5). If we extend the strong inversion approximation all the way to zero inversion layer charge (as required in the preceding definition of the pinchoff voltage), this implies $\phi_0 + V_{CB} = \psi_{sa}$; if we solve this for V_{CB} and call the result V_p , we obtain (3.5.7).

Simple models involving the pinchoff voltage are considered in Probs. 3.16 and 3.17.

3.6 Uses for Three-Terminal MOS Structures

Our main reason for covering the three-terminal MOS structure in this book is that it leads smoothly to the four-terminal MOS transistor, discussed in the next chapter. However, three-terminal MOS structures have interesting properties of their own, which can lead to applications. For example, the body terminal has been used as a control terminal for a MOS varactor in very low voltage circuits.²²

[†] As already mentioned, in some treatments ϕ_0 is approximated by $2\phi_F$. In such cases, V_p becomes equal to V_W , as seen from (3.5.6) and (3.5.2).

The three-terminal structure can also be used to achieve the so-called “parametric” operation.²³ In Fig. 3.8a, assume that initially V_{CB} is zero and that a voltage V_{GB} , sufficient to cause operation in strong inversion, is applied. If the connection to the gate is then eliminated, the voltage V_{GB} will be stored on the gate capacitance. If now V_{CB} is raised to a sufficiently positive value, the positive potential on terminal C will attract the electrons, which will leave the channel, thus momentarily reducing the total negative charge below the insulator. Charge balance (2.3.4) is then reestablished through a deepening of the depletion region to supply the missing charge. This implies an increase in potential across that region, from (2.5.3). The total potential on the gate, (2.3.1), will thus increase (recall that the gate is *floating* after the connection to the external source has been broken). We have thus achieved dynamic boosting of the gate voltage. It can be shown that this principle can also be used for small-signal amplification with very low power and very low noise. Interested readers can consult Ref. 23. This type of operation is explored in Prob. 3.20.

Instead of an actual three-terminal structure, one can use advantageously a transistor with its source and drain shorted together, their common connection serving as terminal C.^{22,23}

References

1. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
2. H. C. Pao and C. T. Sah, “Effects of diffusion current on characteristics of metal-oxide (insulator) semiconductor transistors,” *Solid-State Electronics*, vol. 9, pp. 927–937, 1966.
3. Y. Tsividis, “Moderate inversion in MOS devices,” *Solid-State Electronics*, vol. 25, pp. 1099–1104, 1982; see also Erratum, *Solid-State Electronics*, vol. 26, p. 823, 1983.
4. K. V. Arnim, E. Borinski, P. Segebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha, “Efficiency of body biasing in 90-nm CMOS for low-power digital circuits,” *IEEE Journal of Solid-State Circuits*, vol. 40, no. 7, pp. 1549–1556, 2005.
5. R. M. Swanson and J. D. Meindl, “Ion-implanted complementary MOS transistors in low-voltage circuits,” *IEEE Journal of Solid-State Circuits*, vol. SC-7, pp. 146–153, 1972.
6. E. Vittoz and J. Fellrath, “CMOS analog integrated circuits based on weak inversion operation,” *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 224–231, 1977.
7. R. J. Van Overstraeten, G. J. Declerck, and P. A. Muls, “Theory of the MOS transistor in weak inversion—New method to determine the number of surface states,” *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 282–288, 1975.
8. L. L. Lewlyn and J. D. Meindl, “An IGFET inversion charge model for VLSI systems,” *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 434–440, 1985.
9. M. A. Maher and C. A. Mead, “A physical charge-controlled model for MOS transistors,” pp. 211–229, P. Lesleben (editor), *Advanced Research VLSI*, Cambridge, Mass., MIT Press, 1987.
10. C.-K. Park, C.-Y. Lee, B.-J. Moon, Y. H. Byun, and M. Shur, “A unified current-voltage model for long-channel nMOSFETs,” *IEEE Transactions on Electron Devices*, vol. 38, pp. 399–406, 1991.
11. B. Iñiguez and E. G. Moreno, “Explicit C_s -continuous and general model for nMOSFETs,” *Electronics Letters*, vol. 29, pp. 1036–1037, 1993.
12. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, “An explicit physical model for the long-channel MOS transistor including small-signal parameters,” *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
13. K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, N.J., 1993.
14. C. C. Enz, F. Krummenacher, and E. A. Vittoz, “An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications,” *Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, 1995.

15. V. Altschul and Y. Shacham-Diamand, "Modeling of the MOSFET inversion charge and drain current in moderate inversion," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1909–1915, 1990.
16. I. R. M. Mansour, "On the modeling of MOS devices," *Proceedings of the Third International Symposium on Network Theory*, Yugoslavia, 1975, pp. 705–713.
17. I. R. M. Mansour, "Improved modeling of MOS devices," *Proceedings of the European Conference on Circuit Theory and Design*, Italy, 1976.
18. G. Baccarani, M. Rudan, and G. Spadini, "Analytical i.g.f.e.t. model including drift and diffusion currents," *IEEE Journal of Solid-State and Electron Devices*, vol. 2, pp. 62–68, 1978.
19. J. R. Brews, "A charge-sheet model of the MOSFET," *Solid-State Electronics*, vol. 21, pp. 345–355, 1978.
20. F. Van de Wiele, "A long-channel MOSFET model," *Solid-State Electronics*, vol. 22, pp. 991–997, 1979.
21. J. R. Brews, "Physics of the MOS transistor," chap. I, *Silicon Integrated Circuits, Part A*, D. Kahng (editor), Applied Solid State Science Series, Academic Press, New York, 1981.
22. S. Chatterjee, T. Musah, Y. Tsividis, and P. Kinget, "Weak inversion MOS varactors for 0.5 V analog integrated filters," *IEEE Symposium on VLSI Circuits*, Japan, 2005, pp. 272–275.
23. S. Ranganathan and Y. Tsividis, "Discrete-time parametric amplification based on a three-terminal MOS varactor: Analysis and experimental results," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 12, pp. 2087–2093, 2003.

Problems

- 3.1.** (a) For a device with $N_A = 10^{18} \text{ cm}^{-3}$, $t_{\text{ox}} = 2.5 \text{ nm}$, and $V_{FB} = -0.8 \text{ V}$, plot ψ_s vs. V_{GB} , with V_{GB} between 0.5 and 2.5 V and for $V_{CB} = 0.4, 0.8$, and 1.2 V. Show approximately the three regions of inversion on each curve. Assume that $V_{HB} \approx V_{MB} + 0.45 \text{ V}$.
 (b) For the same device, plot ψ_s vs. V_{CB} , with V_{CB} between 0 and 2.5 V and for $V_{GB} = 1.0, 1.5$, and 2.0 V. Show approximately the regions of inversion on each curve.
- 3.2.** Show how the $V_{CB} - V_{GB}$ plane can be separated into five regions, corresponding to accumulation, depletion, weak inversion, moderate inversion, and strong inversion. (*Hint:* In accumulation, take into account the fact that the concentration of holes is practically unaffected by V_{CB} .)
- 3.3.** For the device of Prob. 3.1:
- (a) Find V_{LB} , V_{MB} , and V_{HB} for $V_{CB} = 0 \text{ V}$; assume $V_{HB} \approx V_{MB} + 0.45 \text{ V}$.
 - (b) Plot $|Q'_l|$ vs. V_{GB} for V_{GB} between V_{LB} and $V_{HB} + 1.5 \text{ V}$. Use (3.2.18) and (3.2.21). Give the corresponding plot vs. V_{GC} .
 - (c) Repeat using a logarithmic axis for $|Q'_l|$.
 - (d) On each of the preceding plots, plot $|Q'_l|$ as given by (3.4.9) in strong inversion and as given by (3.4.25) in weak inversion. Comment on the accuracy obtainable.
 - (e) Repeat (a) through (d) for $V_{CB} = 0.8 \text{ V}$.
- 3.4.** Plot $\ln|Q'_l|$ vs. ψ_s , and verify that this plot approaches straight-line behavior in the weak and strong inversion regions. Use the process parameters of Prob. 2.6 and $V_{CB} = 0.5 \text{ V}$. Determine and compare the slopes of the plot in these two regions.
- 3.5.** It is often assumed that there is a point below which $Q'_l(V_{GB})$ is an exponential and above which it is a straight line. Attempt to find such a point on the plots of Prob. 3.3. Is this possible? How large an error is involved in such an approximation?
- 3.6.** It has been seen in Sec. 3.4.2 that if ψ_s is pinned to a fixed value, Q'_l plots as a straight line vs. V_{GB} . Is such pinning a *necessary* condition for obtaining such a straight line? If yes, why? If not, find the necessary condition for such a behavior of ψ_s .
- 3.7.** Show that a better approximation than (3.4.5) for the surface potential in *strong* inversion is

$$\psi_{s,\text{strong}} \approx 2\phi_F + V_{CB} + \phi_t \ln \left\{ \frac{1}{\phi_t} \left[\frac{1}{\gamma^2} (V_{GB} - V_{FB} - V_{CB} - \phi_0)^2 - \phi_0 - V_{CB} \right] \right\}$$

[Hint: Use (3.4.5) as an initial guess and perform one iteration using (3.2.21).] Compare the preceding expression to (3.4.5) and to (3.2.21). What happens if we attempt to use the preceding expression in moderate and weak inversion?

- 3.8. Equation (3.4.25b) was developed by using an approximation around $\psi_s = 2\phi_F + V'_{CB}$. Develop a similar expression using an expansion around $\psi_s = 1.5\phi_F + V'_{CB}$. How do the two expressions compare?
- 3.9. Show that for the structure of Fig. 3.2c, $C'_i < 0.1C'_b$ if the surface potential satisfies $\psi_s < 2\phi_F + V_{CB} - 2.3\phi_t$.
- 3.10. For a given V_{GB} , find *explicit* expressions for ϕ_M , ϕ_H , V_W , and V_Q using the procedure outlined at the end of Appendix D.
- 3.11. Mark the accurate limits of the regions of inversion on the curves of Prob. 3.3 using the definitions of Appendix D.
- 3.12. Prove (3.2.18). [Hint: Find the total semiconductor charge $Q'_C = Q'_I + Q'_B$ using basic electrostatics (Appendix A) by following a procedure similar to that of Prob. 2.12. From that find Q'_I by subtracting (3.2.17a).]
- 3.13. Verify the entries in Table 3.1.
- 3.14. Verify the entries in Table 3.2.
- 3.15. Show that, in the cases of accumulation and depletion, the results given in Sec. 2.5 can be obtained from appropriate equations in Sec. 3.2, by ignoring negligible terms.
- 3.16. This problem investigates the use of the pinchoff voltage, V_p , discussed in Sec. 3.5.2, in simple models.¹⁴ (a) Using (3.5.7), express n in terms of V_p . (b) Plot V_p vs. V_{GB} . Give the critical value of V_{GB} at which V_p becomes zero, and explain this value. Give the slope of the plot at that point in terms of n . (c) Approximate the above plot by a straight line and give its equation with n and V_{T0} as parameters. (d) Consider now the strong inversion approximation for $|Q'_I|$ shown by the broken line in Fig. 3.16. Find the slope $\partial Q'_I / \partial V_{CB}$ at the point $V_{CB} = V_p$ starting with (3.5.3). (e) Using the result in (a), show that this slope is equal to nC'_{ox} . (f) Thus, show that the inversion layer charge in strong inversion can be approximated by $Q'_I \approx -nC'_{ox}(V_p - V_{CB})$. (g) Show that, if instead of linearizing around the point V_p , we had linearized around $V_{CB} = 0$, a large error would occur near the point $V_{CB} = V_Q$ in Fig. 3.16.
- 3.17. Derive a simple equation for Q'_I in weak inversion as a function of V_{CB} and V_p , by taking advantage of (3.5.7).
- 3.18. Use implicit differentiation in (3.2.11) to develop an expression for $\partial \psi_s / \partial V_{GB}$ (keeping V_{CB} fixed). Explain how this expression correlates with the behavior seen in Figs. 3.7 and 3.14 in different regions of operation.
- 3.19. Repeat Prob. 3.18 for $\partial \psi_s / \partial V_{CB}$ (keeping V_{GB} fixed).
- 3.20. Consider the three-terminal structure in parametric operation, as described in Sec. 3.6.²³ Develop the equations that describe such operation. Using the parameters in Prob. 3.1, plot the final gate potential (after boosting) vs. its initial value (before boosting). Show that a small change in the initial value can result in a larger change in the final value.

4

The Four-Terminal MOS Transistor

4.1 Introduction

The MOS transistor is obtained by adding one more terminal to the structure of Chap. 3, so that the inversion layer is contacted at two opposite ends. By applying a voltage between these ends, a current can be caused to flow in the inversion layer. Since the density of carriers available for conduction depends on the gate potential, the latter can be used to either create or eliminate the inversion layer (i.e., turn the device “on” or “off”) for digital applications or to modulate its conduction in a continuous manner for analog applications.

The path from the conception of the basic MOS transistor principle to the demonstration of working devices is spread over a quarter century;^{1–4} see also Ref. 5. The basic theory of MOS transistor operation was developed in the early 1960s.^{6–9} Extensive treatments on the MOS transistor were published late in that decade, including investigations of the role of diffusion current and the role of substrate charge.^{10–18} Our list of references to the work of the 1960s is by no means complete, and many more references can be found in several early texts that helped spread the knowledge of the device.^{11,18–20} Starting in that decade, the industry applied the newly acquired knowledge to the successful fabrication of digital ICs, and the need arose for efficient models for computer-aided design.^{21–25} At the same time, the current in weakly inverted channels received considerable attention.^{26–39} Starting in the 1970s, attempts have been made to produce models valid for all combinations of external bias voltages by including drift and diffusion currents,^{40–47} but in a simpler manner than that of early attempts.¹⁰ These references are only indicative of early work in the topics mentioned in this chapter; further references on such topics will be given along the way. Extensive work has also been done on mobility behavior,

ion-implanted devices, short- and narrow-channel devices, noise, the modeling of charges and capacitances, high-frequency operation, and other topics, as well as on extensive models for computer-aided design, incorporating many different effects. Related references will be given as such topics are discussed in this and in subsequent chapters.

The basic structure of a MOS transistor and a brief overview of its operation and characteristics were given in Sec. 1.6. In this chapter, our goal will be to determine the drain current for any combination of dc terminal voltages. Throughout the chapter, we will assume that the channel is sufficiently *long* and *wide*, so that edge effects are confined to a negligible part of it; this statement will become more quantitative later on. We will also assume that the substrate is *uniformly* doped. (The doping will be assumed to be *p* type unless indicated otherwise.[†]) There have been many models proposed for this basic case, and many readers are likely to encounter several of them in the course of their work. In this chapter, we will introduce these models in a systematic way. Although each model is usually derived in the literature independently of the others, here we will place all of them in context, by showing how they can all be derived from one important model, and how they relate to that model and to each other. Readers who do not desire a very detailed treatment can skip certain sections without loss of continuity.

An *n*MOS transistor with external dc voltages applied is shown in Fig. 4.1. The channel extends from $x = 0$ to $x = L$. In this figure, and all related figures in this book, we show only the part of the depletion region extending into the substrate. The extent of this region into the heavily doped n^+ regions is much narrower and is not shown for simplicity. (The complete picture of the depletion region is qualitatively similar to that shown in Fig. 1.21.) In Fig. 4.1a, we show a common body connection for the voltage sources, whereas in Fig. 4.1b, we show a common source connection; both cases will prove useful in our ensuing discussion. We will be able to apply at the source end of the channel the results we developed for the structure of Fig. 3.2c with $V_{CB} = V_{SB}$. Similarly, at the drain end, we will be able to apply those results with $V_{CB} = V_{DB}$. It should be emphasized that *normal operation of a MOS transistor requires that both pn junctions be reverse-biased*.[‡] Thus, for an *n*MOS transistor it will be implicitly assumed that[§]

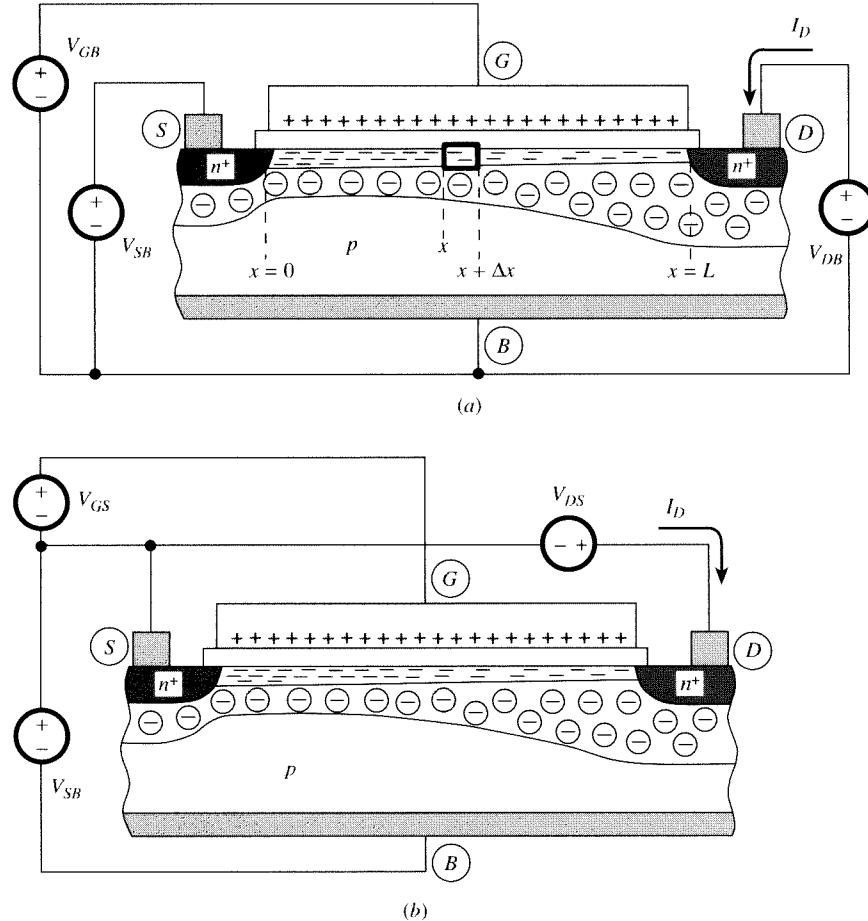
$$V_{SB} \geq 0 \quad (4.1.1)$$

$$V_{DB} \geq 0 \quad (4.1.2)$$

[†] Nonuniform body doping, introduced during fabrication through the process of ion implantation, is used to control several key characteristics of MOS transistors, such as threshold voltage value, capacitances, and short-channel effects. The equations we will develop here are widely used even in such cases, sometimes with simple modifications; these will be mentioned as we go along and will be undertaken in detail in Chap. 9.

[‡] An exception will be discussed later.

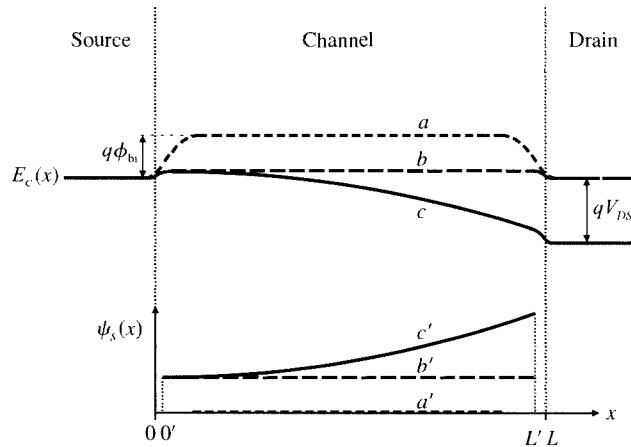
[§] In MOS transistor models used in circuit simulation, diodes must be included between source and body, and between drain and body. This takes care of cases where (4.1.1) and (4.1.2) are not satisfied in the circuit and also models the leakage currents of the reverse-biased n^+p junctions. The n^+p junction currents are assumed to be modeled separately in this way and are not considered here.

**FIGURE 4.1**

An n-channel MOS transistor. (a) Terminal voltages referred to the body; (b) terminal voltages referred to the source.

We assume that the external dc voltages have been applied for a long time, so that all charges have reached steady-state values.

It is useful at this point to consider a simple energy band diagram *at the surface*, as shown in Fig. 4.2. In the following discussion, we use the energy barrier concepts discussed in the beginning of Sec. 3.2. In the example of Fig. 4.2, we assume $V_{SB} = 0$ for simplicity. Also, until further notice, we assume $V_{DB} = V_{SB}$. In the upper part, we show the energy level E_c (bottom of conduction energy band) vs. position x (recall that this energy corresponds to the potential energy of electrons). In the lower part, we show the corresponding surface potential. Assume first that the gate-body voltage V_{GB} is equal to the flat-band voltage. Then the body is a neutral p region, and between the source and the body we have a simple n⁺p junction; similarly, next to the drain we have another such junction. Using the concepts

**FIGURE 4.2**

Upper part: conduction energy band edge (E_c) of the surface for (a) $V_{GB} = V_{FB}$, (b) V_{GB} in strong inversion and $V_{DB} = V_{SB}$, (c) V_{GB} as in (b), but $V_{DB} > V_{SB}$. Lower part: corresponding surface potential (ψ_s).

of Sec. 3.2, we see that the variation of E_c at the surface vs. horizontal position is as shown by curve a ; this variation takes place over the depletion regions of the two junctions. We will concentrate on the channel region *between* these depletion regions, and we will only show the corresponding surface potential for that region. The total variation of E_c is equal to $q\phi_{bi}$, where ϕ_{bi} is the junction built-in potential. It can be seen that the electrons in the n^+ regions face an energy barrier of this height, which makes it difficult for them to enter the channel. Thus the channel stays electrically neutral, which, as we have discussed in Chap. 2, means that the surface potential ψ_s is zero, as shown by curve a' in the figure.

Assume now that V_{GB} is increased; this will cause the surface potential to become positive, as indicated by curve b' . The corresponding electron potential energy at the surface of the body will be reduced. Thus, the energy barrier is lowered (curve b), which allows the electrons in the n^+ regions to enter the channel and form an inversion layer. Another way to look at this phenomenon is to think of the positive surface potential on the p side as *forward-biasing* the junctions, like the case of the junction discussed in Sec. 1.5. Just as was discussed for that junction, the forward bias causes the electrons to diffuse into the channel. Note that the energy barrier has not been eliminated completely; this is because the surface potential, already calculated in Chap. 2, is not exactly equal to the built-in potential of the junction.

Let us assume that the surface potential is enough to cause strong inversion. Maintaining V_{GB} at its previous value, let us raise V_{DB} above V_{SB} . This raises the surface potential near the drain, as shown by curve c' . Increasing the potential at the drain corresponds to lowering the electron energy there, as shown by curve c . From curve c' , we see that the direction of the electric field now is such that it causes electrons to move toward the right; a channel current results.

As already discussed, next to the n^+ regions there are transition regions, corresponding to the difference between the potentials at the source/drain and the channel; for example, for the case (c, c') we have such transition regions extending from 0 to $0'$ along the x axis, and from L' to L . In this chapter, we will only consider the channel between points $0'$ and L' , and our surface potential plots will be limited to this region (see, for example, curve c'). We will assume that the extent of the transition regions is a negligible part of the total channel length L ; for simplicity, then, we will assume that $0'$ coincides with 0, and L' coincides with L . We will postpone a discussion of what happens when the channel is short until Chap. 5.

If $V_{DB} = V_{SB}$ in Fig. 4.1a, we have a situation similar to that in Fig. 3.2c, and the electric field in the semiconductor is perpendicular to the surface (except at points too close to the n^+ regions). If now $V_{DB} \neq V_{SB}$, there will be a nonzero component of the electric field in the horizontal direction. *The horizontal field component will be assumed to be much smaller than the vertical component*; likewise, the variation of the horizontal field component along the x direction will be assumed to be much smaller than the variation of the vertical field component along the y direction. The field direction is assumed to be practically the same as in Fig. 3.2c, and the equations developed in reference to that figure will be appropriately extended for use here. (This approximation is referred to as the *gradual channel approximation* and can be verified with two-dimensional numerical computations; however, we will also encounter cases later where this approximation fails.)

In this chapter, we will view the oxide and the depletion region under the channel as perfectly insulating layers. This means that for the gate current I_G and for the body current I_B we will assume

$$I_G = 0 \quad (4.1.3)$$

$$I_B = 0 \quad (4.1.4)$$

It will be seen in Chap. 5 that this assumption can be violated in devices with very small dimensions. In these, a nonzero I_G can flow, as well as a nonzero I_B , which must be added to the channel current to give the total drain current. Also, the drain-body junction can have a leakage current, which must also be included in the drain current. Thus, in general, the drain current I_D can be different from the channel current, or *drain-to-source current*, I_{DS} . In this chapter, unless noted otherwise, we will assume that these effects are negligible, and thus

$$I_D = I_{DS} \quad (4.1.5)$$

Whereas in Fig. 3.2c the various charges were assumed to be uniformly distributed along the x direction (except for points too close to the n^+ region), this will, in general, not be the case in Fig. 4.1, because potentials will vary with x (curve c' in Fig. 4.2). Thus, charges per unit area should be defined *locally*. For example, consider a small area element ΔA of the channel (as seen from above in Fig. 4.1) centered around x , and the inversion layer charge ΔQ_I contained in it. The inversion layer charge per unit area Q'_I at point x is approximately given by $\Delta Q_I / \Delta A$, provided

that ΔA is small enough (but not too small, in order not to have to consider the discrete nature of charge); similarly for Q'_B and Q'_G . In other words, we define

$$Q'_I = \frac{\Delta Q_I}{\Delta A} \quad (4.1.6)$$

$$Q'_B = \frac{\Delta Q_B}{\Delta A} \quad (4.1.7)$$

$$Q'_G = \frac{\Delta Q_G}{\Delta A} \quad (4.1.8)$$

These quantities will in general depend on x .

4.2 Transistor Regions of Operation

Typical sets of dc current-voltage characteristics for an *n*MOS transistor are shown in Fig. 4.3*a* and *b*, corresponding to Fig. 4.1*a* and *b*, respectively. If the *logarithm* of the current is plotted on the vertical axis, plots of the type shown in Fig. 4.4 are obtained. This type of plot is more fair to the moderate and weak inversion regions, which are very important in low-voltage, low-power applications.

On each curve in Figs. 4.3 and 4.4, the initial, sloped part is termed *nonsaturation*. With sufficiently high drain potential, the curve flattens out and we have *saturation*.[†] The boundary between nonsaturation and saturation will be discussed later.

In general, the drain current depends on the terminal voltages, in a complicated manner. However, for certain combinations of terminal voltages, simplifications become possible and relatively simple expressions can be developed for the current. Anticipating this, we will define transistor regions of operation as shown in Table 4.1. These definitions are consistent with long-established practice, and the rationale behind them will be seen in subsequent sections. As a mnemonic aid it should be noted that:

The name of an inversion region coincides with the level of inversion at the more heavily inverted channel end.

For convenience in drawing certain figures (like Fig. 4.1) or providing certain plots, we will be assuming that the more heavily inverted channel end is the one next to the *source*, unless noted otherwise. This does not restrict generality, since the source and drain terminals in Fig. 1.30 are equivalent.

[†] In some of the literature, the names “nonsaturation” and “saturation” are used in conjunction with strong inversion only. Since the same qualitative behavior is observed also in moderate and weak inversion, we will use these names in any region.

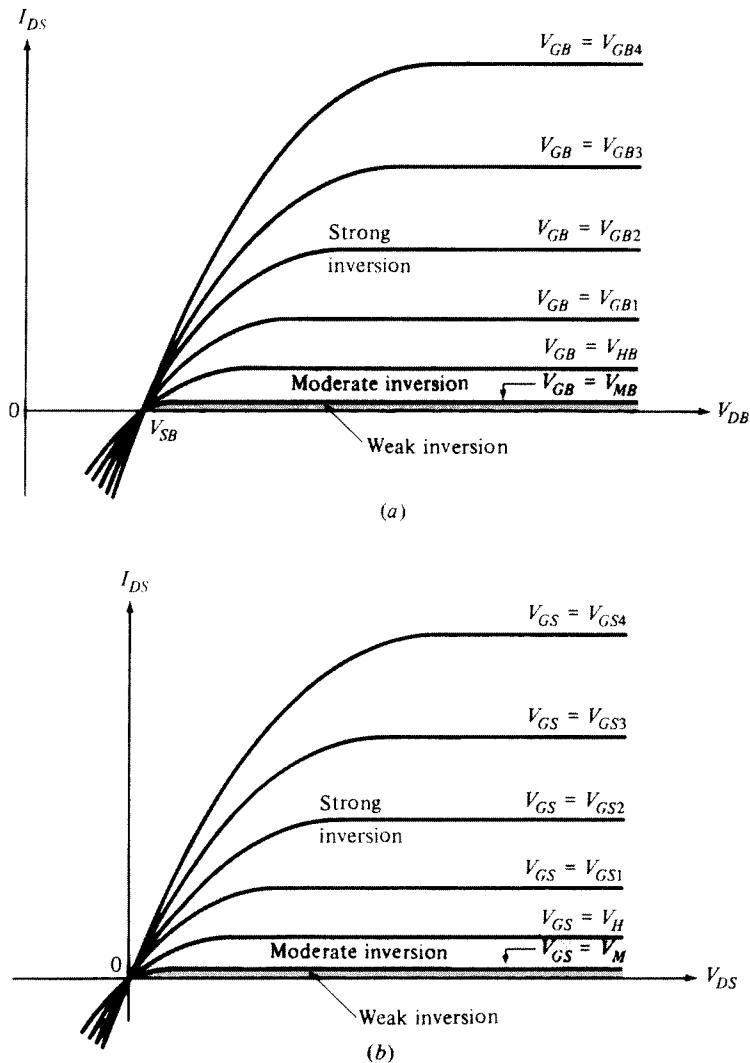


FIGURE 4.3

Current-voltage characteristics corresponding to Fig. 4.1. In both cases V_{SB} is assumed fixed, and $V_{GBi} = V_{GSi} + V_{SB}$.

TABLE 4.1
Transistor regions of operation

Region	Channel condition
Strong inversion	The more heavily [†] inverted channel end is in strong inversion
Moderate inversion	The more heavily [†] inverted channel end is in moderate inversion
Weak inversion	The more heavily [†] inverted channel end is in weak inversion

[†] If both ends are equally inverted, either end can be considered.

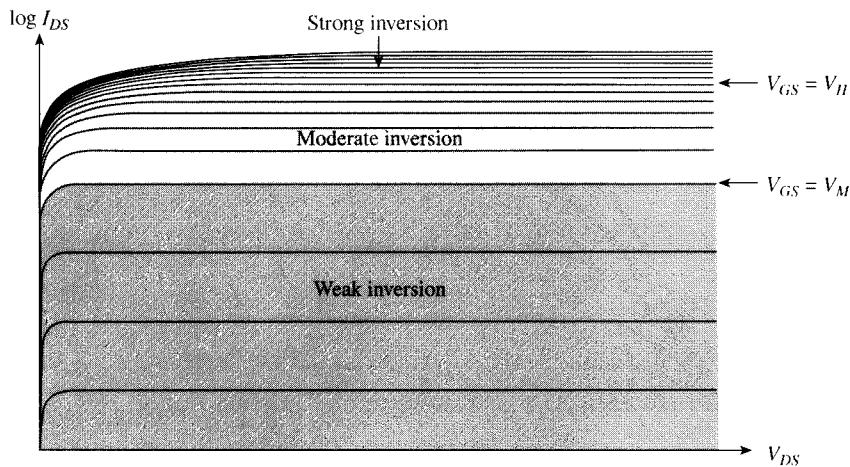


FIGURE 4.4
Logarithm of drain-source current vs. drain-source voltage, with gate-source voltage as a parameter.

There are many models for the MOS transistor. Some are for particular regions of inversion only, and in this book this will be indicated as part of their name. Other models use a single expression for each quantity modeled (e.g., the drain current) that is valid in all regions; these will be called “all-region models” in this book.[†] We now discuss such a model.

4.3 Complete All-Region Model

Current Equations In this section, we derive expressions for the drain current that are valid in all regions of inversion. We can refer to the result as an “all-inversion-region model,” which we will shorten to “all-region model.” (Inversion, which is the only region of interest in this chapter, will be implied.) In the section title, “complete” refers to the fact that no simplifications will be used beyond the basic premises of the model (in contrast with the models in Sec. 4.4, for which further simplifications will be sought). All the models we will be discussing, unless stated otherwise, will be *charge sheet* models; this refers to the basic assumption in the models, i.e., that the inversion layer is of infinitesimal thickness.[‡] We note that we have already made this assumption in Chaps. 2 and 3. We will use expressions from those chapters

[†] Sometimes such models are called “single-piece models” to distinguish them from models that use separate expressions for each region of inversion and “stitch” those expressions together using various techniques. In this book, we will not emphasize models using such stitching. The term *all-region* model will only be used to describe models with expressions valid in all regions.

[‡] Note, nevertheless, that when drawing a figure it is convenient to use a finite thickness for the inversion layer.

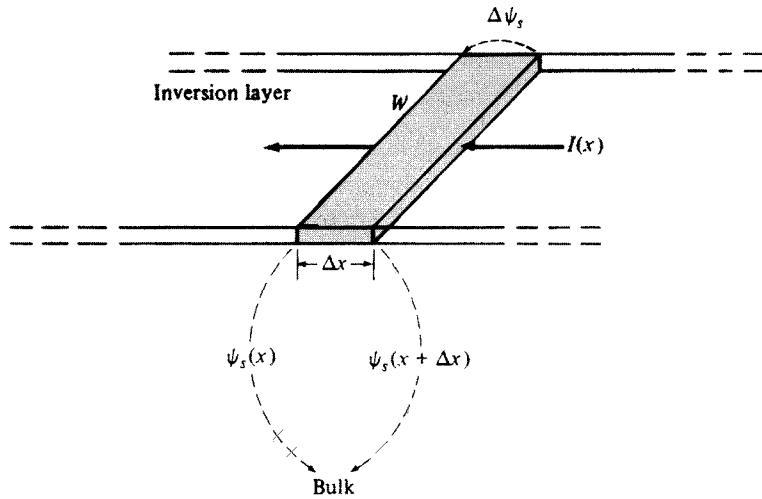


FIGURE 4.5
Small element of the inversion layer in the device of Fig. 4.1.

in our present development. The model we will present has been derived in several references,^{40,41,43–46} but our derivation will be simpler.

A key to the generality of the results we are about to develop is the observation that the current in the channel can be caused by both drift and diffusion (Sec. 1.3). Thus, let x be the horizontal position in the channel, measured from the source end. If the inversion layer current at x is denoted by $I(x)$, we will have

$$I(x) = I_{\text{drift}}(x) + I_{\text{diff}}(x) \quad (4.3.1)$$

To write an expression for the drift component, consider a small element in the inversion layer between x and $x + \Delta x$ in Fig. 4.1a, as shown magnified in Fig. 4.4. The potential difference across this element is $\Delta\psi_s(x) = \psi_s(x + \Delta x) - \psi_s(x)$. Comparing this figure to Fig. 1.12 [and assuming that the electron velocity is proportional to the small horizontal electric field, analogous to (1.3.8)], it is obvious that we can use (1.3.10) with V replaced by $\Delta\psi_s(x)$, a replaced by Δx , b by W , and $|Q'|$ by $-Q'_I$, where Q'_I is the (negative) inversion layer charge per unit area at x . Also, μ_B must be replaced with a *smaller* mobility value because electrons move with difficulty parallel to the surface (the semiconductor-oxide interface), being pulled toward it by the vertical field. The corresponding mobility will be called the *surface mobility* and will be denoted by μ . (This quantity will be considered in more detail in Sec. 4.11; for now it can be assumed that it has roughly half the value of μ_B .) We will thus have

$$I_{\text{drift}}(x) = \mu(-Q'_I) \frac{W}{\Delta x} \Delta\psi_s(x) \quad (4.3.2)$$

which, allowing Δx to approach zero, becomes

$$I_{\text{drift}}(x) = \mu W(-Q'_l) \frac{d\psi_s}{dx} \quad (4.3.3a)$$

The diffusion current component can be obtained as in (1.3.20):

$$I_{\text{diff}}(x) = \mu W \phi_t \frac{dQ'_l}{dx} \quad (4.3.3b)$$

In dc steady state, which is the only case discussed in this chapter, the total current in the channel must be the same for all x and equal to the drain-to-source current (which, under our assumptions of Sec. 4.1, is equal to the drain current). Using this fact and (4.3.3) in (4.3.1), we obtain

$$I_{DS} = \mu W(-Q'_l) \frac{d\psi_s}{dx} + \mu W \phi_t \frac{dQ'_l}{dx} \quad (4.3.4)$$

Let the surface potential at the source end of the channel ($x = 0$) be denoted by ψ_{s0} , and Q'_l there by Q'_{l0} . Let the corresponding quantities at the drain end of the channel ($x = L$) be denoted by ψ_{sL} and Q'_{lL} . Integrating (4.3.4) from $x = 0$ to $x = L$, we obtain

$$\int_0^L I_{DS} dx = W \int_{\psi_{s0}}^{\psi_{sL}} \mu(-Q'_l) d\psi_s + W \phi_t \int_{Q'_{l0}}^{Q'_{lL}} \mu dQ'_l \quad (4.3.5)$$

Since I_{DS} is independent of x it can be moved outside the integral. Thus the left-hand side is equal to $I_{DS}L$ and we have

$$I_{DS} = \frac{W}{L} \left[\int_{\psi_{s0}}^{\psi_{sL}} \mu(-Q'_l) d\psi_s + \phi_t \int_{Q'_{l0}}^{Q'_{lL}} \mu dQ'_l \right] \quad (4.3.6)$$

Thus we can view I_{DS} as consisting of two components I_{DS1} and I_{DS2} .^{48,49}

$$I_{DS} = I_{DS1} + I_{DS2} \quad (4.3.7)$$

where I_{DS1} is due to the presence of drift:

$$I_{DS1} = \frac{W}{L} \int_{\psi_{s0}}^{\psi_{sL}} \mu(-Q'_l) d\psi_s \quad (4.3.8a)$$

and I_{DS2} is due to the presence of diffusion:

$$I_{DS2} = \frac{W}{L} \phi_t \int_{Q'_{l0}}^{Q'_{lL}} \mu dQ'_l \quad (4.3.8b)$$

The interpretation of I_{DS1} and I_{DS2} requires some caution. Note that, in general, there may not be single values of a drift current and a diffusion current in the channel, since $I_{\text{drift}}(x)$ and $I_{\text{diff}}(x)$ in (4.3.3) are functions of position.⁵⁰ Nevertheless,

the preceding development makes it clear that I_{DS1} is there because drift was assumed to be present in the channel; were there no drift, there would be no I_{DS1} . Similarly, I_{DS2} is there because it was assumed that diffusion was present in the channel. It is easy to see that I_{DS1} and I_{DS2} are the mean values, over x , of the drift and diffusion currents in (4.3.3).

We now make the assumption that μ is constant along the channel (the more general case will be discussed in Sec. 4.11). Then μ can be moved outside the integral in (4.3.8), and we have

$$I_{DS1} = \frac{W}{L} \mu \int_{\psi_{s0}}^{\psi_L} (-Q'_L) d\psi_s \quad (4.3.9a)$$

$$I_{DS2} = \frac{W}{L} \mu \phi_t (Q'_{IL} - Q'_{I0}) \quad (4.3.9b)$$

To evaluate now I_{DS1} and I_{DS2} , we need Q'_L as a function of ψ_s . An appropriate expression has been derived in (3.2.19) and is repeated below. (The expression is, of course, assumed valid here because of the “gradual channel approximation” made in Sec. 4.1.)

$$Q'_L = -C'_{ox}(V_{GB} - V_{FB} - \psi_s) - Q'_B \quad (4.3.10)$$

Using this in (4.3.9), we get

$$I_{DS1} = \frac{W}{L} \mu C'_{ox} \left[(V_{GB} - V_{FB})(\psi_{sL} - \psi_{s0}) - \frac{1}{2}(\psi_{sL}^2 - \psi_{s0}^2) \right] + \frac{W}{L} \mu \int_{\psi_{s0}}^{\psi_L} Q'_B d\psi_s \quad (4.3.11a)$$

$$I_{DS2} = \frac{W}{L} \mu C'_{ox} \phi_t (\psi_{sL} - \psi_{s0}) - \frac{W}{L} \mu \phi_t [Q'_B(\psi_{sL}) - Q'_B(\psi_{s0})] \quad (4.3.11b)$$

If the reader follows the development of these two equations, it will be evident that they are not based on any particular form of $Q'_B(\psi_s)$; they have been based only on a few fundamental facts such as charge balance, potential balance, the gate/oxide charge potential relation, and the current equation in (4.3.4). The generality of these results will be found useful in several instances in this book, e.g., in discussing non-uniform substrates (Chap. 9). For now, though, we return to the bulk charge expression we have derived for uniform substrates under the charge sheet assumption, (3.2.17b), repeated for convenience here:

$$Q'_B = -\gamma C'_{ox} \sqrt{\psi_s} \quad (4.3.12)$$

This, from (4.3.10), corresponds to the following form of the inversion layer charge:

$$Q'_L = -C'_{ox}(V_{GB} - V_{FB} - \psi_s - \gamma \sqrt{\psi_s}) \quad (4.3.13)$$

Using (4.3.12) in (4.3.11), we get the current components due to drift, I_{D1} , and due to diffusion, I_{D2} , as follows:

$$I_{DS1} = \frac{W}{L} \mu C'_{ox} \left[(V_{GB} - V_{FB})(\psi_{sL} - \psi_{s0}) - \frac{1}{2}(\psi_{sL}^2 - \psi_{s0}^2) - \frac{2}{3}\gamma(\psi_{sL}^{3/2} - \psi_{s0}^{3/2}) \right] \quad (4.3.14a)$$

$$I_{DS2} = \frac{W}{L} \mu C'_{ox} \left[\phi_t(\psi_{sL} - \psi_{s0}) + \phi_t \gamma(\psi_{sL}^{1/2} - \psi_{s0}^{1/2}) \right]^{\dagger} \quad (4.3.14b)$$

Evaluating ψ_{s0} and ψ_{sL} The only step remaining is the evaluation of ψ_{s0} and ψ_{sL} from the externally applied voltages in Fig. 4.1a. Comparing this figure with Fig. 3.2c, we note that the expressions developed for the latter can be used at the source end of the channel by replacing V_{CB} with V_{SB} . Similarly, such expressions can be used at the drain end of the channel by replacing V_{CB} with V_{DB} . Thus, writing (3.2.21) for the source end and the drain end of the channel, we obtain^{40–46}

$$\psi_{s0} = V_{GB} - V_{FB} - \gamma \sqrt{\psi_{s0} + \phi_t e^{(\psi_{s0} - 2\phi_F - V_{SB})/\phi_t}} \quad (4.3.15a)$$

$$\psi_{sL} = V_{GB} - V_{FB} - \gamma \sqrt{\psi_{sL} + \phi_t e^{(\psi_{sL} - 2\phi_F - V_{DB})/\phi_t}} \quad (4.3.15b)$$

These equations can be solved for ψ_{s0} and ψ_{sL} by iteration. This can easily be done with a computer. Methods to speed up the computation of ψ_{s0} and ψ_{sL} can be found in the literature, where several implementations of the complete all-region model are discussed.^{51–62}

Two techniques have emerged as the preferred approaches to compute ψ_s . The first uses an iterative numerical procedure, in which a good initial estimate is made of ψ_s and then this is refined using a first- or second-order Newton-Raphson scheme.^{55,56,58} Convergence to tight tolerances can be achieved in two to three iterations. The second approach is based on analytical approximations. These start from very good approximations to ψ_s for specific regions of operation, and then use either smoothing functions (Chap. 10) or fixed refinement steps to produce highly accurate, explicit approximations to ψ_s , with errors compared to accurate numerical solutions of less than 1 nV.^{52,59–61} The computation time and numerical robustness of both the iterative and analytical approaches is similar,⁵⁸ although it has been noted that changes in the number of iterations for different biases can lead to “micro” discontinuities in iterative solutions.⁶²

[†] The reader is warned that this equation, although theoretically sound, can cause numerical difficulties in weak inversion. This is explained later in this section. An improvement is suggested in Prob. 4.2.

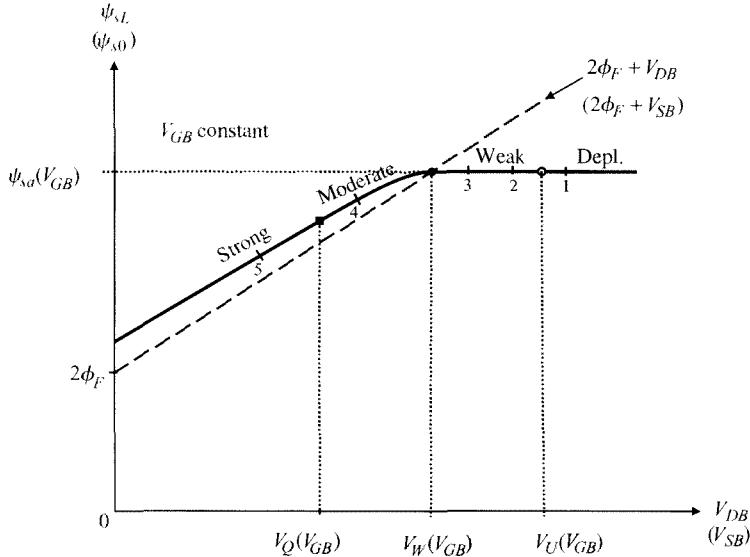


FIGURE 4.6
Surface potential at the drain (source) end of the channel vs. drain-substrate (source-substrate) voltage for a given gate-substrate voltage.

Some expressions in the all-region model involve the difference $\psi_{sL} - \psi_{s0}$, which is small in weak inversion even though neither ψ_{s0} nor ψ_{sL} is small, and so is sensitive to the accuracy of the procedure used to calculate the surface potentials (Prob. 4.2). Numerical robustness can be improved by rewriting the expressions in a numerically robust way,[†] and/or by directly solving for $\psi_{sL} - \psi_{s0}$.⁶³

A plot of ψ_{sL} vs. V_{DB} (or of ψ_{s0} vs. V_{SB}) is shown in Fig. 4.6. The symbols used on the figure are defined in Chap. 3. The plot saturates at a value dependent on V_{GB} , as given by (3.2.24). The numbered points are in relation to an upcoming discussion.

Note that by neglecting the edge effects near the drain we have ended up with a potential ψ_{sL} at the drain end of the channel that will, in general, be different from the potential in the n^* region. As expected from the associated discussion in Sec. 4.1, a more detailed picture would include a transition region containing the boundary, over which the potential would change continuously and, eventually, would become equal to the potential in the n^* region. The length of this transition region may be

[†] For example, consider $I_{DS} = I_{DS1} + I_{DS2}$ as obtained using (4.3.14). We can factor out $\psi_{sL} - \psi_{s0}$; this gives

$$I_{DS} = \frac{W}{L} \mu C'_{ox} \left[V_{GB} - V_{FB} - \frac{1}{2}(\psi_{sL} + \psi_{s0}) - \frac{2}{3} \gamma \frac{\psi_{sL} + \psi_{s0}^{1/2} \psi_{s0}^{1/2} + \psi_{s0}}{\psi_{sL}^{1/2} + \psi_{s0}^{1/2}} + \phi_i \left(1 + \frac{\gamma}{\psi_{sL}^{1/2} + \psi_{s0}^{1/2}} \right) \right] (\psi_{sL} - \psi_{s0})$$

This form is less sensitive than (4.3.14) to limitations caused by finite precision arithmetic.⁶⁴

expected to be roughly the same as the depth of the depletion region under the drain. In this region, one cannot assume that the electric field is practically vertical. Similar comments apply to the source end of the channel. Based on such a picture, one could interpret L in the preceding analysis as representing the length of the channel, excluding the source and the drain transition regions, and ψ_{s0} and ψ_{sL} as the surface potentials to the right and left of these regions, respectively. Then (4.3.15) would again be used, assuming the effect of V_{SB} and V_{DB} at the two points remained essentially the same.[†] However, for the long channels assumed in this chapter, we can continue viewing L as the total source-drain spacing. Note that in our analysis we are forced to neglect the preceding effects since they cannot be handled by the “gradual channel approximation,” which we have adopted to be able to obtain simple analytical results. An investigation of the preceding effects would have made necessary the use of two-dimensional analysis requiring a numerical solution, or at least a pseudo-two-dimensional analysis (Chap. 5).

We now fix V_{SB} , keep V_{GB} as a parameter, and vary V_{DB} . We find the corresponding ψ_{s0} and ψ_{sL} from (4.3.15) and substitute them in (4.3.14). Adding I_{DS1} and I_{DS2} produces I_{DS} as in (4.3.7), and gives the plots of Fig. 4.3a, where the current is shown vs. V_{DB} , with V_{GB} as a parameter. These plots can be converted to the ones in Fig. 4.3b, where we show the current vs. $V_{DS} = V_{DB} - V_{SB}$, with $V_{GS} = V_{GB} - V_{SB}$ as a parameter. Plotting $\ln I_{DS}$, rather than I_{DS} , produces Fig. 4.4. In these figures, the regions of operation are marked according to Table 4.1, with the values of V_{GB} or V_{GS} delineating these regions denoted by the symbols defined in Chap. 3. These values are not of interest at this point. What is important is to note that the *single* expression (4.3.7) predicts the current in all these regions. Experiments agree well with this expression.^{43–46,51–56}

Note that all curves in Fig. 4.3a saturate for large V_{DB} . This can be understood by relating the drain current for one of the V_{GB} values (say, V_{GB4} in Fig. 4.3a) to the corresponding surface potential ψ_{sL} at the drain. This is shown in Fig. 4.7: increasing V_{DB} eventually drives the drain end of the channel into weak inversion and then depletion, where ψ_{sL} becomes practically constant at a value that depends only on $V_{GB} = V_{GB4}$, as seen in Fig. 4.6. Increasing V_{DB} further has little effect on ψ_{sL} . Thus I_{DS1} and I_{DS2} in (4.3.14) also become independent of V_{DB} . It is seen then that, although the source end of the channel can be, say, strongly inverted with a large $|Q'_I|$ and with ψ_{s0} strongly dependent on V_{SB} , the drain end of the channel can be weakly inverted with a small $|Q'_I|$ and with ψ_{sL} practically independent of V_{DB} . However, the current at the two ends (and throughout the channel) is the same. This should not be surprising since, as follows from (4.3.4), the current at any point in the channel is not determined only by the value of Q'_I at that point. This will be discussed further later on.

Consider now a value of V_{DB} corresponding to the saturation part of the curves in Fig. 4.3 or 4.4. If I_{DS} and its components are plotted for that value vs. V_{GB} , we obtain the curves of Fig. 4.8.⁴⁹ To include a large range of currents, a logarithmic vertical axis is used. The regions of inversion across the horizontal axis have been marked according to the definitions of Table 4.1. Here V_{DB} is larger than V_{SB} , so that the more heavily inverted

[†] This assumption is equivalent to assuming that the electron quasi-Fermi levels remain essentially constant over the transition regions; a corresponding assumption was discussed in relation to the *p*n junction in Sec. 1.5.^{16,65}

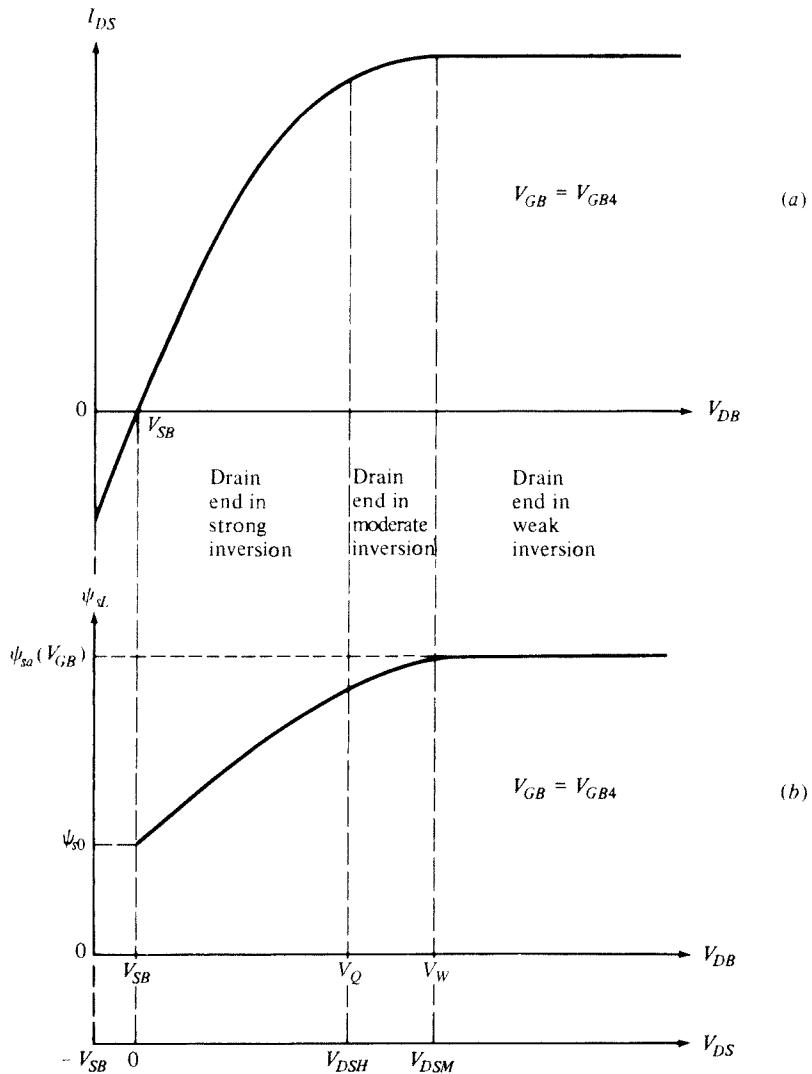
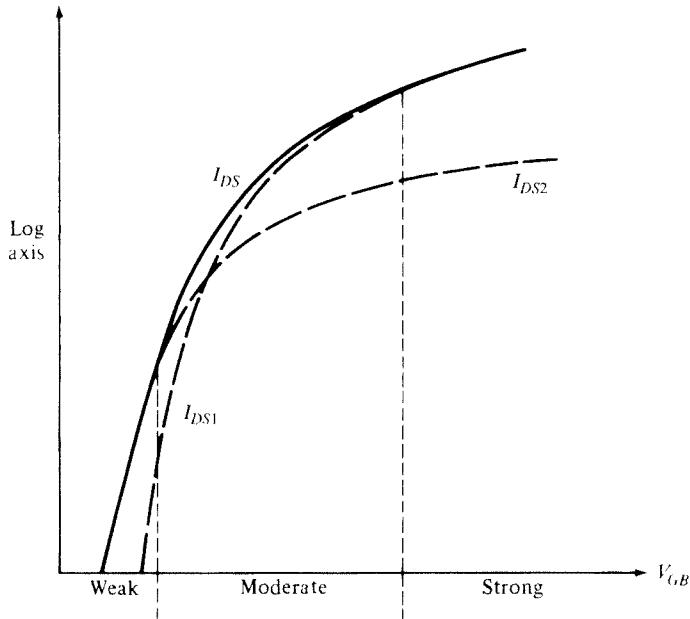


FIGURE 4.7

(a) Drain-source current for a transistor operating in strong inversion vs. drain-body voltage for fixed source-body and gate-body voltages; (b) surface potential at the drain end of the channel, corresponding to (a). A drain-source voltage axis is shown at the bottom of the figure.

channel end is the one next to the source. The region of inversion at that end can be determined as in Chap. 3. We note here that in transistor literature the weak inversion region is often not bounded from below; i.e., everything below moderate inversion is called weak inversion, as long as the current does not become so low as to be masked by leakage currents (Sec. 4.8).

It is seen in Fig. 4.8 that in strong inversion $I_{DS} \approx I_{DS1}$, so the current is mainly due to the presence of drift. In weak inversion, the current is mainly due to the

**FIGURE 4.8**

Drain-source current I_{DS} , its component due to drift I_{DS1} , and its component due to diffusion I_{DS2} vs. gate-body voltage for a large drain-body voltage corresponding to the saturation part of the characteristics in Fig. 4.3a.⁴⁹ A logarithmic axis is used for the current.

presence of diffusion, since $I_{DS} \approx I_{DS2}$. However, in moderate inversion *both* I_{DS1} and I_{DS2} are important; both drift and diffusion play an important role in this region. Similar conclusions can be reached for other V_{DB} values.

In Fig. 4.9, we show a comparison of the preceding model (lines) to full computer solution of the semiconductor equations (dots), for two values of V_{SB} . The simulation is a two-dimensional numerical solution of the Poisson and drift-diffusion equations, allowing for the spreading of the inversion layer below the surface. The parameters used are shown in the caption. The values of W and L used were on purpose chosen very large, to minimize edge effects, which are not supposed to be taken into account in the preceding model (these will be dealt with in Chap. 5). The flat-band voltage and mobility were determined to obtain the best fit of the complete all-region model to the data.

A Graphical Interpretation We now give a graphical interpretation of the preceding results, with the help of Fig. 4.10.[†] In this figure, we show $-Q'_I$ vs. ψ_s , as obtained from (4.3.13), for a given V_{GB} . As can be verified from the way that

[†] Our approach involves the surface potential for the all-region model under discussion. Related interpretations have been given using the external voltages for the strong inversion region,⁶⁶ and the quasi-Fermi potential.⁶⁷

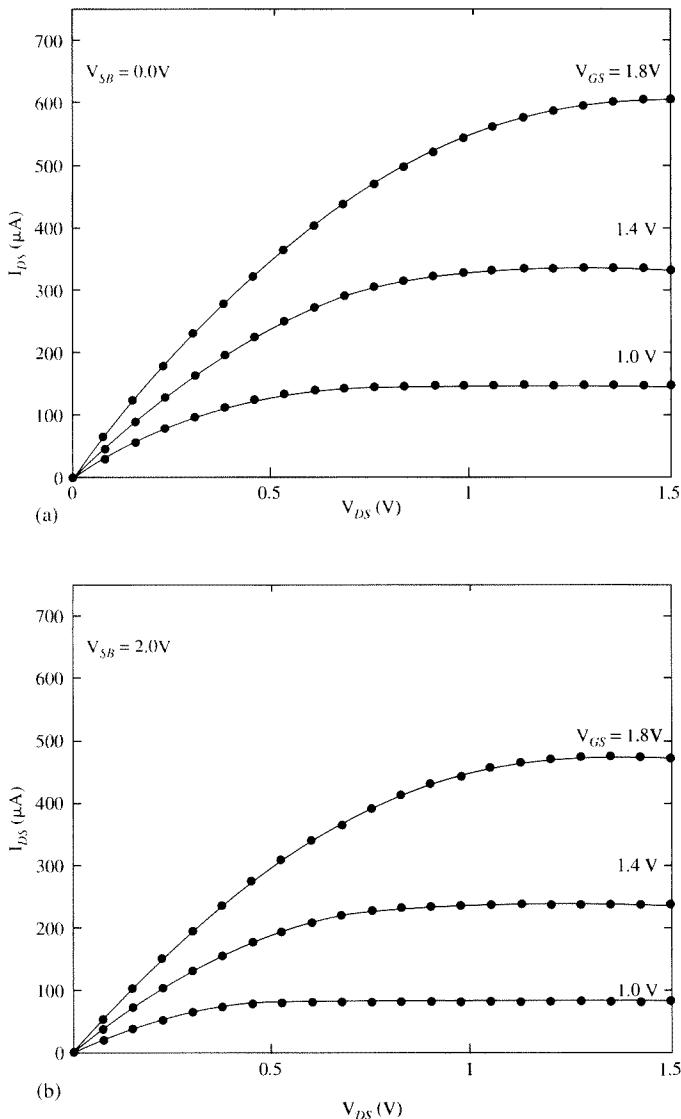


FIGURE 4.9
Comparison of the complete all-region model (lines) to a full computer solution of the semiconductor equations (dots), for two values of V_{SB} . $N_A = 7.0 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2.0 \text{ nm}$, $W = L = 10 \mu\text{m}$.

equation is derived, this curve describes how $-Q'_I$ must behave in order to satisfy charge balance, potential balance, and the gate/oxide and bulk charge-potential relations. The only part of this curve that is relevant to us is that between ψ_{s0} and ψ_{sL} , the values of the surface potential at the source and the drain. To determine these values, we note that $-Q'_I$ vs. ψ_s must also satisfy (3.2.18), which describes the semiconductor physics in the inversion layer, evaluated at the source and the drain. In the

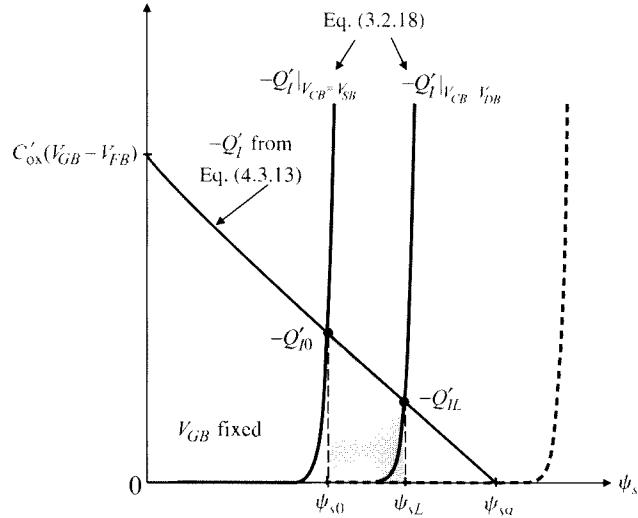


FIGURE 4.10
Graphical interpretation of (4.3.9)
(see text).

figure, we show plots of this equation for $V_{CB} = V_{SB}$ and $V_{CB} = V_{DB}$. Thus the solution at the two ends of the channel is given by the points of intersection, shown by the dots in the figure. These determine the surface potentials at the source and the drain, ψ_{s0} and ψ_{sL} , and the corresponding charges, $-Q'_{l0}$ and $-Q'_{lL}$, as indicated. In between, the negative of the inversion layer charge is represented by the line corresponding to (4.3.13). Thus, the shaded area of the trapezoid is the integral in (4.3.9a). Further, the difference between the two vertical sides of the shaded trapezoid is the difference in (4.3.9b).

Before we continue, let us determine the point where the curve representing (4.3.13) intersects the horizontal axis. This can be found by setting that equation equal to zero and solving for ψ_s . The result is nothing but ψ_{sa} , a quantity we have encountered for the first time in Sec. 2.5; this makes sense, given the definition of ψ_{sa} in that section as the surface potential when the inversion layer charge is negligible.

The graphical construction in Fig. 4.10 facilitates a feel for the behavior of I_{DS1} and I_{DS2} in (4.3.9). The particular operating point assumed in Fig. 4.10 happens to be in the strong inversion nonsaturation region. This is easy to see, as when V_{DB} is raised a little, the corresponding $-Q'_l$ curve will move to the right in the figure; the value of ψ_{sL} will increase, and so will the currents in (4.3.9) (see their graphical interpretation above). Thus $I_{DS} = I_{DS1} + I_{DS2}$ will increase. Eventually, as V_{DB} is raised further, the corresponding $-Q'_l$ plot will move far to the right, as illustrated by the broken-line curve; ψ_{sL} will asymptotically reach the value ψ_{sa} and will be unable to rise further. $-Q'_{lL}$ now becomes essentially zero, and the trapezoid becomes a triangle; I_{DS1} and I_{DS2} stay fixed as V_{DB} is raised further. We have now reached saturation, and the current flattens out at its maximum value, as seen in Fig. 4.7.

Symmetry It is clear from (4.3.7) and (4.3.14) that I_{DS} can be written in the form

$$I_{DS} = \frac{W}{L} [f(\psi_{sL}) - f(\psi_{s0})] \quad (4.3.16)$$

where

$$f(\psi_s) = \mu C'_{ox} \left[(V_{GB} - V_{FB} + \phi_t) \psi_s - \frac{1}{2} \psi_s^2 - \frac{2}{3} \gamma \psi_s^{3/2} + \phi_t \gamma \psi_s^{1/2} \right] \quad (4.3.17)$$

Equation (4.3.16) is in a form that emphasizes the symmetry of the transistor. If the potentials at the source and drain are interchanged, the only difference will be that I_{DS} will change sign. That I_{DS} must be in the form of (4.3.16) can also be deduced directly from (4.3.6). In fact, this form, with $f(\psi_s)$ an appropriate function, will be valid *even* if μ depends on ψ_s in that equation.

The complete all-region model we have presented is often developed in the literature by using the concept of quasi-Fermi potentials (Sec. 4.5 and Appendix F). This results in a more complex derivation^{40,41,43-46} than the one we have presented. It is also possible to calculate the current without making the charge sheet approximation, allowing for the spreading of the inversion layer below the surface, and even allowing for the presence of holes in the depletion region; this was done quite early by Pao and Sah¹⁰ (Appendix F). Although their analysis is recognized to be very general and accurate, it includes the numerical evaluation of a double integral and is, thus, computationally inefficient. It has been shown that the double-integral formulation can be reduced to an equivalent single-integral form,⁴⁷ but numerical integration is still required. The all-region model we have presented is known to be in excellent agreement with these general formulations and does not require numerical integration.

Surface Potential, Charge, and Currents vs. Position For visualizing transistor operation and for calculating certain quantities, it will be useful to relate the surface potential ψ_s to the position along the channel x . We now describe how this can be done in a general way. The expression for $I_{DS} = I_{DS1} + I_{DS2}$ for the model of (4.3.7) and (4.3.14) (and, in fact, also for the models discussed in the next section) is of the form

$$I_{DS} = \frac{W}{L} F(\psi_{sL}, \psi_{s0}) \quad (4.3.18)$$

where F is an appropriate function. The part of the device in Fig. 4.1a between point x and the source can be viewed as a transistor by itself, with point x playing the role of the drain and $\psi_s(x)$ playing the role of the surface potential at the drain end; the channel length of this transistor will be x . Thus, in lieu of (4.3.18), we will have for this transistor

$$I_{DS} = \frac{W}{x} F[\psi_s(x), \psi_{s0}] \quad (4.3.19)$$

where, of course, the current is the same as in the complete device owing to current continuity in the channel. Eliminating I_{DS} between the preceding two equations we obtain

$$\frac{x}{L} = \frac{F[\psi_s(x), \psi_{s0}]}{F(\psi_{sL}, \psi_{s0})} \quad (4.3.20)$$

This equation gives the relation between x and $\psi_s(x)$. The easy way to get results from it is to give values to $\psi_s(x)$ between ψ_{s0} and ψ_{sL} and determine x . In strong inversion, plots obtained in this way have the form shown in Fig. 4.11. In moderate inversion, the variation of ψ_s with x for $V_{DB} > V_{SB}$ is less pronounced. In weak inversion, the two curves would practically coincide because, in that region, the variations of the surface potential along the channel is negligible even when $V_{DB} > V_{SB}$. The variation of charges with position in the channel can now be determined. One can, for example, consider (4.3.13) and (4.3.20) as a parametric representation of Q'_l vs. x . For each value of $\psi_s(x)$ (between ψ_{s0} and ψ_{sL}) given to these equations, a point (x, Q'_l) is obtained. Thus one can plot Q'_l vs. x . For the case illustrated by the upper curve of Fig. 4.11, one finds from (4.3.13) that $|Q'_l|$ as predicted by the all-region model decreases monotonically along the channel as we go from the source toward the drain (see also Fig. 4.10).

One can similarly evaluate $d\psi_s/dx$ and dQ'_l/dx as a function of position x , and substitute them into (4.3.3). It is then found that $I_{\text{drift}}(x)$ decreases with x , whereas $I_{\text{diff}}(x)$ increases.⁵⁰ The values of these two current components at every x are such that their sum is constant and equal to I_{DS} . This is why we have avoided calling I_{DS1} and I_{DS2} “drift” and “diffusion” currents; the drift and diffusion currents *vary* with x . As already mentioned, I_{DS1} and I_{DS2} are just the *average* values of those currents over x .

An Energy Barrier Point of View To complete our general picture, we now return to our energy barrier viewpoint, discussed at the beginning of this section with the help

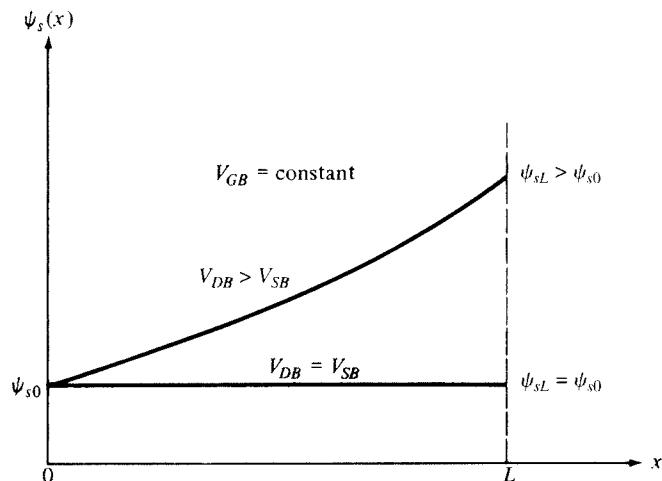


FIGURE 4.11

Surface potential vs. distance from the source for given V_{GB} and V_{SB} , for two values of V_{DB} (strong inversion is assumed).

of Fig. 4.2, and complete the picture given there. As shown in Fig. 4.2, if the drain-source voltage is larger than zero, the potential rises toward the right (curve c'), which corresponds to a lowering of E_c (curve c). The total amount of E_c lowering is qV_{DS} , as shown. Source electrons with enough energy to enter the channel are accelerated by the field toward the drain. As they move toward the right, the difference between their energy and the potential energy E_c increases, which means that their kinetic energy increases; they suffer collisions, and the scattering mechanisms we have mentioned in Sec. 1.3.2 (and which we will discuss in more detail in Sec. 4.11) result in a certain mobility and drift velocity. We have already calculated the current in the channel by considering how the potential and the electron density vary with position. As we have seen, in strong inversion the current in the channel is mainly due to drift. However, note that at the peak of the E_c plot in curve c , we have zero gradient for E_c , and thus for ψ_s , with respect to position; thus the electric field there is zero, and there is no drift at this point. This means that the electrons cross the barrier by diffusion; once they are in the channel, they are subjected to the electric field, and drift takes hold. These points were not part of our description so far, as we limited our attention to points *in the channel*, i.e., to the right of the barrier's peak on curve c of Fig. 4.2.

The preceding discussion suggests a different point of view of the current in a MOS transistor.^{68,69} Thus, the population of electrons that enter the channel is governed by an energy barrier at the source. The role of the gate voltage is to lower that barrier and allow more electrons to enter. To see that this picture is consistent with our detailed analysis, recall that the density of electrons able to cross the barrier and enter the channel is proportional to $\exp(\psi_s/\phi_r)$ (Sec. 3.2). In weak inversion, changes in ψ_s are almost proportional to changes in the gate voltage (Fig. 2.16), and thus the density of electrons crossing the barrier changes nearly exponentially with the gate voltage; this can be linked to the exponential dependence of drain current on gate voltage in weak inversion, apparent in Figs. 4.4 and 4.8. In strong inversion, changes in ψ_s are logarithmically related to changes in the gate voltage (Fig. 2.16); thus, the density of electrons crossing the barrier varies linearly with changes in gate voltage, and this can be linked to the fact that, in nonsaturation, the drain current changes linearly with changes in the gate voltage (see, for example, Fig. 4.9 at small V_{DS} values).⁶⁸ The functional dependence of the drain current on the gate voltage in each region of inversion will be discussed in more detail in the next several sections.

At this point one may wonder how it is possible that the energy barrier lowering viewpoint leads to results consistent with those we have derived by concentrating on the channel, without considering such lowering. The answer is that *it had better!* Both energy barrier lowering and channel phenomena are simultaneously present in a transistor, so they should be expected to produce consistent results.

Unconventional Operation of the MOSFET The above discussion, and in fact this entire book, deals with MOSFETs in which the entire gate is at the potential applied to its external terminal, and a similar statement holds for the body. There have been attempts to apply a potential along the gate,⁷⁰ and along the gate and the body,⁷¹ to achieve a variety of aims. The operation of a six-terminal MOS transistor as an extremely linear voltage-controlled resistor has been described elsewhere.⁷¹

4.4 Simplified All-Region Models

4.4.1 Linearizing the Depletion Region Charge

The “complete” all-region model of (4.3.7), (4.3.14), and (4.3.15) is very accurate, but is too complicated for some applications; this becomes apparent when the same formulation is extended to charge evaluation, which is needed in transient response calculations (Chap. 6). Part of this complexity is due to the 3/2 power and 1/2 power terms in (4.3.14). It is clear from the development presented that the origin of these terms is the square root in the expression for Q'_B in (4.3.12). Figure 4.12 shows a plot of $-Q'_B/C'_{\text{ox}}$ vs. ψ_s from that equation. The surface potential in the channel will be between ψ_{s0} and ψ_{sL} (see our discussion on surface potential vs. position, and Fig. 4.11). Since the slope of the plot in Fig. 4.12 does not vary much, it is reasonable to attempt to approximate $-Q'_B/C'_{\text{ox}}$ by the first two terms of its Taylor expansion around a convenient expansion point $\psi_s = \psi_{se}$ (choices for ψ_{se} will be discussed shortly). This gives, from (4.3.12):

$$-\frac{Q'_B}{C'_{\text{ox}}} = \gamma\sqrt{\psi_{se}} + \frac{\gamma}{2\sqrt{\psi_{se}}}(\psi_s - \psi_{se}) \quad (4.4.1)$$

where $\gamma/2\sqrt{\psi_{se}}$ is the slope of the curve in Fig. 4.12. Let us define, for later use, a quantity α as follows:

$$\alpha = 1 + \frac{\gamma}{2\sqrt{\psi_{se}}} \quad (4.4.2)$$

In terms of this quantity, (4.4.1) can be written as

$$-\frac{Q'_B}{C'_{\text{ox}}} = \gamma\sqrt{\psi_{se}} + (\alpha - 1)(\psi_s - \psi_{se}) \quad (4.4.3)$$

Using Q'_B from this equation, the last terms in (4.3.11) yield simple results; specifically, we have

$$\int_{\psi_{s0}}^{\psi_{sL}} Q'_B d\psi_s = -C'_{\text{ox}} \left[\gamma\sqrt{\psi_{se}} + (1 - \alpha) \left(\psi_{se} - \frac{\psi_{s0} + \psi_{sL}}{2} \right) \right] (\psi_{sL} - \psi_{s0}) \quad (4.4.4a)$$

$$Q'_B(\psi_{sL}) - Q'_B(\psi_{s0}) = C'_{\text{ox}}(1 - \alpha)(\psi_{sL} - \psi_{s0}) \quad (4.4.4b)$$

Using these in (4.3.11), we get

$$I_{DS1} = \frac{W}{L} \mu C'_{\text{ox}} \left[\left(V_{GB} - V_{FB} - \psi_{se} - \gamma\sqrt{\psi_{se}} \right) + \alpha \left(\psi_{se} - \frac{\psi_{s0} + \psi_{sL}}{2} \right) \right] (\psi_{sL} - \psi_{s0}) \quad (4.4.5a)$$

$$I_{DS2} = \frac{W}{L} \mu C'_{\text{ox}} \alpha \phi_t (\psi_{sL} - \psi_{s0}) \quad (4.4.5b)$$

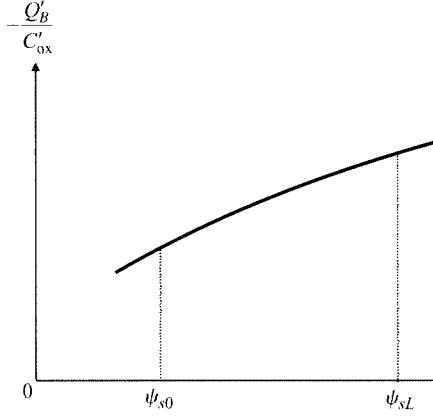


FIGURE 4.12
The quantity $-Q'_B/C'_ox$ vs. surface potential.

The potentials ψ_{s0} and ψ_{sL} in these expressions can be found from (4.3.15). Although it is tempting to further simplify (4.4.5a) by taking advantage of the special form of (4.4.2), we will not do so as the value of α will have to be modified in some cases, as we will see.

For future reference, we also give the expression for the inversion layer charge that corresponds to the preceding simplification. Using Q'_B from (4.4.3) in (4.3.10), we have

$$Q'_I = -C'_{ox}(V_{GB} - V_{FB} - \psi_s) - Q'_B \quad (4.4.6a)$$

$$= -C'_{ox} \left[V_{GB} - V_{FB} - \psi_{se} - \gamma \sqrt{\psi_{se}} - \alpha(\psi_s - \psi_{se}) \right] \quad (4.4.6b)$$

Our discussion so far has been in terms of an arbitrary expansion point, $\psi_s = \psi_{se}$. Several choices for ψ_{se} have been proposed in the literature. These are considered in the following sections.

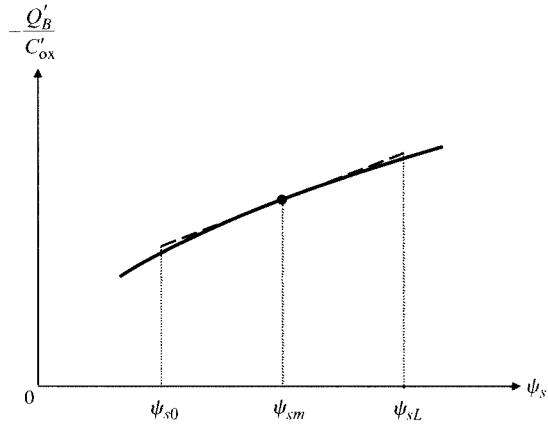
4.4.2 Body-Referenced Simplified All-Region Models

Consider (4.4.4a) and (4.4.5a). It can be seen that a simplifying cancellation occurs if we choose as ψ_{se} the arithmetic mean of the potential in the channel:^{67,72-74}

$$\psi_{se} = \psi_{sm} = \frac{\psi_{s0} + \psi_{sL}}{2} \quad (4.4.7)$$

This corresponds to approximating $-Q'_B/C'_ox$ as shown in Fig. 4.13. Note that ψ_{sm} is the midpoint of the *potential*, not the physical midpoint of the channel (the two coincide only when $V_{DS} = 0$). The resulting value of α is, from (4.4.2):

$$\alpha = \alpha_m = 1 + \frac{\gamma}{2\sqrt{\psi_{sm}}} \quad (4.4.8)$$

**FIGURE 4.13**

Solid line: $-\frac{Q'_B}{C'_\text{ox}}$ vs. surface potential. Broken line: approximation (4.4.9) with (4.4.7) and (4.4.8) (symmetric linearization).

The corresponding equation for Q'_B becomes, using the preceding equation in (4.4.3):

$$Q'_B = -C'_\text{ox} \left[\gamma \sqrt{\psi_{sm}} + (\alpha_m - 1)(\psi_s - \psi_{sm}) \right] \quad (4.4.9)$$

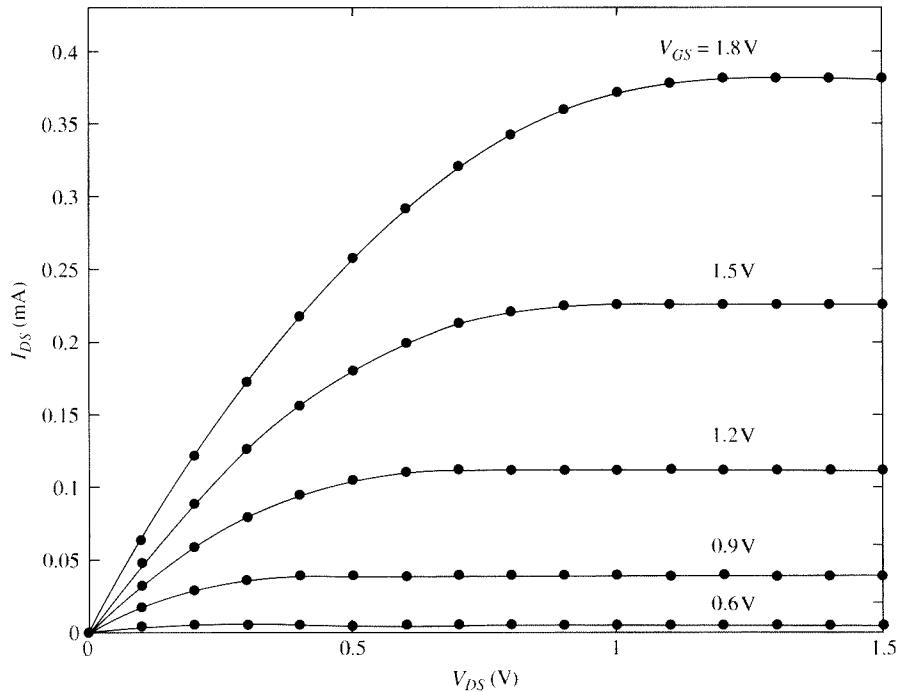
and Q'_I is still given by (4.4.6). Equation (4.4.5) gives the corresponding current components due to drift and due to diffusion as follows:

$$I_{DS1} = \frac{W}{L} \mu C'_\text{ox} (V_{GB} - V_{FB} - \psi_{sm} - \gamma \sqrt{\psi_{sm}})(\psi_{sL} - \psi_{s0}) \quad (4.4.10a)$$

$$I_{DS2} = \frac{W}{L} \mu C'_\text{ox} \alpha_m \phi_i (\psi_{sL} - \psi_{s0}) \quad (4.4.10b)$$

The inherent source-drain symmetry of this model is apparent. As before, ψ_{s0} and ψ_{sL} can be found from (4.3.15). Note that ψ_{sm} and α_m change when ψ_{sL} changes, as seen from (4.4.7) and (4.4.8).

To get some intuition for the current due to drift, (4.4.10a), consider for simplicity a case where ψ_{s0} and ψ_{sL} vary in opposite directions by equal amounts, so that their average value in (4.4.7) is constant. As seen from (4.4.6) and (4.4.7), the quantity $C'_\text{ox}(V_{GB} - V_{FB} - \psi_{sm} - \gamma \sqrt{\psi_{sm}})$ is the magnitude of the inversion charge (per unit area), evaluated at the average potential between source and drain, and for the case we are considering it is fixed for a fixed V_{GB} . Thus, (4.4.10a) can be interpreted as follows: the “vertical” field caused by the applied voltage V_{GB} induces a conducting channel to form underneath the gate, with an “effective” charge as above, and the “lateral” field caused by $\psi_{sL} - \psi_{s0}$ causes current to flow through this conducting channel. If ψ_{s0} and ψ_{sL} vary arbitrarily, this picture gets complicated, as now the above effective channel charge will vary together with those potentials.

**FIGURE 4.14**

Comparison of drain-source current between the general all-region model (lines) and the body-referenced simplified general model (dots), where $V_{FB} = -0.8$ V, $N_A = 5 \times 10^{17}$ cm $^{-3}$, $t_{ox} = 2.5$ nm, $\mu = 400$ cm 2 /(V · s), $W = L = 10$ μ m, $V_{SB} = 0$ V.

Although square-root terms are still present in (4.4.10), the expressions are analytically and computationally simpler than (4.3.14) and numerically more stable to evaluate; yet, they are essentially indistinguishable in terms of accuracy. Figure 4.14 compares $I_{DS}(V_{DS})$ characteristics from the above simplified model (lines) to a detailed solution of the semiconductor equations (dots). The agreement is practically as good as for the complete all-region model of Sec. 4.3. The preceding symmetric linearization^{72–74} is the basis of the PSP model.⁶⁸

Ensuring Continuity outside Inversion The simplified all-region model presented in this section and the complete model of Sec. 4.3 give the drain current behavior of MOSFETs in all regions of inversion. However, the related charge expressions used in these models are not applicable in depletion and accumulation. It is desirable, both to maintain a single physical formulation and to maintain continuity of analysis for all biases, to have a single set of equations valid irrespective of the specific region of operation. This is important for charge modeling for dynamic operation, as is discussed in Chap. 6.

MOSFET operation is controlled by a balance of charges in the gate and in the silicon; the latter charge comprises inversion, depletion, and accumulation charges.

In inversion operation, the bulk charge is primarily due to the depletion region formed under the gate, but in accumulation there is no inversion charge and the bulk charge comes primarily from the induced accumulation of majority carriers (holes in the *p*-type substrate of an *n*MOS device). To capture this with a single bulk charge equation, one can replace Q'_B in (4.3.12) by the right-hand side of (2.5.1):

$$Q'_B = -\text{sgn}(\psi_s)\sqrt{2q\epsilon_s N_A} \sqrt{\phi_t e^{-\psi_s/\phi_t} - \phi_t + \psi_s} \quad (4.4.11)$$

and rework the development of the model equations accordingly. Needless to say, this yields more complicated expressions than the ones we have presented. The resulting accuracy for the drain current is not affected, because in inversion the preceding expression essentially reduces to (4.3.12); however, in depletion and accumulation, where the drain current is negligible, (4.3.12) fails whereas (4.4.11) continues to be valid. The symmetric linearization^{67,72} procedure of (4.4.8) and (4.4.9) is a simplified version (valid for $\psi_s \geq 3\phi_t$) of actual equations used in the PSP model.^{73,74}

4.4.3 Source-Referenced Simplified All-Region Models

A different point of expansion used sometimes is:^{75,76}

$$\psi_{se} = \psi_{s0} \quad (4.4.12)$$

This corresponds to approximating $-Q'_B/C'_{ox}$ [see (4.4.1)–(4.4.3)] as shown by line *a* in Fig. 4.15. This choice of ψ_{se} was originally suggested for a strong inversion model in Ref. 24. Using the preceding value in (4.4.5), we obtain

$$I_{DS1} = \frac{W}{L} \mu C'_{ox} \left[\left(V_{GB} - V_{FB} - \psi_{s0} - \gamma \sqrt{\psi_{s0}} \right) (\psi_{sL} - \psi_{s0}) - \frac{\alpha}{2} (\psi_{sL} - \psi_{s0})^2 \right]$$

(4.4.13a)

$$I_{DS2} = \frac{W}{L} \mu C'_{ox} \phi_t \alpha (\psi_{sL} - \psi_{s0})$$

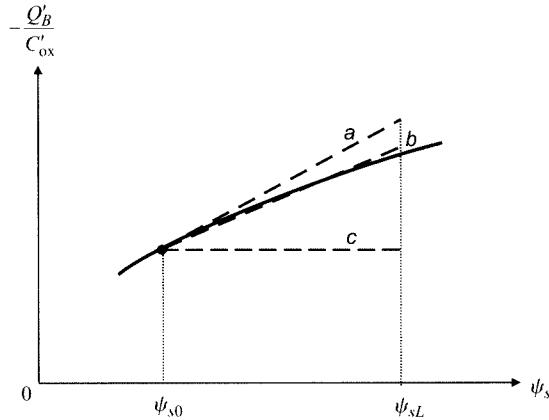
(4.4.13b)

where the nominal value to be used for α , which we will denote by α_1 , can be found by using (4.4.12) in (4.4.2):

$$\alpha = \alpha_1 = 1 + \frac{\gamma}{2\sqrt{\psi_{s0}}} \quad (4.4.14)$$

As seen in the current equations, the drain surface potential appears only in a *difference* with the source surface potential. This is typical of *source-referenced* models and will be seen to lead to a popular strong-inversion model in Sec. 4.7. On the other hand, the inherent symmetry of the ideal MOS transistor is not apparent in the preceding equations. Certain refinements are needed before this approach can lead into a fully developed model.[†]

[†] As seen from line *a* in Fig. 4.15, the choice in (4.4.12) results in an expansion that is most accurate near the source. Near the drain, Q'_B and thus Q'_t will be somewhat in error. In the complete all-region model, as V_{DB} was

**FIGURE 4.15**

Solid line: $-Q'_B/C'_\text{ox}$ vs. surface potential. Line *a*: approximation for (4.4.3) with (4.4.12) and (4.4.14) (linearization with respect to the source). Line *b*: improved approximation obtained by slightly lowering the value of α . Line *c*: approximation for $\alpha = 1$.

It has been seen that (4.4.3), with α as in (4.4.14), results in the approximation shown by the upper broken line in Fig. 4.15. One can improve the overall accuracy of Q'_B vs. ψ_s , by replacing α above by a slightly smaller value. This corresponds to the type of approximation shown by line *b* in Fig. 4.15. We will see examples of this practice in Sec. 4.7. The slope of line *b* can be made a function of $\psi_{sL} - \psi_{s0}$ for even better accuracy.⁷⁷ For future use, Fig. 4.15 also shows the poor approximation resulting from assuming $\alpha = 1$ in (4.4.3) (line *c*).

4.4.4 Charge Formulation of Simplified All-Region Models

In Sec. 4.4.1, we have introduced approximations for $Q'_B(\psi_s)$ in the general form of (4.4.3). In these, Q'_B varies linearly with ψ_s . For such approximations, the corresponding inversion charge Q'_I for a fixed V_{GB} also varies linearly with ψ_s , as can be seen from (4.4.6b). This approximation for Q'_I is more accurate than the one for Q'_B , as Q'_I already contained a dominant linear term to begin with [see (4.4.6a)], which is maintained when Q'_B is approximated. The nearly linear variation of Q'_I with ψ_s for a fixed V_{GB} can be also be observed in Fig. 3.16. Differentiating (4.4.6b) with respect to ψ_s gives

$$\frac{\partial Q'_I}{\partial \psi_s} = \alpha C'_\text{ox} \quad (4.4.15)$$

raised, ψ_{sL} approached ψ_{sa} and Q'_{IL} approached zero, leading to “saturation” of the I_{DS} curves as in Fig. 4.7a. Here, though, because of the preceding small error, Q'_{IL} approaches zero at a value of ψ_{sL} slightly different from ψ_{sa} . This should be taken into account.⁷⁵ It turns out that this is not a serious problem, as for real devices the saturation value of ψ_{sL} has to be modified anyway⁷⁶ to take into account other effects such as velocity saturation (Chap. 5).

This is a key relation for expressing the current in terms of the inversion layer charge.⁷⁸ Indeed, using it in (4.3.9a) to perform a change of the integration variable from ψ_s to Q'_I , we obtain

$$I_{DS1} = \frac{W}{L} \mu \int_{Q'_{I0}}^{Q'_{IL}} (-Q'_I) \frac{1}{\alpha C'_{ox}} dQ'_I \quad (4.4.16)$$

where Q'_{I0} and Q'_{IL} are the values of Q'_I at the source and drain ends, respectively. This gives the drain current component due to drift:

$$I_{DS1} = \frac{W}{L} \frac{\mu}{2\alpha C'_{ox}} (Q'^2_{I0} - Q'^2_{IL}) \quad (4.4.17a)$$

whereas the component due to diffusion is still given by (4.3.9b), repeated here:

$$I_{DS2} = \frac{W}{L} \mu \phi_i (Q'_{IL} - Q'_{I0}) \quad (4.4.17b)$$

Adding the two current components gives a result similar to ones derived by various means elsewhere.^{76,78-82} The model in the previous equations is completely equivalent to the surface potential-based simplified model of Sec. 4.4.1; in fact, using in it Q'_{I0} and Q'_{IL} as obtained from (4.4.6), we obtain identically (4.4.5). Thus, the only difference is that rather than being formulated in terms of the surface potentials at the source and drain, the expressions in (4.4.17) are formulated in terms of the inversion charge densities at the source and drain, Q'_{I0} and Q'_{IL} . If the model of (4.4.17) is used with α and Q'_I as in Sec. 4.4.2, it becomes the body-referenced simplified all-region model described in that section; if used with α and Q'_I as in Sec. 4.4.3, it becomes the source-referenced model described there.

The linearization usually used in conjunction with the model of (4.4.17) is shown in Fig. 4.16. The depletion region charge is expanded around the point⁸¹⁻⁸⁶

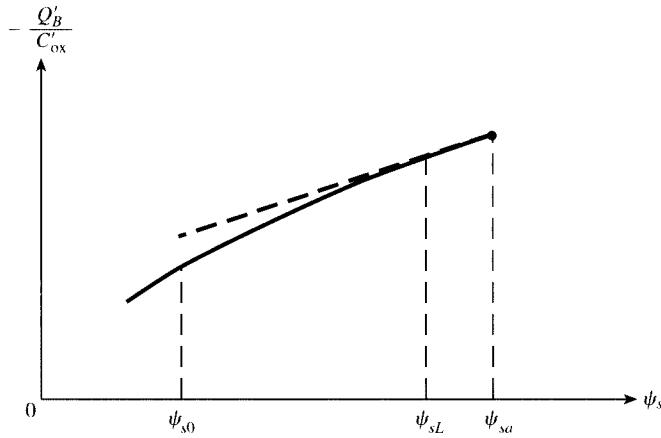
$$\psi_{se} = \psi_{sa} \quad (4.4.18)$$

We remind the reader that ψ_{sa} is the maximum possible value of the surface potential for a given V_{GB} (see Fig. 4.6). The above choice implies, from (4.4.2) and (3.2.26):

$$\alpha = n = 1 + \frac{\gamma}{2\sqrt{\psi_{sa}}} \quad (4.4.19)$$

Various modifications of the preceding linearization point are in use. This type of model is often called an “inversion charge based” model, “charge-controlled” model,^{78,85-87} or “unified charge control model” (UCCM),^{79,80,88} although this last name is often used in conjunction with a particular way for evaluating the charges involved, summarized in the following paragraphs. Significant further development of the charge-based approach has led to the well-known EKV^{84,85} and ACM^{81,86} models.

The charges in (4.4.17) can be evaluated by finding the surface potentials at the source and the drain from (4.3.15), and using them in (4.3.13) to obtain Q'_{I0} and Q'_{IL} .

**FIGURE 4.16**

Solid line: $-Q'_B/C'_\text{ox}$ vs. surface potential. Broken line: approximation (4.4.1) with (4.4.18).

However, the intermediate step of finding the surface potentials can be bypassed as follows. One can solve (3.2.20) for ψ_s in terms of Q'_I , and substitute the result into (3.2.18), to eliminate ψ_s . This gives an implicit equation in Q'_I . This equation can be used to find Q'_{I0} and Q'_{IL} by using in it $V_{CB} = V_{SB}$ and $V_{CB} = V_{DB}$, respectively, and solving numerically. This procedure gives identical results to the solution involving the surface potentials.

Because the implicit equations just described are complicated, simplifications are used.^{79–83,85,88} One can take advantage of the fact that, at the point of linearization in (4.4.18), the inversion layer charge can be assumed to be zero (see the definition of ψ_{sa} in Sec. 2.5). This, together with (4.4.15) and (4.4.19), leads to the following simplified expression for Q'_I :

$$Q'_I = nC'_\text{ox}(\psi_s - \psi_{sa}) \quad (4.4.20)$$

If this expression is used in (3.2.18) to eliminate ψ_s , and further simplifications are made, the following implicit equation results:^{83,85}

$$\frac{-Q'_I}{n\phi_t C'_\text{ox}} + \ln \frac{-Q'_I}{2n\phi_t C'_\text{ox}} = \frac{V_p - V_{CB}}{\phi_t} \quad (4.4.21)$$

where V_p is the pinchoff voltage, discussed in Sec. 3.5.2 and given by (3.5.6).[†] As before, one can find Q'_{I0} and Q'_{IL} from the equation given above by using in it $V_{CB} = V_{SB}$ and $V_{CB} = V_{DB}$, respectively, and solving numerically. The results can be

[†] The value of the parameter ϕ_0 in (3.5.6) is sometimes adjusted for best fit in charge-based modeling.

used in (4.4.17). A similar approach using a somewhat different set of simplifications^{81–83,85,88} results in (4.4.21) again, but with 2 replaced by e .

In weak inversion, the first term in (4.4.21) becomes negligible compared with the second term; thus, the equation gives Q'_L explicitly as an exponential. In strong inversion, the first term becomes large and the second can be neglected; Q'_L now is given explicitly by a linear equation. Both of these limit cases agree with independent development of V_p -based models, which are summarized in Appendix G. In moderate inversion, both terms in 4.4.21 play a significant role.

Forward and Reverse Components By adding the current components in (4.4.17) we obtain

$$I_{DS} = I_{DS1} + I_{DS2} \quad (4.4.22a)$$

$$= \frac{W}{L} \mu \left(\frac{1}{2\alpha C'_{ox}} \left(Q'^2_{l0} - Q'^2_{lL} \right) + \phi_t \left(Q'_{lL} - Q'_{l0} \right) \right) \quad (4.4.22b)$$

$$= \frac{W}{L} \mu \left(\frac{Q'^2_{l0}}{2\alpha C'_{ox}} - \phi_t Q'_{l0} \right) - \frac{W}{L} \mu \left(\frac{Q'^2_{lL}}{2\alpha C'_{ox}} - \phi_t Q'_{lL} \right) \quad (4.4.22c)$$

This equation can be written in the form:^{81–84,89,90}

$$I_{DS} = I_F - I_R \quad (4.4.23)$$

where

$$I_F = \frac{W}{L} \mu \left(\frac{Q'^2_{l0}}{2\alpha C'_{ox}} - \phi_t Q'_{l0} \right) \quad (4.4.24a)$$

$$I_R = \frac{W}{L} \mu \left(\frac{Q'^2_{lL}}{2\alpha C'_{ox}} - \phi_t Q'_{lL} \right) \quad (4.4.24b)$$

Consider operation in strong inversion, with $V_{SB} = V_{SB}$. Clearly the inversion level is then identical at the source and drain, hence $Q'_{lL} = Q'_{l0}$; therefore $I_F = I_R$ and, from (4.2.23), $I_{DS} = 0$. If the source voltage is held fixed and the drain voltage is raised, the magnitude of the inversion charge density at the drain drops (see Fig. 3.16); therefore, I_R decreases below I_F , and $I_{DS} > 0$. A positive current thus flows from the drain to the source. If the drain voltage continues to increase, the drain end of the channel will drop out of strong inversion and Q'_{lL} will asymptotically approach zero (Fig. 3.16). Therefore, the drain-source current asymptotically approaches

$$I_{DS,\text{saturation}} = I_F \quad (4.4.25)$$

The inversion charge-based model therefore naturally captures the saturation behavior seen in Figs. 4.3 and 4.4. Note that the value of the drain current in saturation is determined by Q'_{l0} , the inversion charge density at the source, which depends only on V_{GB} and V_{SB} , and *not* on V_{DB} . This is an important point: *in (forward) saturation, the*

saturation drain current is determined only by the level of inversion at the source. It does not depend on the voltage at the drain. The same behavior is exhibited by the other all-region models we have discussed in the previous sections. We note that the previous conclusion is valid for *long-channel* devices, which are the only devices considered in this chapter; short-channel devices, which will be covered in Chap. 5, deviate from this behavior.

If the source voltage is raised rather than the drain voltage, I_F , in (4.4.23) will decrease below I_R , and $I_{DS} < 0$. A positive current now flows from the source to the drain. If the source voltage is raised high enough, Q'_{l0} will approach zero and, therefore, so will I_F . We can interpret this as *reverse saturation* current, where

$$I_{DS,\text{rev.saturation}} = -I_R \quad (4.4.26)$$

We thus see then that the drain current in (4.4.23) can be interpreted as the sum of a *forward* component I_F and a *reverse* component I_R . The first of these components depends only on V_{GB} and V_{SB} ; the second depends only on V_{GB} and V_{DB} .[†] A similar interpretation can be developed for the models of Secs. 4.3 and 4.4.2.

4.5 Models Based on Quasi-Fermi Potentials

All of the previous models are based on (4.3.4), which expresses the total channel current as a sum of a drift component and a diffusion component. It can be shown that both of these components can be combined into a single term, proportional to the gradient of V_{CB} with position, where V_{CB} is the quantity discussed in Chap. 3, corresponding to the quasi-Fermi level split. Specifically, it is shown in Appendix F that we can write[‡]

$$I_{DS} = \mu W(-Q'_l) \frac{dV_{CB}}{dx} \quad (4.5.1)$$

[†] Similarly, other quantities of interest in the transistor, such as charges and small-signal parameters, can also be expressed in terms of these currents.^{83,84,89–91} For example, one can express⁹¹ Q'_{l0} and Q'_{h0} in terms of I_F and I_R , respectively, by solving (4.4.24). One thus ends up with a model in which all quantities of interest, instead of being functions of terminal voltages, are functions of I_F and I_R . Because these are, respectively, the forward and reverse saturation currents, they are easily measurable and are the currents that are often used to bias transistors in practical circuit design. In this approach, the level of inversion at the source and drain channel end is judged by comparing I_F and I_R , respectively, to a specific current value,^{83,84,89,90} given by $I_Z = (W/L)\mu C'_{ox}(2n\phi_i^2)$. Currents less than 1/10 of this value suggest weak inversion operation, whereas currents larger than 10 times this value suggest strong inversion operation.⁹² A different specific current value, equal to 1/4 of the above, is used elsewhere.^{86,91}

[‡] A rigorous way of proving this result requires integrating the inversion layer charges over depth below the surface, as is done in Appendix F. A heuristic argument can be given to make this result plausible for the case of the all-region model, as follows. Recall that the volume density of electrons at the surface is proportional to $\exp[(\psi_s - V_{CB})/\phi_i]$, according to (3.2.2). Our assumption of an inversion charge sheet of negligible thickness, adjacent to the surface, is consistent with assuming that the inversion layer *charge per unit area*, Q'_l , will be proportional to the preceding exponential. Thus, if we differentiate Q'_l with respect to x , we will have $dQ'_l/dx = (Q'_l/\phi_i)(d\psi_s/dx - dV_{CB}/dx)$. Using this in (4.3.4) results in (4.5.1). Again, this is not meant to replace the rigorous proof of Appendix F.

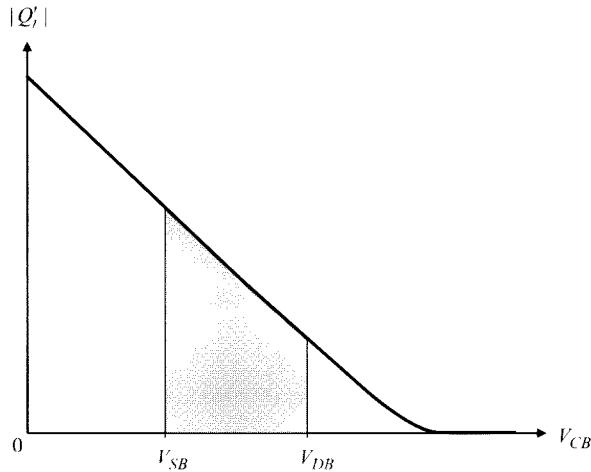


FIGURE 4.17
Graphical interpretation of (4.5.2).

Integrating this equation over the length of the channel, assuming the mobility is constant, and recognizing $\int_0^L I_{DS} dx$ as $I_{DS}L$, we obtain

$$I_{DS} = \frac{W}{L} \mu \int_{V_{SB}}^{V_{DB}} (-Q'_I) dV_{CB} \quad (4.5.2)$$

Q'_I has been plotted as a function of V_{CB} in Fig. 3.16, repeated here as Fig. 4.17. The preceding result shows that the current is proportional to the shaded area in the figure; the particular values chosen for V_{SB} and V_{DB} are on the strong-inversion part of the curve, and thus the transistor is in the strong-inversion, nonsaturation region. As V_{DB} is raised, the shaded area (and thus the current) increases and it eventually saturates asymptotically, as expected. This is the same feature as for the surface potential-based models, only here no surface potential is explicitly involved.

The preceding formulation is quite general, in that it is also valid without making the assumption that the inversion layer is a charge sheet of negligible thickness.¹⁰ Unfortunately, no simple expression is known for Q'_I as a function of V_{CB} , valid in all regions of inversion and at the same explicitly integrable. The above integral is sometimes used to derive various versions of the all-region model,⁴³⁻⁴⁷ as well as charge-based models.^{78.81.85.86.93} Finally, it is used to derive the Pao-Sah model,¹⁰ which is rigorous but computationally inefficient; this model is summarized in Appendix F.

Simple expressions of Q'_I as a function of V_{CB} are possible in particular regions, and when used in (4.5.2) result in well-known models. In strong inversion non-saturation, a very simple interpretation is possible: V_{CB} can be viewed as the effective reverse bias between the inversion layer and the body, as will be discussed in Sec. 4.7.1. Notice, though, that such an interpretation is *not* possible in moderate and weak inversion.

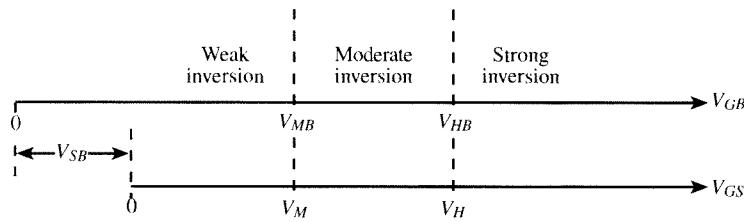


FIGURE 4.18

Definitions of regions of inversion in terms of gate voltage, for a given V_{SB} (assuming that $V_{DB} \geq V_{SB}$).

4.6 Regions of Inversion in Terms of Terminal Voltages

It was seen in Sec. 4.2 that the transistor is said to operate in a certain region of inversion, if its most heavily inverted channel end is in that region of inversion. On the basis of this definition, the regions of inversion have been defined as in Table 4.1. We now define these regions in terms of externally applied voltages.

Definitions in Terms of Gate Voltage Let us assume arbitrarily that the drain end is no more heavily inverted than the source end. This is equivalent to assuming[†]

$$V_{DB} \geq V_{SB} \quad (4.6.1)$$

This assumption can be written in terms of V_{DS} as follows:

$$V_{DS} \geq 0 \quad (4.6.2)$$

Then we can give the condition for operation in a certain inversion region, in terms of the inversion level at the *source* end of the channel. We can use appropriate conditions developed in Chap. 3 for the structure of Fig. 3.2c or d, simply by replacing terminal C by the source terminal S. Thus, from Fig. 3.6 or Table 3.2 (second and third rows), the regions can be defined in terms of V_{GB} or V_{GS} as shown in Fig. 4.18 (in drawing this figure, we have arbitrarily assumed that V_M is greater than zero). We remind the reader that the quantities marked on the V_{GS} axis are given by (Table 3.1):

$$V_M = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{SB}} \quad (4.6.3)$$

[†] We make this convenient assumption just to be able to draw some figures and give some equations in a specific way. This does not really restrict the generality of our discussion. If $V_{DB} < V_{SB}$ ($V_{DS} < 0$), the role of the source and drain are interchanged; then simply replace V_{DB} by V_{SB} and vice versa, and V_{DS} by V_{SD} , in the equations that follow. See Prob. 4.7 for other ways to define regions of operation.

$$V_H = V_M + V_Z \quad (4.6.4)$$

with the value of V_Z being several tenths of a volt (0.5 to 0.6 V at room temperature for practical devices; V_Z decreases slightly with increasing V_{SB}). The corresponding V_{GB} limits can be found by simply adding V_{SB} to these values, as indicated in Fig. 4.18.

Recall that in Chaps. 2 and 3, a *lower* limit for the weak inversion region was also defined at $V_{GB} = V_{LB}$. However, when dealing with a complete transistor, this lower limit is rather academic, for two reasons. (1) Nothing really special happens at that limit. As seen in Fig. 3.6, the various plots are smooth, and their shape does not change drastically below $V_{GB} = V_{LB}$. For this reason, some of the equations we will develop in this section can hold even below that limit. (2) In a real MOS transistor, the observable (and usable) drain current is the sum of the real drain current and a leakage current (which can include the leakage of the drain-substrate n^+p junction as well as other leakage currents discussed in Chap. 5). Thus, a “pragmatic” lower limit for weak-inversion operation can be taken to be at a point where the leakage current can be neglected, requiring

$$I_D \gg I_l \quad (4.6.5)$$

where I_l is the leakage current.

Definitions in Terms of Source and Drain Voltages For a given V_{GB} , the surface potential at the drain (or source) end of the channel can be plotted vs. the drain (or source) voltage with respect to the body, as shown in Fig. 4.6. It is clear that one can determine the level of inversion at the source and drain ends of the channel by comparing each of V_{SB} and V_{DB} to the quantities V_Q and V_W , which were discussed in Sec. 3.5 and given in the last row of Table 3.1. For convenience, we repeat the values of these quantities below:

$$V_Q = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB} - V_Z} \right)^2 - 2\phi_F \quad (4.6.6)$$

$$V_W = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 - 2\phi_F \quad (4.6.7)$$

where, in (4.6.6), V_Z is the same quantity discussed after (4.6.4).

Thus, for example, it follows from Table 4.1 that if the conditions at the source and drain correspond respectively to points 3 and 2 in Fig. 4.6, the transistor operates in weak inversion, and the same will be true for points 3 and 1. For points, respectively, 4 and 2 or 4 and 1, it operates in moderate inversion. The transistor operates in strong inversion for points 5 and 4, or 5 and 3, or 5 and 1, etc. Note that if point 1 corresponds to the source and point 5 to the drain, the transistor is *still* said to operate in strong inversion, according to Table 4.1. However, now the most heavily inverted

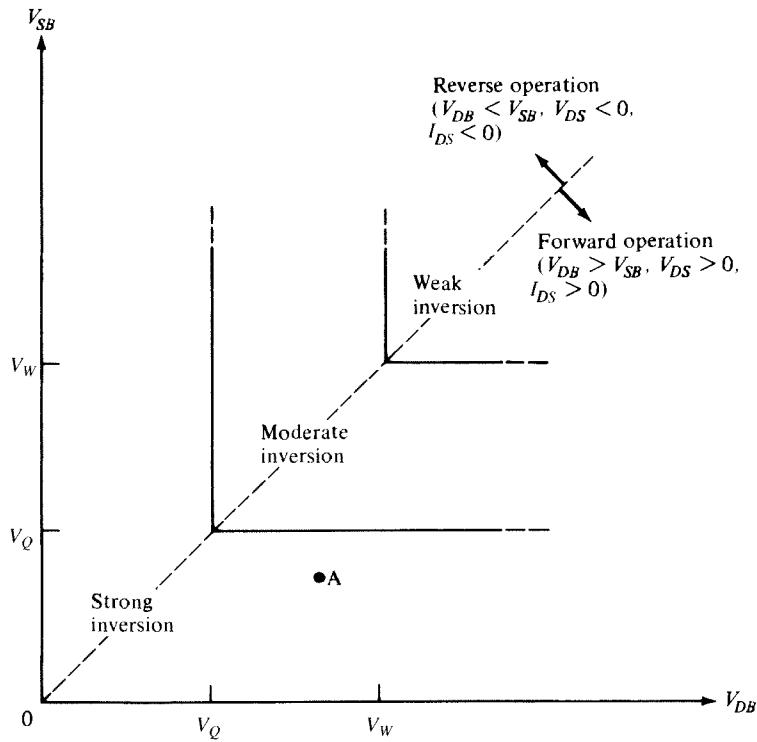


FIGURE 4.19
Definition of regions of operation for a MOS transistor in terms of source-substrate and drain-substrate voltages for a given V_{GB} .

channel end is the one next to the drain; here $V_{DB} < V_{SB}$ ($V_{DS} < 0$), which results in $I_{DS} < 0$. This sort of operation can be termed as *reverse*, whereas *forward* operation can refer to the opposite case, for which most of our equations have been written.

Using these observations, it is easy to see that a general way to define the regions of inversion is as shown in Fig. 4.19. As an example, a transistor with $V_{SB} < V_Q$ and $V_Q < V_{DB} < V_W$ can correspond to, say, point A. The transistor is then in (forward) strong inversion.

Note that in accordance with the preceding discussion, we prefer not to specify a limit between weak inversion and depletion. Rather, we take a transistor to operate in weak inversion if it corresponds to a point in the upper right-hand corner in Fig. 4.19 as long as the current remains well above the leakage current.

In the next three sections, we will discuss each region of inversion separately, looking for possible simplified expressions for the drain-source current. Such simplification is sought for several reasons. An obvious one is computational speed for simple calculations, which often accompany circuit design. Another reason for seeking simplification is that, by concentrating on the dominant phenomena in each region of inversion and by making certain approximations, the resulting form of the drain

current expression can “display” these phenomena in an obvious manner in some cases. Certain easily identifiable or measurable parameters can be made to appear in the expression. Finally, if the form of the functional dependence of the drain current on the terminal voltages is made apparent in a simple, explicit expression, a circuit designer may be able to take advantage of this form in creating new circuits.

4.7 Strong Inversion

A transistor is said to operate in strong inversion if at least one of the two channel ends is strongly inverted. Several models have been developed for the strong-inversion region. Each has something going for it. Historically, each strong-inversion model was developed independently starting with appropriate strong-inversion assumptions. However, in order to make clear where these models fit in the overall modeling framework, we will develop them starting from the all-region models of Sec. 4.3.

4.7.1 Complete Strong-Inversion Model

Nonsaturation With strong inversion guaranteed at the source end of the channel, if $V_{DB} = V_{SB}$ ($V_{DS} = 0$), the drain end will also be strongly inverted. If now the drain potential is raised, the level of inversion there will decrease and, eventually, strong inversion at that point will cease. For the present, we assume that the drain potential is sufficiently low so that this does *not* occur. With both channel ends strongly inverted, the corresponding surface potentials at these ends will be, from (3.4.5), using $V_{CB} = V_{SB}$ and $V_{CB} = V_{DB}$, respectively,

$$\psi_{s0} \approx \phi_0 + V_{SB} \quad (4.7.1a)$$

$$\psi_{sL} \approx \phi_0 + V_{DB} \quad (4.7.1b)$$

The most commonly used value for ϕ_0 is $2\phi_F$. However, this is not accurate, for reasons discussed near the beginning of Sec. 2.6.2. Thus, $\phi_0 = 2\phi_F + \Delta\phi$ should be used; for the *uniform substrates* we are assuming here, $\Delta\phi = 6\phi_t$ is a good compromise value for ϕ_0 , for common ranges of substrate doping concentration and oxide thickness.

Strong inversion at both ends ensures strong inversion throughout the channel since the surface potential varies monotonically from ψ_{s0} at the source to ψ_{sL} at the drain (Fig. 4.11). As established in Sec. 4.3, in strong inversion, the current is almost totally due to drift. Thus, we use (4.3.14a) to model it, with ψ_{s0} and ψ_{sL} as in (4.7.1). Denoting by I_{DSN} the current in the case we are presently considering (*both* channel ends strongly inverted), we have then the classical model^{8,14}

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left\{ (V_{GB} - V_{FB})(V_{DB} - V_{SB}) - \frac{1}{2} [(V_{DB} + \phi_0)^2 - (V_{SB} + \phi_0)^2] - \frac{2}{3} \gamma [(\phi_0 + V_{DB})^{3/2} - (\phi_0 + V_{SB})^{3/2}] \right\} \quad (4.7.2a)$$

or, after some manipulations,

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left\{ (V_{GB} - V_{FB} - \phi_0)(V_{DB} - V_{SB}) - \frac{1}{2} (V_{DB}^2 - V_{SB}^2) - \frac{2}{3} \gamma \left[(\phi_0 + V_{DB})^{3/2} - (\phi_0 + V_{SB})^{3/2} \right] \right\} \quad (4.7.2b)$$

Thus, the drain current becomes an *explicit* function of the terminal voltages, a very desirable result, which is to be contrasted with the general models of Sec. 4.3. Note also that I_{DSN} is of the form²³

$$I_{DSN} = \frac{W}{L} [g(V_{GB}, V_{DB}) - g(V_{GB}, V_{SB})] \quad (4.7.3)$$

which shows the symmetry of source and drain.[†] Another form of (4.7.2) is considered in Prob. 4.8. Equation (4.7.2) is the basis of the “level 2” model in the Berkeley SPICE simulator.⁹⁴

Derivation Directly in Strong Inversion In classical treatments, (4.7.2) is derived directly^{8,14} (rather than from the general case of Sec. 4.3), as follows. For any point x in the channel, where the surface potential is $\psi_s(x)$, a quantity $V_{CB}(x)$ is defined such that

$$\psi_s(x) = \phi_0 + V_{CB}(x) \quad (4.7.4)$$

Thus, using (4.7.1), we obtain $V_{CB}(0) = V_{SB}$ and $V_{CB}(L) = V_{DB}$, which is consistent with (4.7.1).

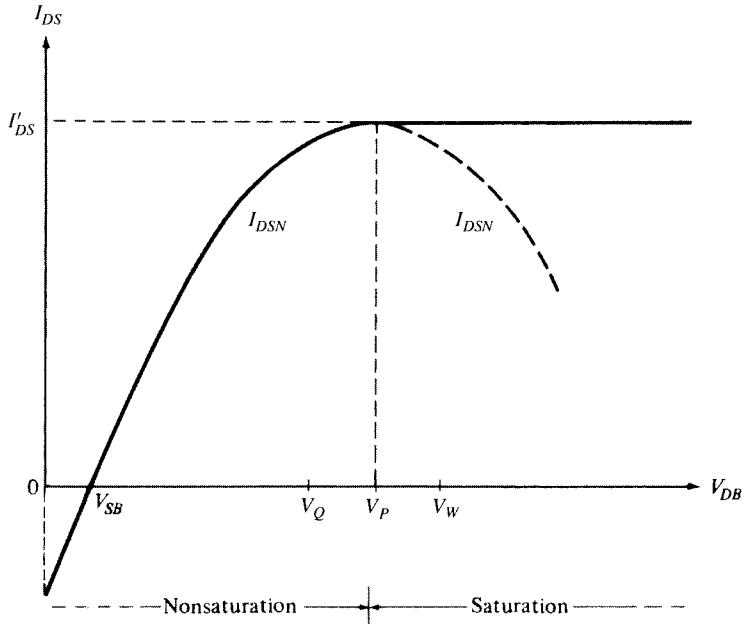
Recalling now the view of the *strong*-inversion layer as an n^+ region, which along with the body forms a field-induced n^+p junction (see Sec. 3.3), an easy interpretation becomes possible for $V_{CB}(x)$. It can be viewed as the “effective reverse bias” between the inversion layer at point x and the body, changing from V_{SB} at the source to V_{DB} at the drain. We have used the symbol V_{CB} so that the equations of Sec. 3.4.2 become directly applicable here.

Since ϕ_0 is assumed constant, from (4.7.4) we have $d\psi_s(x)/dx = dV_{CB}(x)/dx$. Thus, the drain current, assuming it is due only to drift, can be written from (4.3.3a) as follows:

$$I_{DSN} = \mu W(-Q'_L) \frac{dV_{CB}}{dx} \quad (4.7.5)$$

Integrating from $x = 0$ (where $V_{CB} = V_{SB}$) to $x = L$ (where $V_{CB} = V_{DB}$) produces

[†] Expressions (4.7.2) can be written in terms of source-referenced voltages by using in it $V_{GB} = V_{GS} + V_{SB}$ and $V_{DB} = V_{DS} + V_{SB}$. This, however, destroys the symmetry in the equations. We will postpone our discussion of source-referenced models until Sec. 4.7.3.

**FIGURE 4.20**

I_{DSN} computed from (4.7.2), and extension of its maximum value to the range $V_{DB} > V_P$. The solid curves taken together represent the complete strong-inversion model of (4.7.11).

$$I_{DSN} = \frac{W}{L} \int_{V_{SS}}^{V_{DB}} \mu(-Q'_l) dV_{CB} \quad (4.7.6)$$

The expression for Q'_l is, from (4.3.10) and (4.7.4):

$$Q'_l = -C'_{ox}(V_{GB} - V_{FB} - \phi_0 - V_{CB}) - Q'_B \quad (4.7.7)$$

where Q'_B is, from (4.3.12) and (4.7.4),

$$Q'_B = -\gamma C'_{ox} \sqrt{\phi_0 + V_{CB}} \quad (4.7.8)$$

Using (4.7.8) and (4.7.7), (4.7.6) gives exactly (4.7.2) again, assuming μ is constant along the channel. (The more general case where μ varies along the channel is discussed in Sec. 4.11.)

A graphical interpretation of I_{DSN} based on (4.7.6) is examined elsewhere.^{14,66,95} Equation (4.7.2) is plotted in Fig. 4.20 for the same V_{GB} and V_{SS} values as Fig. 4.7. We have extended the plot to V_{DB} values smaller than V_{SS} to show that the current equation is valid for such values. In fact, the smaller the value of V_{DB} , the stronger the level of inversion at the drain end. Thus, the channel remains strongly

inverted everywhere for such values, and our development remains valid. Note that V_{DB} is not used below zero, as then the drain-substrate junction would become forward-biased.[†]

Forward Saturation Assume now that $V_{DB} > V_{SB}$ and consider increasing values of V_{DB} . On the horizontal axis in Fig. 4.20, we show the values of V_{DB} from Fig. 4.7. Strictly speaking, (4.7.2) is valid *only* for $V_{DB} < V_Q$. Above V_Q , the channel is not in strong inversion near the drain, and thus that equation is not valid. Sometimes, nevertheless, (4.7.2) is used up to the point where the slope of the curve becomes zero; the resulting error may be tolerable in some applications. By setting $dI_{DS}/dV_{DB} = 0$ from (4.7.2), we find that the slope becomes zero at the following V_{DB} value:

$$V_P = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 - \phi_0 \quad (4.7.9)$$

The value of the drain current predicted by (4.7.2) with $V_{DB} = V_P$ is denoted by I'_{DS} , as shown in Fig. 4.20. Thus,

$I'_{DS} = I_{DSN} \Big|_{V_{DB}=V_P}$

(4.7.10)

The quantity V_P is the “pinchoff voltage,” which has been encountered in Sec. 3.5.2. According to the derivation and discussion there, it is easy to show that V_P is simply the value of V_{DB} that makes the gate-substrate threshold voltage at the *drain* end of the channel, $V_{TB}(V_{DB})$, equal to the applied gate-substrate voltage V_{GB} .²⁵ It is clear that the value of V_P depends on V_{GB} but has *nothing* to do with V_{SB} , as confirmed by (4.7.9). The value of V_P is close to the quantity V_W in (3.5.2) (which, we remind the reader, is the value of V_{DB} needed to bring the drain end of the channel to the boundary between moderate and weak inversion), as can also be seen in Fig. 4.20. If $\phi_0 = 2\phi_F$ is assumed, V_P reduces to V_W in (4.4.7).

As may be concluded from the preceding derivations, using V_{DB} values up to V_P in (4.7.2) implicitly assumes that the strong-inversion charge expressions are valid for Q'_l arbitrarily close to zero, predicting $Q'_l = 0$ at the drain end of the channel at $V_{DB} = V_P$. (The channel is said to be “pinched off” at that point.) This corresponds to assuming that the broken line in Fig. 3.16, which is a plot of (3.5.3) and (3.5.4), actually represents Q'_l all the way down to $Q'_l = 0$, which is clearly incorrect. In addition, if at the drain end of the channel $Q'_l = 0$ at $V_{DB} = V_P$, the carriers would have to travel with infinite drift velocity in order for a nonzero current to be possible, as seen from (1.3.7). (We talk about drift velocity in order to be consistent with the assumption on which the strong inversion model is based, i.e., that all current is due to drift.) To make more physical sense, we will allow for a very small, but nonzero, value

[†] It should be noted, though, that the models we are presenting work even with V_{SB} and/or V_{DB} somewhat negative, provided that the junction currents are negligible.

of Q'_l at the drain end of the channel in the following discussion; then the carrier speed must be large but finite nevertheless. For $V_{DB} > V_p$, a narrow region is taken to exist between the pinched-off tip of the channel and the drain, with very small $|Q'_l|$ in it; the carriers pass through this region with very high speed. This region is actually viewed practically as a depletion region, with the excess voltage $V_{DB} - V_p$ dropped across it. The inversion layer then need only support along its length the voltage it supported when $V_{DB} = V_p$. As V_{DB} is increased, the length of the depletion region must increase to support the excess voltage, but its length is still assumed very small in comparison to the channel length. Thus, the inversion layer length remains almost at the value L , and, since the voltage across it is the same as when $V_{DB} = V_p$, the current is still I'_{DS} .[†] We thus have, assuming always that the source end of the channel is strongly inverted,

$$I_{DS} = \begin{cases} I_{DSN}, & V_{DB} \leq V_p \\ I'_{DS}, & V_{DB} > V_p \end{cases} \quad (4.7.11a)$$

$$(4.7.11b)$$

corresponding to the solid curve in Fig. 4.20. Note that I_{DSN} should *not* be used for $V_{DB} > V_p$, because then a completely meaningless behavior is obtained, as shown by the broken curve in the figure. The region $V_{DB} \leq V_p$ is the nonsaturation region and the region for $V_{DB} > V_p$ is the saturation region.[‡] Note that despite the fact that in saturation the drain end of the channel is not strongly inverted, strong-inversion theory has been used in this simple model to predict I_{DS} over the whole range covered by (4.7.11). This explains the reason for the common practice of saying that the transistor “operates in strong inversion” for the entire range.

The preceding explanations concerning the current around $V_{DB} = V_p$ and in saturation are obviously not very satisfying. They have to be used because of the oversimplifying assumptions inherent to the model, as already discussed. Obviously, when this model is used in the transition region between V_Q and V_W or above it, some error can be expected. This will be more apparent in applications where the slope dI_{DS}/dV_{DB} in this transition region must be known accurately.⁹⁶ Let us recall, after all, that the electric field distribution next to the drain region is two-dimensional, and the “gradual channel approximation” does not hold there in the first place. However, no simple equation exists for the transition region between V_Q and V_W , and so (4.7.11) is often used. A comparison of (4.7.11) to the complete all-region model of Sec. 4.3 is considered in Prob. 4.11.

If the horizontal depletion region length discussed previously cannot be neglected in comparison to L , increasing $V_{DB} - V_p$ will increase the length of that

[†] This accuracy in modeling I'_{DS} is good despite the fact that the modeling of the situation near the drain leaves something to be desired. As discussed following (4.4.25), I_{DS} in saturation is determined by the level of inversion at the source. In saturation, the source inversion level has more control over I_{DS} than the details of operation near the drain. Thus, for the long-channel devices we are considering here, the effect of imprecision in modeling conditions at the drain does not affect much the overall accuracy of modeling I'_{DS} .

[‡] The nonsaturation region in strong inversion is alternatively called the *triode* region, because the $I_{DS}-V_{DS}$ characteristics in it remind one of the characteristics of a triode electron tube. Similarly, the saturation region in strong inversion is (rarely) referred to as the *pentode* region.

region, and thus it will decrease the “effective” channel length, resulting in non-negligible increase in the drain current. Then (4.7.11) will not be adequate. Such effects become important when L is not too large, and they are treated in Sec. 5.3 where a more careful look at the saturation region is taken.

Reverse Saturation In previously describing (forward) saturation, we have assumed $V_{DB} > V_{SB}$. If, instead, $V_{SB} > V_{DB}$ and V_{SB} is raised to V_P and beyond, the phenomena discussed previously will take place at the source rather than the drain. This is termed *reverse saturation*. In that case, what we have said previously will apply to the current entering the *source* terminal with the role of V_{DB} played by V_{SB} , and vice versa. The reverse saturation current, in analogy with the previous results, will be given by

$$I''_{DS} = I_{DSN} \Big|_{V_{SB}=V_P} \quad (4.7.12)$$

Complete Strong-Inversion Model Combining the three expressions we developed previously (for nonsaturation, forward saturation, and reverse saturation), we have a complete strong-inversion model as shown in Fig. 4.21. The quantities V_P and V_Q are fixed for a given V_{GB} . Notice that, to guarantee strong inversion at least at one channel end, we keep insisting that at least one of V_{DB} and V_{SB} be smaller than V_Q . In some of the literature this is not done; instead, the only restriction imposed is that one

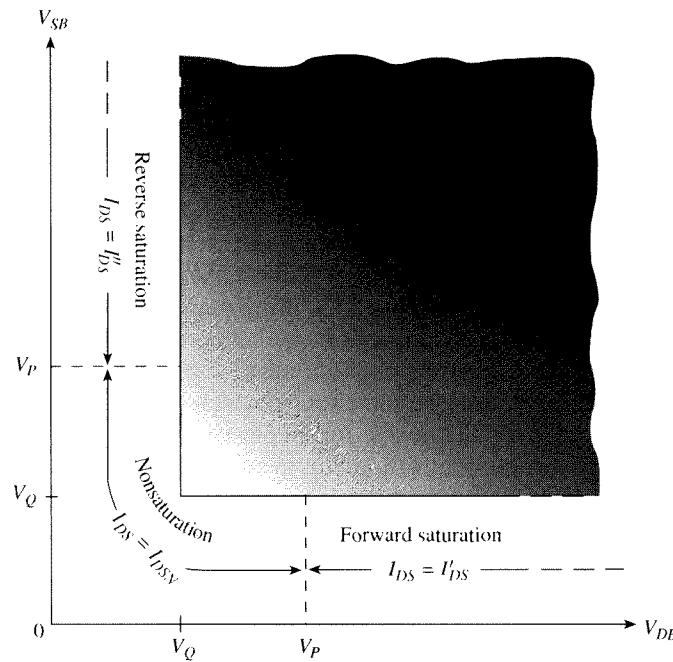


FIGURE 4.21
Complete strong-inversion model. The quantities V_P and V_Q depend on V_{GB} .

of V_{DB} and V_{SB} be smaller than V_p . With such relaxed assumptions, one cannot expect much accuracy of the model we have presented.

4.7.2 Body-Referenced Simplified Strong-Inversion Model

We can develop a strong-inversion model corresponding to the symmetric linearization model of Sec. 4.4.2. We assume again that the diffusion current is negligible in strong inversion. For the nonsaturation region, we can use (4.7.1) in (4.4.7) and (4.4.10a) to obtain

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left(V_{GB} - V_{FB} - \phi_0 - \frac{V_{SB} + V_{DB}}{2} - \gamma \sqrt{\phi_0 + \frac{V_{SB} + V_{DB}}{2}} \right) (V_{DB} - V_{SB}) \quad (4.7.13)$$

This model produces current values almost identical to (4.7.2) but is computationally simpler. On the other hand, the preceding expression can no longer be written in the form of (4.7.3); the terms containing V_{SB} and V_{DB} are now coupled through the square root. We will not carry this model further; extending it to saturation is left for Prob. 4.10. However, we will use the preceding form to show how different forms of equations can sometimes provide different insights, especially in circuit design.

Consider a situation in which the variations in V_{SB} and V_{DB} are balanced; i.e., one goes down by as much as the other goes up. It then becomes obvious from (4.7.13) that the entire quantity in the brackets will be fixed; thus, the current becomes proportional to $(V_{DB} - V_{SB})$, the voltage across the transistor. The transistor becomes a linear resistor. What is more, the conductance of this resistor, that is, the entire quantity in front of $(V_{DB} - V_{SB})$, can be controlled through the gate-body voltage, V_{GB} . This turns out to be a useful function for circuit design.⁹⁷

Notice that it would not have been easy to see the preceding property from (4.7.2). In fact, working with that equation (which is in principle more accurate), one can show that, under balanced operation, there are some residual odd-order nonlinearities. This is a good example of another important fact: one cannot judge the adequacy of an equation for all purposes by only looking at the overall error in the current; the aforementioned residual odd-order nonlinearities are not revealed by (4.7.13). Although such nonlinearities are minute in terms of contributing to the total current error, they may be important in predicting distortion in circuits, which in some cases must be kept very low. Nevertheless, (4.7.13) is valuable for the initial insight; once such insight is obtained, one can always refine things by going to more sophisticated equations. For the case we are discussing, further nonlinearities can be introduced by the variation of mobility with the gate field—a subject that will be discussed in Sec. 4.11.

4.7.3 Source-Referenced Simplified Strong-Inversion Model

A simple strong-inversion model²⁴ can be developed by substituting $\psi_{s0} = \phi_0 + V_{SB}$ and $\psi_{sL} = \phi_0 + V_{DB}$ from (4.7.1) into the expression for the drift component of the *simplified* source-referenced all-region model, (4.4.13a). This gives

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left[(V_{GB} - V_{SB} - V_{FB} - \phi_0 - \gamma\sqrt{\phi_0 + V_{SB}}) (V_{DB} - V_{SB}) - \frac{\alpha}{2} (V_{DB} - V_{SB})^2 \right] \quad (4.7.14)$$

where α has nominally the value given by (4.4.14) with $\psi_{s0} = \phi_0 + V_{SB}$; the value of this quantity will be discussed in detail later.

Substituting $V_{DB} - V_{SB}$ by V_{DS} and $V_{GB} - V_{SB}$ by V_{GS} in (4.7.14) gives

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left[(V_{GS} - V_T|_{V_{SB}}) V_{DS} - \frac{\alpha}{2} V_{DS}^2 \right] \quad (4.7.15)$$

where $V_T|_{V_{SB}}$ is the gate-source extrapolated threshold voltage, obtained from (3.4.13) with $V_{CB} = V_{SB}$.

Derivation Directly in Strong Inversion The preceding model can also be derived directly in strong inversion using (4.7.6). Q'_B is the same as in (4.7.7), repeated here:

$$Q'_B = -C'_{ox} (V_{GB} - V_{FB} - \phi_0 - V_{CB}) - Q'_B \quad (4.7.16)$$

For Q'_B , we use (4.4.1) with ψ_s from (4.7.4) and, from (4.4.12) and (4.7.1a), $\psi_{se} = \psi_{s0} = \phi_0 + V_{SB}$. In other words

$$Q'_B = -C'_{ox} \left[\gamma\sqrt{\phi_0 + V_{SB}} + (\alpha - 1)(V_{CB} - V_{SB}) \right] \quad (4.7.17)$$

Using this approach, and assuming μ is constant, (4.7.7) with (4.7.6) give (4.7.14) and thus (4.7.15) again.

One could also have derived (4.7.15) directly from the accurate current expression in (4.7.2). This is indeed possible but amounts to little more than a mathematical exercise, not revealing the underlying physical assumptions in the simplified model. The approach is considered in Prob. 4.12.

In most of the literature, the fact that the threshold voltage is taken at the source is *not* indicated explicitly. For simplicity, we will comply with this practice. Thus, we will write (4.7.15) as

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left[(V_{GS} - V_T) V_{DS} - \frac{\alpha}{2} V_{DS}^2 \right] \quad (4.7.18)$$

where, for the rest of this book, V_T will be *defined* as the gate-source extrapolated threshold voltage given by

$$V_T = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0 + V_{SB}} \quad (4.7.19)$$

or

$$V_T = V_{T0} + \gamma(\sqrt{\phi_0 + V_{SB}} - \sqrt{\phi_0}) \quad (4.7.20a)$$

with

$$V_{T0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0} \quad (4.7.20b)$$

Clearly, V_T increases with V_{SB} owing to the body effect. In fact, in many treatments the term *body effect* implies only the dependence of V_T on V_{SB} . A more general point of view was provided in Sec. 3.3. The parameter ϕ_0 is often taken equal to $2\phi_F$, but this is not justifiable, as has already been explained near the beginning of Sec. 2.6.2.

The value of V_{SB} is normally zero or positive. Sometimes, in low-voltage circuits work, somewhat negative values for V_{SB} are used to reduce the value of V_T , but it is made sure that V_{SB} is not too negative, so that no significant forward-bias current flows in the source-body junction. Equation (4.7.20) remains approximately valid in such cases.⁹⁸ This is because, although we have been assuming that V_{SB} is normally nonnegative, if the development of this equation is traced, it will be seen that no such assumption plays a role in it (always provided that the junction current is negligible).

A plot of I_{DSN} from (4.7.18) is shown in Fig. 4.22. The value V'_{DS} of V_{DS} at which the maximum occurs is found from (4.7.18) by setting $dI_{DSN}/dV_{DS} = 0$ and solving for $V_{DS} = V'_{DS}$:

$$V'_{DS} = \frac{V_{GS} - V_T}{\alpha} \quad (4.7.21)$$

The corresponding value of the drain current I'_{DS} is found by using $V_{DS} = V'_{DS}$ from the preceding equation in (4.7.18):

$$I'_{DS} = \frac{W}{L} \mu C'_{ox} \frac{(V_{GS} - V_T)^2}{2\alpha} \quad (4.7.22)$$

As in the accurate model, this is taken to be the value of I_{DS} for $V_{DS} > V''_{DS}$ also. The complete model then becomes, in analogy to (4.7.11):

$$I_{DS} = \begin{cases} I_{DSN}, & V_{DS} \leq V'_{DS} \\ I'_{DS}, & V_{DS} > V'_{DS} \end{cases} \quad (4.7.23a)$$

$$(4.7.23b)$$

Thus, using the equations developed for I_{DSN} and I'_{DS} , we have

$$I_{DS} = \begin{cases} \frac{W}{L} \mu C'_{ox} [(V_{GS} - V_T)V_{DS} - \frac{\alpha}{2} V_{DS}^2], & V_{DS} \leq V'_{DS} \\ \frac{W}{L} \mu C'_{ox} \frac{(V_{GS} - V_T)^2}{2\alpha}, & V_{DS} > V'_{DS} \end{cases} \quad (4.7.24a)$$

$$(4.7.24b)$$

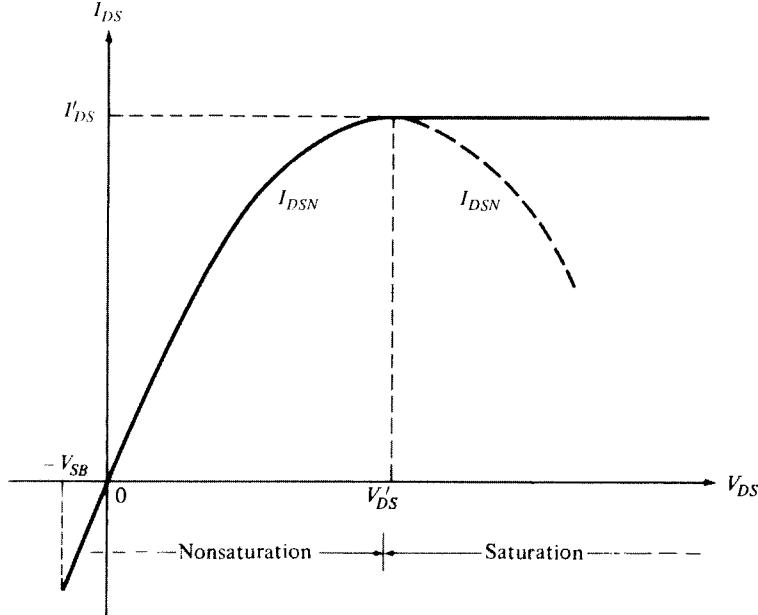


FIGURE 4.22
 I_{DSN} as computed from (4.7.18), and extension of its maximum value to the range $V_{DS} > V'_{DS}$. The solid curves taken together represent the strong-inversion model in (4.7.24).

This corresponds to the solid line in Fig. 4.22. A set of I_{DS} - V_{DS} characteristics for various V_{GS} values is shown in Fig. 4.23. Such characteristics depend on V_{SB} through V_T in (4.7.20).

A comparison of the source-referenced strong-inversion model to the symmetric linearization all-region model is shown in Fig. 4.24. The parameters ϕ_0 and α have been adjusted for best fit in the saturation region. Results are shown for two values of V_{SB} ; as seen, the increase in the threshold voltage caused by the body effect results in a “shrinkage” of the characteristics.

The above model²⁴ is the basis of the “level 3” model in the Berkeley SPICE simulator.⁹⁴

Equation (4.7.24) can be put in a very compact form by defining a convenient quantity η as follows:^{99,100}

$$\eta = \begin{cases} 1 - \frac{V_{DS}}{V'_{DS}}, & V_{DS} \leq V'_{DS} \\ 0, & V_{DS} > V'_{DS} \end{cases} \quad (4.7.25a)$$

$$(4.7.25b)$$

This quantity is plotted in Fig. 4.25. The drain current can now be expressed as follows:

$$I_{DS} = I'_{DS} (1 - \eta^2), \text{ both nonsaturation and saturation} \quad (4.7.26)$$

where I'_{DS} is given by (4.7.22).

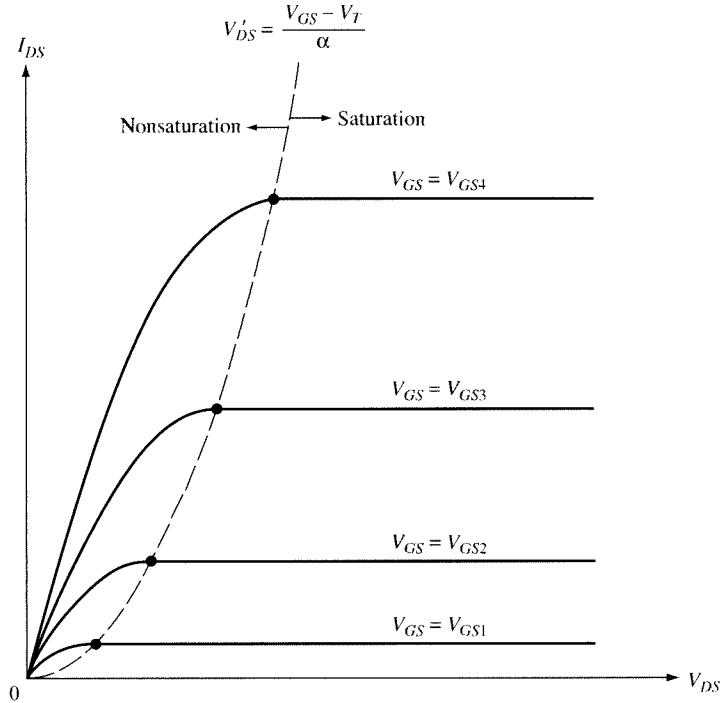


FIGURE 4.23
 I_{DS} - V_{DS} characteristics as obtained from the simplified strong-inversion model of (4.7.24).

It is easy to verify that at $V_{DS} = V'_{DS}$ (4.7.16) with (4.7.17) predict $Q'_l = 0$ at the drain end of the channel. The reason for this unrealistic result is as discussed for the complete strong-inversion model and will not be repeated. Similarly applicable here are the comments made for that model concerning possible inaccuracies in the neighborhood of $V_{DS} = V'_{DS}$. Special “smoothing functions” have been used to improve the transition from nonsaturation to saturation.^{101–103}

We now return to the problem of choosing an appropriate value for α in (4.7.24). The derivation of MOSFET characteristics in the early days implicitly assumed $\alpha = 1$. This corresponds to line *c* in Fig. 4.15, which is a very poor approximation to the solid line in the same figure. This approximation is equivalent to assuming that the depletion region depth is the same all along the channel and equal to its actual value at the source. The result is that $|Q'_B|$ is underestimated everywhere except at the source. This, from (4.7.7), results in an overestimation of $|Q'_l|$ and a resulting overestimation of I_{DS} . The error in I_{DS} can be *large* for devices in which γ is not small; we will come back to this point later. Note also that, since $|Q'_l|$ is overestimated, an artificially large value of V_{DS} would be needed to reduce $|Q'_l|$ to zero at the drain. This value of V_{DS} is V'_{DS} , as already mentioned, under the simplified assumptions in the present model. The overestimation of V'_{DS} is also seen from (4.7.21) with $\alpha = 1$, rather than with a more realistic value of α larger than unity.

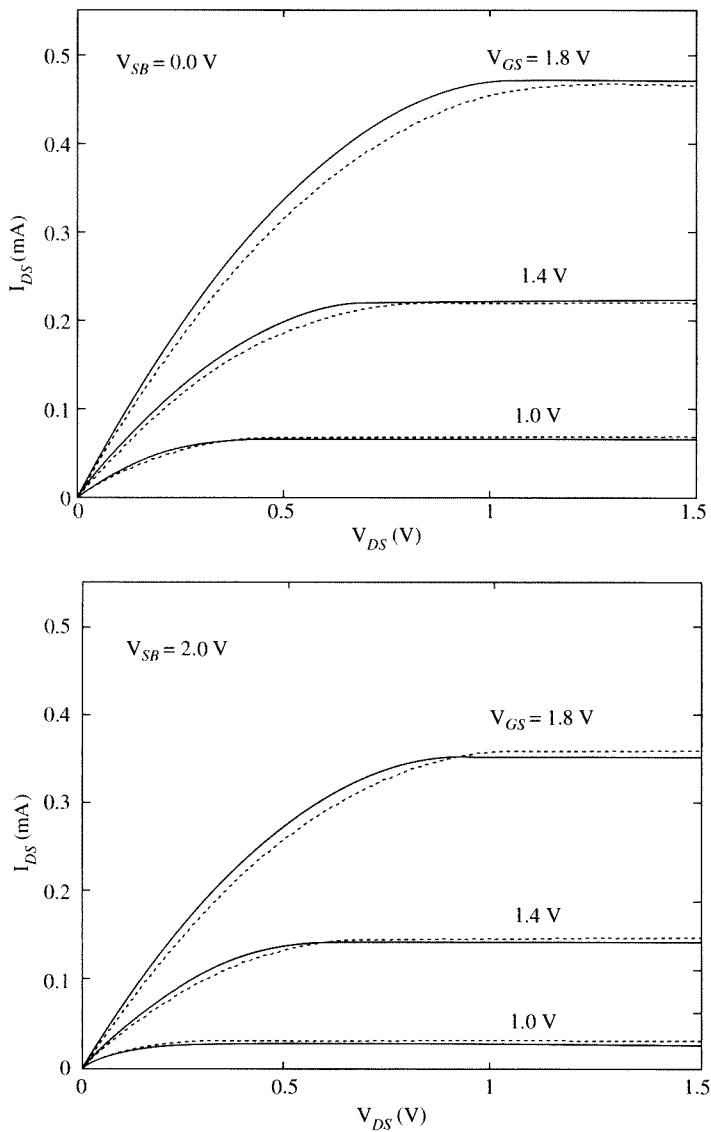


FIGURE 4.24
 I_{DS} - V_{DS} characteristics for two different V_{SB} values. Solid lines: source-referenced simplified strong-inversion model (4.7.24). Dashed lines: simplified all-region model of (4.4.10). ϕ_0 and α are chosen for a best fit between the two models in saturation. $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2.5 \text{ nm}$, $W = L = 10 \mu\text{m}$, $\mu = 400 \text{ cm}^2/(\text{V}\cdot\text{s})$.

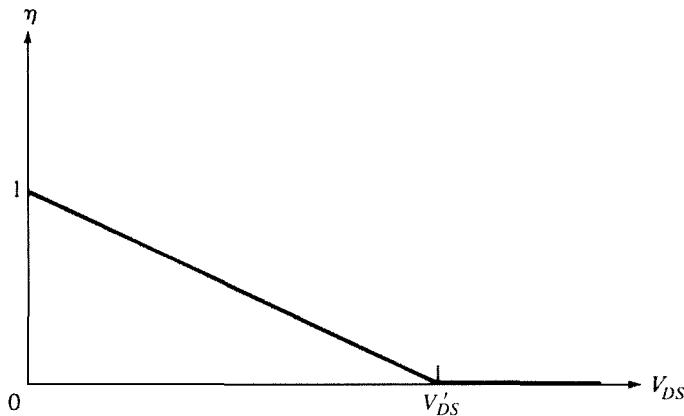


FIGURE 4.25
Quantity η vs. V_{DS} .

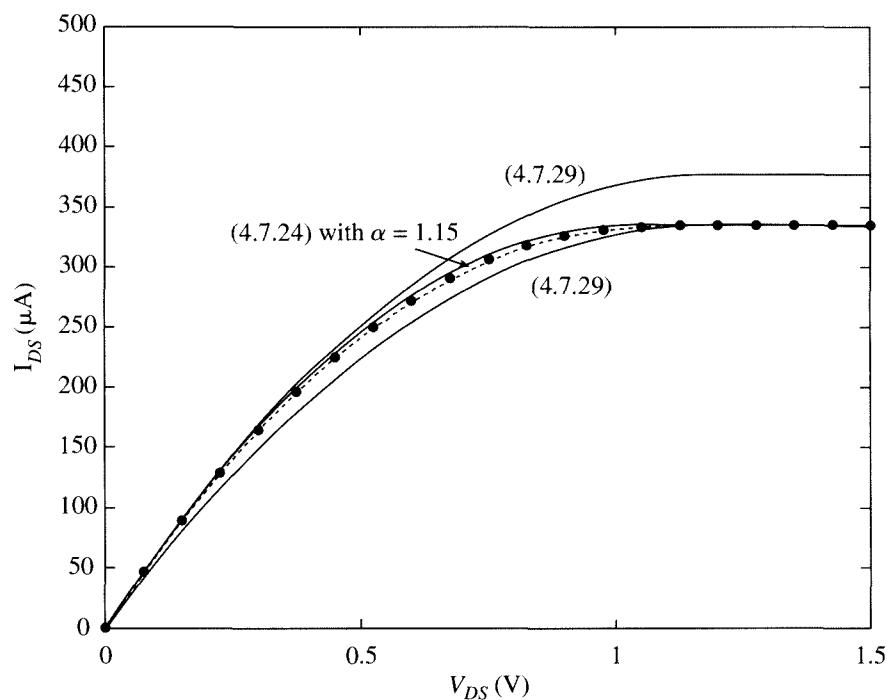


FIGURE 4.26
 I_{DS} - V_{DS} curves for $V_{GS} = 1.4$ V and $V_{SB} = 0.0$ V. Dots: full numerical solution of semiconductor equations. Lower solid line: (4.7.29), adjusted for best match in saturation. Upper solid line: (4.7.29), adjusted for best match at very low V_{DS} values. Middle solid line: (4.7.24), with $\alpha = 1.15$ chosen for good matching at very low V_{DS} and in the saturation region. $N_A = 7.0 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2.0 \text{ nm}$, $W = L = 10 \mu\text{m}$.

The next value we will consider for α is the one that results from the Taylor expansion leading to (4.4.3) with (4.4.12).²⁴ This value corresponds to line *a* in Fig. 4.15 and is given by[†]

$$\alpha_1 = 1 + \frac{\gamma}{2\sqrt{\phi_0 + V_{SB}}} \quad (4.7.27)$$

It is clear from the figure that this would be a good value to use for α only if $V_{DS} = V_{DB} - V_{SB}$ were small. In the general case, though, it results in overestimating $|Q'_B|$ and, hence, underestimating $|Q'_I|$. Thus, I_{DS} and V'_{DS} will be underestimated, i.e., the errors here are in the opposite direction from those for $\alpha = 1$. Clearly, one should seek values for α between 1 and α_1 , resulting in the approximation indicated by line *b* in Fig. 4.15. Several such values have been proposed in the literature, and they usually correspond to modifications of (4.7.27) to lower its value. Such modifications are of the following general form:

$$\alpha = 1 + d_2 \frac{\gamma}{2\sqrt{\phi_0 + V_{SB} + \phi_3}} \quad (4.7.28)$$

where various values for the quantities d_2 and ϕ_3 are used. In one approach,^{24,104} $\phi_3 = 0$ is used, with d_2 between 0.5 and 0.8. The quantity d_2 is sometimes allowed to depend on V_{SB} for better fit.¹⁰⁵ This, however, detracts from the simplicity of the model. Another approach¹⁰⁶ is to use $d_2 = 1$, and a fixed ϕ_3 value of the order of 1 V. Elsewhere, ϕ_3 is made a function of V_{GS} .^{77,107} In general, the exact nature of the modifications used depends on the range of biases over which reasonable results are expected.

The reader may wonder whether, following the several approximations that had to be used to derive the preceding simple models, these models provide satisfactory accuracy. The answer depends on how the parameters in the equations are chosen. If the values dictated from physics are used in them, the accuracy these models provide may not be adequate. If, instead, an optimizer is used to vary these parameters, as if they were free variables, until minimum I_{DS} error is obtained, the accuracy may be improved. This is often done during “parameter extraction,” a process considered in Chap. 10. However, this practice is not always without penalty, as discussed in Sec. 4.17 and, in more detail, in Chap. 10. A good model has just enough parameters to model all effects considered; this gives it enough degrees of freedom, if the value of these parameters must be adjusted. Some models do not have enough degrees of freedom, so that no matter what value is chosen for the parameters in them, no global accuracy is possible. We now present an example of such a situation.

[†] It is interesting to note here that, as can be verified by using (4.7.19), we have

$$\alpha_1 = 1 + \frac{dV_T}{dV_{SB}}$$

Issues with a Popular Model for Circuit Design When γ is *small*, the value of α is close to 1, as seen from the previous equations. In this case, (4.7.24) becomes

$$I_{DS} = \begin{cases} \frac{W}{L}\mu C'_{ox} \left[(V_{GS} - V_T) V_{DS} - \frac{1}{2} V_{DS}^2 \right], & V_{DS} \leq V'_{DS}, \text{ small } \gamma \\ \frac{W}{L}\mu C'_{ox} \frac{(V_{GS} - V_T)^2}{2}, & V_{DS} > V'_{DS}, \text{ small } \gamma \end{cases} \quad (4.7.29a)$$

$$I_{DS} = \begin{cases} \frac{W}{L}\mu C'_{ox} \frac{(V_{GS} - V_T)^2}{2}, & V_{DS} > V'_{DS}, \text{ small } \gamma \end{cases} \quad (4.7.29b)$$

where $V'_{DS} = V_{GS} - V_T$. These equations have been used widely for approximate hand calculations for circuit design, and even for quick computer calculations,²¹ and are the basis of the historical “level 1” model in SPICE.⁹⁴ However, in circuits literature these equations are often used indiscriminately even for devices with large γ , which can result in *significant* error. As an example, consider the dots in Fig. 4.26, obtained by full numerical solution of the semiconductor equations for the parameters given in the caption. The lower curve represents (4.7.29), with the value for $(W/L)\mu C'_{ox}$ adjusted for good matching in saturation. The model clearly fails in nonsaturation, and it predicts a V'_{DS} which is too large. If $(W/L)\mu C'_{ox}$ is instead adjusted for a good fit in the low nonsaturation region, then the model is inaccurate in saturation, as shown by the upper curve. If the model of (4.7.24) is used instead, we have the additional flexibility of choosing α ; with $\alpha = 1.15$, we obtain the middle curve, which agrees well with the accurate calculations (dots). *The inadequacy of (4.7.29) is even more apparent in processes with large body effect.*

For very crude modeling work, the strong-inversion equations are sometimes taken to be valid for V_{GS} as low as V_T , where they predict $I_{DS} = 0$. In such cases, the moderate- and weak-inversion regions are not considered.

Potential vs. Position In strong inversion, one can develop a rather simple relation between the position x in the channel and the effective reverse bias $V_{CB}(x)$ at that point, extending an idea developed in Sec. 4.3. All strong-inversion expressions we have given for the nonsaturation current can be written in the form

$$I_{DSN} = \frac{W}{L} H(V_{GB}, V_{SB}, V_{DB}) \quad (4.7.30)$$

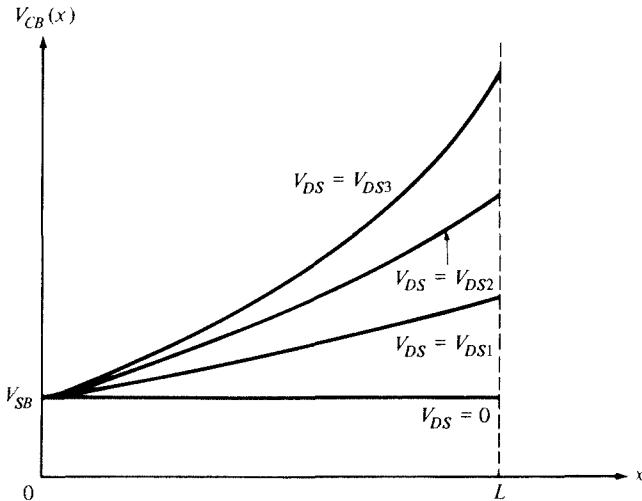
with the function H depending on which model is being used. If we consider point x in the channel as the drain of a fictitious transistor with length x , we will have

$$I_{DSN} = \frac{W}{x} H(V_{GB}, V_{SB}, V_{CB}(x)) \quad (4.7.31)$$

Eliminating I_{DSN} among the two equations gives

$$\frac{x}{L} = \frac{H(V_{GB}, V_{SB}, V_{CB}(x))}{H(V_{GB}, V_{SB}, V_{DB})} \quad (4.7.32)$$

This equation gives the relation between x and $V_{CB}(x)$. As an example, using the approximate model developed previously results in a simple expression. Extending this expression to include the saturation region, and using the parameter η in (4.7.25), we can put it in the form (Prob. 4.17)

**FIGURE 4.27**

Channel “effective reverse bias” with respect to the body vs. distance from the source for the simplified strong-inversion model and for four V_{DS} values.

$$V_{CB}(x) = V_{SB} + \frac{V_{GS} - V_T}{\alpha} \left[1 - \sqrt{1 - \frac{x}{L}(1 - \eta^2)} \right] \quad (4.7.33)$$

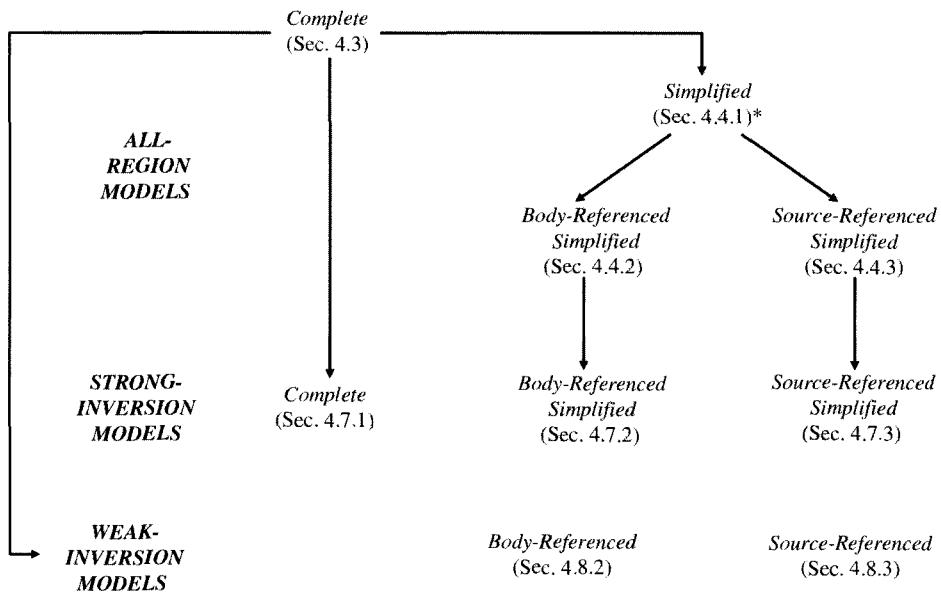
This relation is plotted for various values of V_{DS} in Fig. 4.27 ($0 < V_{DS1} < V_{DS2} < V_{DS3}$). As seen, for small V_{DS} values, V_{CB} is nearly proportional to x . This is because for such V_{DS} values the inversion layer is nearly uniform, so the voltage is distributed almost evenly along its length. At larger V_{DS} values, this is no longer the case. This can be explained easily. For a given V_{DS} , say $V_{DS} = V_{DS3}$, one can find from the plot the voltage drop across a section of the inversion layer of length Δx between two points x_1 and x_2 . The drop will simply be $\Delta V_{CB} = V_{CB}(x_2) - V_{CB}(x_1)$. If another section of the same length Δx is taken closer to the drain, the voltage ΔV_{CB} across it will be larger, for the following reason. The second section contains a smaller inversion layer charge per unit area, since $|Q'_I|$ decreases, as one goes from the source toward the drain [this follows, for example, from (4.7.16) and (4.7.17)]. Thus a larger voltage will be needed across the second section to support the same current (Sec. 1.3.2). From this argument, it is clear why the slope dV_{CB}/dx increases with increasing x . A related observation concerns the drift velocity of electrons. As follows from (1.3.7), to carry the same current the electrons must move faster at points where the inversion charge magnitude per unit area is smaller. Since the electrons are in abundance at the source end of the channel, they can carry the required charge per unit time (current) while moving slowly. In a given small element of length Δx near the drain, fewer electrons are to be found at any given instant than within a corresponding element of the same length near the source; thus, the electrons move faster as they approach the drain. If V_{DS} is large, the electrons can be traveling so fast that “velocity saturation” can occur near the drain. The effect of this on the transistor I - V characteristics can become significant in short-channel devices and is considered in Chap. 5.

If $V_{CB}(x)$ were plotted for $V_{DS} = V'_{DS}$ using (4.7.33), the slope would be infinite at $x = L$. This physically impossible result, corresponding to an infinite electric field at that point, is simply a result of the limitations of the model as $V_{DS} = V'_{DS}$ is approached. As already discussed, the model implicitly assumes $Q'_L = 0$ at the drain for $V_{DS} = V'_{DS}$; this is wrong as already explained. In addition, the model fails to take into account the presence of diffusion currents. Whereas in strong inversion, the I_{DS} component due to diffusion is small, the portion of the channel current that is due to diffusion, $I_{\text{diff}}(x)$ in (4.3.3b), increases with increasing x at the expense of $I_{\text{drift}}(x)$, as has already been remarked.⁴⁹ Thus, any conclusion reached at $V_{DS} = V'_{DS}$, based on the assumption of drift current only, should be scrutinized to make sure it leads to reasonable results.

Comparison to the Complete Strong-Inversion Model In strong inversion, if the complete strong-inversion model introduced in Sec. 4.7.1 is compared with the complete all-region model of Sec. 4.3, the agreement is found to be very good. Compared with these models, the simplified strong-inversion model of (4.7.24) is somewhat in error (e.g., up to 5 percent or more in terms of current values). In practice, though, several factors contribute to the simplified model being favored over the complete one in a variety of situations. These factors are summarized as follows:

1. The simplified strong-inversion model is simple. This is a desirable feature when quick hand calculations are needed in the course of circuit design.
2. Practical devices exhibit high-order effects that are not taken into account by either model; for example, the body is never exactly uniform, owing to fabrication details (Chap. 9). This tends to make the difference between the two models less significant, in terms of accuracy.
3. The assumptions behind each model will be seen later to lead to corresponding expressions for the total charges (Chap. 6) and the capacitances (Chap. 7) of the device. In the case of the simplified strong-inversion model, the expressions for these quantities are simple. In contrast to this, the expressions for some of these quantities, as derived for the complete strong-inversion model, are so complicated as to be totally impractical. In addition, the assumptions behind the simplified model are known to lead to simple expressions suitable for very high-frequency operation (Chap. 8); this is not known to be possible for the complete model.
4. The simplified model contains explicitly the threshold voltage V_T . This is an important parameter that is widely discussed, used, and measured.[†] In fact, even for short-channel and/or narrow-channel devices, use is made of the simplified model with V_T replaced by an effective threshold (Chap. 5). A large amount of work has been done on how to estimate and measure this quantity in such cases. In contrast, the complete model does not contain the threshold

[†] In this book, the term *threshold voltage* will always imply *extrapolated* threshold voltage unless stated otherwise. Unfortunately, in the literature on MOS devices the term “threshold voltage” is used with several different meanings. This is discussed in Sec. 4.17.



*Charge formulation in Sec. 4.4.4.

FIGURE 4.28
Model hierarchy.

explicitly, and it is more difficult to adopt it for use with short- and/or narrow-channel devices.

Thus, the more versatile simplified model is favored for general use in strong inversion. In cases, however, where the nuances of strong inversion need to be studied, the complete model is to be preferred.

4.7.4 Model Origin Summary

Several models have already been discussed. To help place these models, and their conceptual origin, as derived in this book, in perspective, we provide the chart of Fig. 4.28.[†] All models mentioned in it are in use, and one or the other may be preferable depending on the application.

In Fig. 4.28, we have also included models for weak inversion, which are the subject of the next section.

[†] We note that, historically, the derivation of these models did not follow the flow indicated in Fig. 4.28. References to the historical origins of each model have been given in the corresponding sections.

4.8 Weak Inversion

4.8.1 Special Conditions in Weak Inversion

As seen from the last entry in Table 4.1, for a transistor operating in weak inversion^{26–39} no part of the channel is moderately or strongly inverted. The conditions for weak inversion operation have been given in terms of terminal voltages in Sec. 4.6.

The general model developed in Sec. 4.3 can be used to produce a simplified equation for I_{DS} in the weak-inversion region. This will be discussed later. For the present, we prefer to show the classical approach for obtaining such an equation to provide some independent intuition about this region of operation.

As follows from the material in Sec. 3.4.3, for a weakly inverted point in the channel, the surface potential satisfies

$$\psi_s \approx \psi_{sa}(V_{GB}) \quad (4.8.1)$$

where

$$\psi_{sa}(V_{GB}) = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 \quad (4.8.2)$$

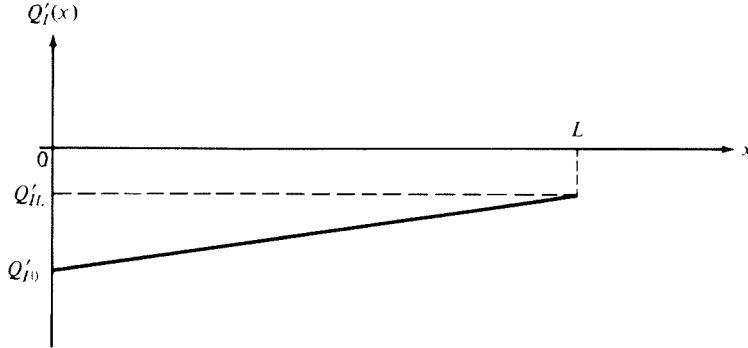
Let us assume for the sake of simplicity that ψ_s is *exactly* equal to $\psi_{sa}(V_{GB})$; this will not affect the forthcoming analysis. Since the surface potential depends only on V_{GB} , it is independent of the position along the channel. This implies two important facts:

1. Q'_B will be independent of position along the channel, as seen from (4.3.12). This means that the depletion region depth does not change along the channel.
2. Since all points at the surface are assumed at the same potential with respect to the substrate, the potential difference between such points is zero. Therefore, the electric field has a zero horizontal component. If there is current through the channel, then, it cannot be caused by drift; thus, all current must be caused by diffusion.

It follows then from the material in Sec. 1.3.3 that the plot of Q'_I vs. x must be a straight line (Fig. 4.29). Thus, (1.3.22) gives, with $Q' = Q'_F$, $b = W$, and $a = L$:

$$I_{DS} = -\frac{W}{L} \mu \phi_i (Q'_{I0} - Q'_{IL}) \quad (4.8.3)$$

Indeed, this is the same as (4.3.9b), which was seen to give the contribution of diffusion current to I_{DS} . Thus, (4.8.3) can be considered as resulting directly from the all-region model of Sec. 4.3.

**FIGURE 4.29**

Inversion layer charge per unit area vs. distance from the source in weak inversion.

4.8.2 Body-Referenced Model

The two values of Q'_I in (4.8.3) can be found from (3.4.20), which is valid in weak inversion and even in depletion:

$$Q'_{I0} = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_{sa}(V_{GB})}} \phi_t e^{[\psi_{sa}(V_{GB}) - 2\phi_F]/\phi_t} \cdot e^{-V_{SB}/\phi_t} \quad (4.8.4)$$

$$Q'_{IL} = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_{sd}(V_{GB})}} \phi_t e^{[\psi_{sd}(V_{GB}) - 2\phi_F]/\phi_t} \cdot e^{-V_{DB}/\phi_t} \quad (4.8.5)$$

Using these in (4.8.3) we obtain

$$I_{DS} = \frac{W}{L} \hat{I}(V_{GB}) (e^{-V_{SB}/\phi_t} - e^{-V_{DB}/\phi_t}) \quad (4.8.6)$$

where

$$\hat{I}(V_{GB}) = \mu \frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\psi_{sa}(V_{GB})}} \phi_t^2 e^{[\psi_{sa}(V_{GB}) - 2\phi_F]/\phi_t} \quad (4.8.7)$$

Readers familiar with bipolar transistors will recognize the similarity of (4.6.6) to the Ebers-Moll equations.⁶⁵ This should not be surprising, as similar mechanisms are responsible for current flow in the bipolar transistor (under common assumptions) and the weakly inverted MOS transistor.⁶⁸

Since (3.4.20), which was used in the above development, is valid even in depletion, (4.8.6) will be valid even if V_{DB} is so large that the drain end of the channel is less than “weakly” inverted.

We can now return to the question of obtaining simplified weak-inversion expressions starting from the general model of Sec. 4.3. Since in weak inversion the current is predominately caused by diffusion, we can work with (4.3.14b) and assume $I_{DS} \approx I_{DS2}$. Unfortunately, if we use (4.8.1) and substitute $\psi_{s0} = \psi_{sa}$ and $\psi_M = \psi_{sa}$ in (4.3.14b), we obtain $I_{DS} = 0$! This is owing to a weakness of the form of this equation. In weak inversion, it relies on differences of nearly equal quantities; thus, approximations such as the preceding are inadequate. Of course, ψ_{s0} and ψ_{sL} are *not exactly* equal; ψ_s only reaches ψ_{sa} asymptotically in Fig. 4.6. To preserve appropriately the minute difference between the two quantities, we can employ (4.3.15). The procedure is outlined in Prob. 4.22 and produces exactly (4.8.6) and (4.8.7) again.

4.8.3 Source-Referenced Model

Equation (4.8.3) can be rewritten as follows:

$$I_{DS} = -\frac{W}{L} \mu \phi_t Q'_{I0} \left(1 - \frac{Q'_{IL}}{Q'_{I0}} \right) \quad (4.8.8)$$

From (4.8.4) and (4.8.5) we have

$$\frac{Q'_{IL}}{Q'_{I0}} = e^{-(V_{DB} - V_{SB})/\phi_t} = e^{-V_{DS}/\phi_t} \quad (4.8.9)$$

Thus, I_{DS} becomes

$$I_{DS} = \frac{W}{L} \mu \phi_t (-Q'_{I0}) (1 - e^{-V_{DS}/\phi_t}) \quad (4.8.10)$$

which can be evaluated by using (4.8.4). A compact expression³⁵ can be developed by using the approximation in (3.4.25b), with subscript *C* replaced by *S*. Using this in (4.8.10) gives

$$I_{DS} = \frac{W}{L} I'_M e^{(V_{GS} - V_M)/(n\phi_t)} (1 - e^{-V_{DS}/\phi_t}), \quad \text{fixed } V_{SB} = V'_{SB} \quad (4.8.11)$$

where V_M is the upper limit of weak inversion in terms of V_{GS} , for the given value of V_{SB} :[†]

$$V_M = V_{FB} + 2\phi_F + \gamma \sqrt{2\phi_F + V'_{SB}} \quad (4.8.12)$$

and

$$I'_M = \mu \frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{2\phi_F + V'_{SB}}} \phi_t^2 \quad (4.8.13)$$

[†] Sometimes V_I is used in lieu of V_M in (4.8.11); in this case, the value of I'_M needs to be modified accordingly.

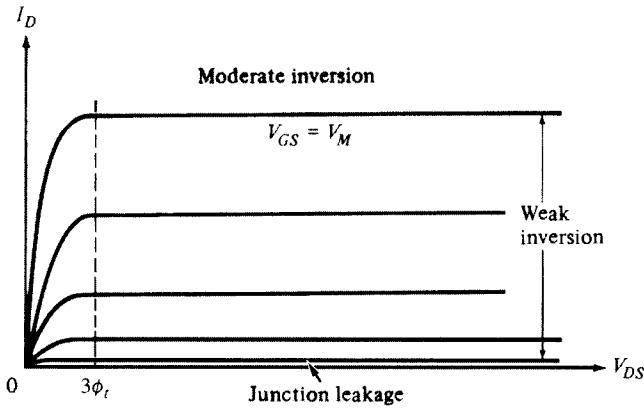


FIGURE 4.30
 I_D - V_{DS} characteristics in weak inversion, with V_{GS} as a parameter.

and with the quantity n evaluated from (3.4.22) at $\psi_{sa} = 2\phi_F + V'_{SB}$ (see Fig. 3.11):

$$n = 1 + \frac{\gamma}{2\sqrt{2\phi_F + V'_{SB}}} \quad (4.8.14)$$

It is emphasized that this equation, along with (4.8.11), is useful for studying the approximate variation of I_{DS} with V_{GS} , when V_{SB} is fixed at some value denoted by V'_{SB} . Trying to use this equation to study the variation of I_{DS} with V_{SB} is not advisable, since several quantities in the equation (V_M , I_M , n) depend on V'_{SB} in a complicated manner. In such cases, (4.8.6) and (4.8.7) should be used. It is clear from (4.8.6) that I_{DS} depends on V_{SB} only through the quantity $\exp(-V_{SB}/\phi_i)$.

I_D is plotted vs. V_{DS} by using (4.8.11) in Fig. 4.30, with V_{GS} as a parameter, for a fixed V_{SB} . I_D is practically the same as I_{DS} if the latter is much larger than the leakage currents (see the following). As seen, the curves become horizontal for V_{DS} larger than a few ϕ_i , since the last exponential in the equation becomes negligible compared to 1. This happens at V_{DS} values which are independent of V_{GS} , a fact which is in sharp contrast to the case of strong-inversion operation (Fig. 4.23). See also Fig. 1.36.

Using equal V_{GS} steps, the vertical spacing of successive curves in Fig. 4.30 for a given V_{DS} increases nearly exponentially. This exponential behavior is brought out clearly by plotting $\log I_D$ vs. V_{GS} with V_{DS} fixed, as shown in Fig. 4.31. The result is nearly a straight line in weak inversion. Above this region we have moderate inversion, where I_{DS} does not vary nearly exponentially with V_{GS} . At the bottom, I_{DS} is so small that it is masked by the leakage current I_l , which increases in general with temperature. At common temperatures, the weak-inversion region can span several decades of current.

The slope of $\log I_{DS}$ vs. V_{GS} determines how well a transistor can be turned off by reducing V_{GS} , for digital applications. An often-used measure for this is the “subthreshold swing” S ,[†] defined as the amount by which V_{GS} must be reduced in order for the weak inversion current to be reduced by 1 order of magnitude (usually

[†] Also called *subthreshold slope*. The term *subthreshold* is often used to mean weak-inversion operation.

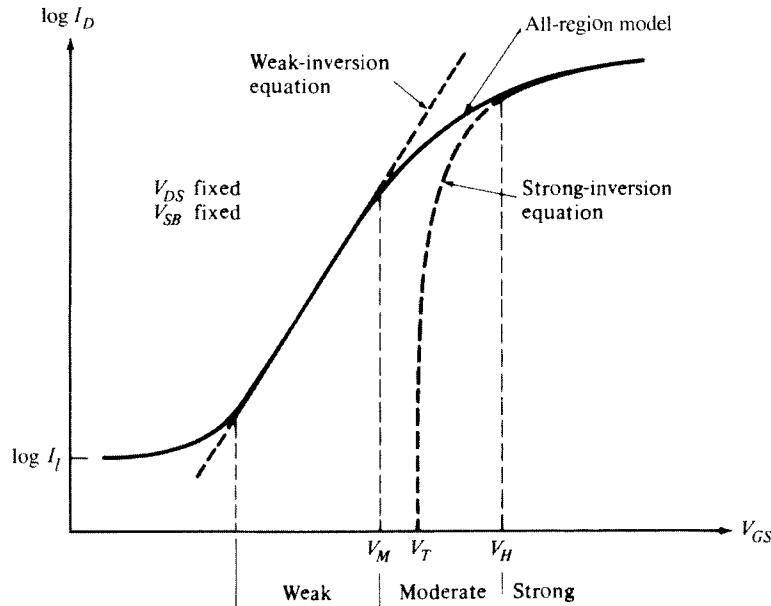


FIGURE 4.31
 $\log I_D$ vs. V_{GS} for fixed V_{DS} and V_{SB} , and comparison to weak- and strong-inversion equations.

expressed in “mV/decade”). It is easily seen that S is the inverse of the slope of the weak-inversion part of the plot in Fig. 4.31, i.e.,

$$S = \frac{dV_{GS}}{d(\log I_{DS})} \quad (4.8.15)$$

which, using (4.8.11), gives

$$S = 2.3n\phi_i \quad (4.8.16)$$

We emphasize that (4.8.11) is only approximate, and that I_{DS} is not exactly an exponential in weak inversion, as is evident from the more accurate equations, (4.8.6) and (4.8.7). The minute deviations from exponentiality are important to keep in mind when small-signal parameters are considered (Chap. 7).

In Fig. 4.31, we also show the behavior of the strong-inversion model of Sec. 4.7.3, as well as of the all-region model of Sec. 4.3.

4.9 Moderate-Inversion and Single-Piece Models

As defined in Table 4.1, a MOS transistor is said to “operate in moderate inversion” if its more heavily inverted channel end is moderately inverted.

The I_{DS} - V_{DS} characteristics of the transistor in this region have a shape roughly similar to that in strong inversion, but are not described accurately by strong inversion equations since, as shown in Sec. 4.3, in moderate inversion, both drift *and* diffusion contribute significantly to the value of the drain current. Convenient simplifications are not known for this region. One can use the all-region models of Secs. 4.3 and 4.4, or use semiempirical models.

The moderate-inversion region used to be completely ignored in the literature. This region was instead assumed to be the bottom of the strong-inversion region, and the strong-inversion equations of Sec. 4.7 were used in it. However, not even a single point in the channel is strongly inverted in this region; hence, equations such as (4.7.1) will be in error, and so will the models based on them. In an attempt to “stretch” the validity of the strong-inversion equations, sometimes the parameters in them are made functions of the gate voltage, for example, ϕ_0 or V_T in (4.7.19).³⁴ In developing expressions for the moderate-inversion region, care should be taken that the drain current be continuous at the boundaries between weak and moderate inversion, and between moderate and strong inversion. It is also very important to make sure that the *derivatives* of the drain current with respect to any terminal voltage are *also* continuous at these boundaries; otherwise, severe errors can result in the evaluation of small signal parameters at these boundaries (Chap. 7), and convergence problems can occur during computer simulation (Chap. 10).

Due to the difficulties mentioned previously in developing moderate inversion expressions, and the need for continuity of I_{DS} and its derivatives, several semiempirical “single-piece” expressions have been proposed.^{83,84,89,101–103,108,109} These are designed to do an acceptable, or even good, job in moderate inversion and to reduce to acceptable weak- and strong-inversion expressions in the corresponding regions.

We show here an example of such a model, which uses interpolation functions^{104,108} and is source-referenced:¹¹¹

$$I_{DS} = \frac{W}{L} \mu C'_{ox}(2n) \phi_t^2 \left\{ \left[\ln \left(1 + e^{(V_{GS}-V_t)/[2n\phi_t]} \right) \right]^2 - \left[\ln \left(1 + e^{(V_{GS}-V_t-nV_{DS})/[2n\phi_t]} \right) \right]^2 \right\} \quad (4.9.1)$$

This model makes explicit use of the threshold voltage V_T , given for each V_{SB} value by the familiar expression (4.7.20). Since V_T is explicitly used in the model, if desired, one can use for this quantity a value found from more elaborate V_T models, e.g., to take into account effects related to ion implantation (Chap. 9) or to small-channel dimensions (Chap. 5). The model is valid in all regions, including moderate inversion for both nonsaturation and saturation. In weak inversion, with both exponentials much smaller than 1 and using $\ln(1+x) \approx x$ for $|x| \ll 1$, the model can be shown to reduce to (4.8.11) within a multiplicative constant. Consider now strong inversion. Here the first exponential is much larger than 1. If V_{DS} is very small

(i.e., we are in deep nonsaturation), the second exponential is much larger than 1, too. Using then the approximation $[\ln(1 + e^y)]^2 \approx (\ln e^y)^2 = y^2$, valid for $e^y \gg 1$, and after some algebra, we obtain (4.7.24a), with n in lieu of α . If, in strong inversion, V_{DS} is raised, the second exponential becomes negligible, whereas the first exponential remains much larger than 1. We are then in saturation, and the expression simplifies to (4.7.24b), again with n in lieu of α .[†]

Independent of the region of inversion, as V_{DS} is raised the second exponential in (4.9.1) becomes negligible. We thus have

$$I_{DS} = \frac{W}{L} \mu C'_{ox} (2n) \phi_i^2 \left[\ln \left(1 + e^{(V_{GS} - V_T)/(2n\phi_i)} \right) \right]^2 \text{ saturation, all regions of inversion} \quad (4.9.2)$$

In analog circuit design, where most devices operate in saturation, this expression is useful for approximate calculations in any region of inversion. However, it is mostly appropriate for low-voltage work, unless n is allowed to reduce to α as strong inversion is entered.

In circuit design, interpolation models of the type previously discussed can be a significant improvement⁹² compared with models that ignore moderate inversion altogether. They achieve a balanced compromise between accuracy and simplicity and are sometimes preferred to complete all-region models (see Secs. 4.3 and 4.4) for reasons of computational efficiency. Details of common interpolation models are discussed in Chap. 10.

4.10 Source-Referenced vs. Body-Referenced Modeling

In this chapter we have discussed a variety of models. Some were symmetric with respect to the substrate and gate, and the independent variables in them were V_{GB} , V_{DB} , and V_{SB} . Others were source-referenced, with V_{GS} , V_{DS} , and V_{SB} as the independent variables. Several advantages are often claimed for each approach, but it should be noted that, in principle, the two approaches are, or can be, equivalent. This is obvious, since a body-referenced model can be converted to a source-referenced one by replacing V_{GB} by $V_{GS} + V_{SB}$ and V_{DB} by $V_{DS} + V_{SB}$. Thus, all “advantages” of the body-referenced model would carry over to the new “source-referenced” model. However, in practice each approach is often associated with a different set of practices, and it is those practices that are responsible for some of the ways each approach is viewed. Let us now list the advantages of each approach. We will take this opportunity to include issues that pertain not only to modeling, but also to points of view of the MOSFET as a circuit element.

[†] The accuracy of the model can be extended if one uses $n = 1 + \gamma/(2\sqrt{\phi + V_{SB}})$, where ϕ is a quantity that is made to increase as V_{GS} is raised, so that n changes from the value in (4.8.14) to the value in (4.7.28).^{77,107}

Advantages of source-referenced models

1. Physically and intuitively, the real “driving forces” that cause the current to flow in a MOSFET are V_{GS} (to create the inversion layer) and V_{DS} (to apply a potential across that layer). They are not V_{GB} and V_{DB} per se. For a device with a desirably very small body effect, variations in V_{GB} , with V_{GS} constant, have a very small effect on I_{DS} , whereas variations of V_{GS} , with V_{GB} fixed, have a major effect. In principle, for a substrate with negligible doping concentration the B terminal would become irrelevant, and it would not make much sense to refer the voltages to it. Some silicon-on-insulator devices have no body connection, and the same is true for multiple-gate devices (which are under development; see Chap. 5). The semiconducting substrate is a layer we would ideally do without; referring all voltages to a terminal we would rather see disappear is not easily acceptable.
2. In the history of electronics, all previous major devices (vacuum tubes and bipolar transistors) were usually viewed with “input” voltages corresponding to V_{GS} (the grid-cathode voltage in tubes, and the base-emitter voltage in bipolar transistors). There was no “fourth” terminal (unless the parasitic substrate on which bipolar devices sit is counted). Using V_{GS} as the variable for MOS devices is a natural carryover, and the intuition developed over many decades of experience with other devices can be extended to use with MOSFETs. All these types of devices can then be viewed in a unified way for the purposes of circuit design (this can be especially convenient in BiCMOS technology, where both bipolar and MOS transistors must be dealt with).
3. By referring all voltages to the source, and going through a linearization of the bulk charge, the threshold voltage occurs naturally in the device equations, as we have seen [see, for example, (4.7.24)]. The threshold voltage is one of the most extensively studied and characterized MOSFET parameters, and it is convenient to have it as an explicit model parameter. In particular, its functional dependence on V_{SB} has been extensively studied, and the associated expressions for $V_T(V_{SB})$ can be used directly in the model equations.
4. As we have seen, the body acts as a second gate and is also referred to as a “back gate” for this reason. It is thus natural to treat it similar to the way we treat the gate. In fact, there are analog circuits in which the body is actually used as a second gate (given that we cannot get rid of the body, we might as well use it for something). With the source as the reference, the “inputs” now are V_{GS} and V_{BS} . This can be handled naturally by a source-referenced model. For a body-referenced model, using the reference (or “ground”) node B as an input is not as convenient.
5. Effects related to carrier transport from source to drain, such as velocity saturation (Chap. 5), are physically more natural to formulate in terms of V_{DS} , rather than in terms of V_{DB} and (separately) of V_{SB} .
6. Although, if a transistor is laid out symmetrically, it has symmetric characteristics, often transistors are not laid out symmetrically. For example, some gate geometries are not rectangular, and the source is made to partially surround the drain (or vice versa). In such cases, the plots of current vs. drain voltage may not be exactly the same as the plots of current vs. source voltage, and the drain capacitances are certainly very different from the source capacitances. In such

cases, keeping track of which terminal is the source is important, and it may be necessary to extract parameters with this in mind. Such distinction is natural to a source-referenced model. A body-referenced model can also be employed, but it will have to be made asymmetric for this purpose.

7. For work at very high frequencies, several modeling approaches have been proposed for source-referenced models (Chap. 8), whereas very little has been done in this respect for body-referenced models. Standard radio-frequency (RF) design techniques and measurements are also source-reference-based.

Advantages of body-referenced models

1. The symmetry inherent in body-referenced models is physically correct for symmetric MOSFETs.
2. The symmetrical nature of symmetrically laid out MOSFETs is sometimes taken advantage of in analog circuits. A body-referenced model is well suited to use in such cases.
3. In body-referenced long-channel models, current saturation is attained at values of V_{DB} that are independent of V_{SB} , as expected from basic considerations.
4. Weak inversion is well handled by body-referenced models. For example, the quantity ψ_{sa} in Sec. 4.8 depends only on V_{GB} . A design culture has developed using such models for circuits operating in weak inversion.³⁷
5. Some effects, like effective mobility, which are associated with transverse fields, can be well handled in a body-referenced model.
6. A body-referenced model does not, by its nature, exhibit discontinuities of I_{DS} or its derivatives at $V_{DS} = 0$. This property is critical for simulation of certain types of circuits. In contrast, source-referenced models can present such problems. Note, though, that effects related to carrier transport from source to drain (such as velocity saturation, Chap. 5) are often handled through V_{DS} even in body-referenced models.

Despite these differences, and as already indicated in the beginning of this subsection, source-referenced and body-referenced models can in principle be equivalent. Either approach can do a good job, and the user of a model implemented in a circuit simulator does not even need to know whether the internal structure of the model is source-referenced or body-referenced. It is the practices usually associated with each type, as well as differences in circuit design cultures, that are responsible for preferring one or the other type of model in several instances. We find it appropriate to discuss both types of model in this book.

4.11 Effective Mobility

In Sec. 1.3.2, we briefly looked at the mobility of electrons in the bulk. We mentioned that the value of this quantity is determined by several scattering mechanisms, through which the electrons exchange momentum and kinetic energy with their environment. One such mechanism is due to the energy of lattice vibrations, which is modeled by using energy quanta called *phonons*. The electron-phonon interaction

is referred to as *phonon scattering*. Another mechanism is related to the electrically charged ionized impurity atoms and is called *Coulomb scattering*.

In a MOS transistor, electrons in the inversion layer flow near the semiconductor-oxide interface (the “surface” of the semiconductor). The electric field component perpendicular to the direction of current flow (referred to as the *transverse*, or *normal*, component) tends to accelerate the inversion layer electrons toward the surface and subjects them to additional scattering.^{17,112} Now there is Coulomb scattering not only due to ionized impurity atoms, but also due to interface trapped charges and to charges trapped within the oxide. It is interesting to note that the inversion layer charge itself, if it is significant, tends to partly “screen” itself from the effects of Coulomb scattering. Additional scattering occurs due to *surface roughness*. All these tend to lower the mobility of electrons in the inversion layer (called *surface mobility*) to values smaller than the bulk mobility considered in Sec. 1.3.2.[†]

Surface mobility has been the subject of many studies.^{17,55,76,112–158} In most of these, surface mobility is studied as a function of an effective transverse field; while this quantity can be defined in more than one way, we will assume for now that it is the mean between the values of the transverse field at the top and at the bottom of the inversion layer. A typical behavior of surface mobility vs. effective transverse field is as shown in Fig. 4.32. Qualitatively, this behavior can be described by taking into

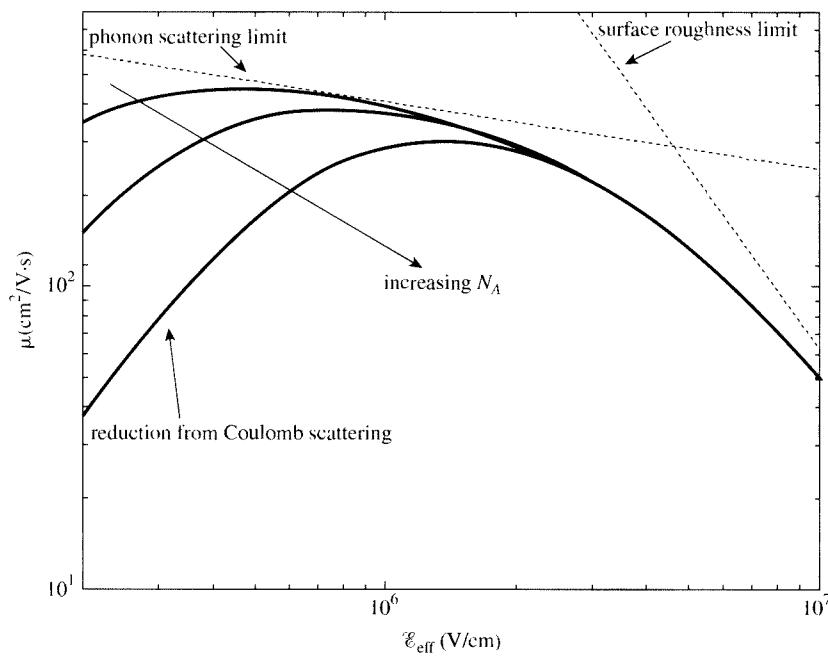


FIGURE 4.32
Surface mobility vs. average transverse field.

[†] In short-channel devices, things get further complicated due to the large field along the surface (the parallel, or “longitudinal” field); this will not concern us until we get to Chap. 5.

account the fact that, while all the preceding mechanisms may be present simultaneously, and while each tends to limit the surface mobility value, different mechanisms tend to dominate the behavior in different ranges of field values. At low fields, Coulomb scattering due to impurity atoms and oxide charges tends to be dominant in limiting mobility.^{140,150–152} As the field is increased, the inversion layer charge becomes significant and tends to screen the impurity and oxide charges, thereby lessening their effect; this is why the mobility shows an initial increase in the figure. This trend does not continue as the field is increased further, since phonon scattering becomes dominant and mobility begins to drop. Finally, at still higher fields, the strong “pull” of electrons toward the surface makes surface roughness the dominant scattering mechanism, and mobility begins to drop faster, as seen toward the right.

Increasing the doping concentration tends to shift the point where Coulomb scattering becomes nondominant to higher field values. This is because there are now more ionized impurity atoms, which makes Coulomb scattering more prominent; also, higher substrate dopings mean that a larger substrate charge must be depleted [see (3.2.17a)] before an inversion layer can be formed, which means that the inversion layer charge will not become significant (so that it can screen the effects of Coulomb scattering) until higher field values.¹⁴⁶ Thus the peak of the mobility curve shifts to the right if doping is increased, as shown. We should note here that the literature is not consistent as to how serious the mobility drop is at low electric fields, and in fact some or all of this drop, as observed by some authors, has been claimed by others^{127,143} to be the result of experimental error, or of errors in data interpretation.

It has been reported that to the right of their peak, the curves for different doping concentrations and different oxide thicknesses tend to merge and to follow a universal curve,^{118,120,121,123,128,146,158} although surface orientation and surface preparation can make a difference.¹⁴⁷ A simple approximate equation for this “universal” behavior will be presented later in this section.

Effective Transverse Field and Mobility Several different ways have been used to define $\mathcal{E}_{y,\text{eff}}$, the “effective” transverse field. For example, this quantity is sometimes assumed to be a simple arithmetic mean:^{29,114}

$$\mathcal{E}_{y,\text{ave}} = \frac{\mathcal{E}_{ys} + \mathcal{E}_{yb}}{2} \quad (4.11.1)$$

where \mathcal{E}_{ys} is the value of the normal field at the surface and \mathcal{E}_{yb} is its value “just below” the inversion layer. The electric field at the surface can be easily related to the total charge per unit area below the surface from basic electrostatics (Appendix A):[†]

$$\mathcal{E}_{ys} = -\frac{Q'_I + Q'_B}{\epsilon_s} \quad (4.11.2)$$

[†] Due to the gradual channel approximation (Sec. 4.1), the electric field component parallel to the direction of current flow is assumed negligible compared with the transverse component. Hence, a one-dimensional analysis can be used.

where Q'_I and Q'_B are the inversion layer and depletion region charges per unit area, respectively, and ϵ_s is the permittivity of the semiconductor. Similarly, the field just below the inversion layer can be determined by assuming a very thin inversion layer, so that the total charge per unit area below it is practically all of Q'_B :

$$\mathcal{E}_{yb} = -\frac{Q'_B}{\epsilon_s} \quad (4.11.3)$$

Using the preceding equations in (4.11.1), we obtain

$$\mathcal{E}_{y,\text{eff}} = -\frac{Q'_B + 0.5Q'_I}{\epsilon_s} \quad (4.11.4)$$

More elaborate definitions of $\mathcal{E}_{y,\text{eff}}$, taking into account weighting by the density of the inversion layer vs. depth, give a similar result, only with the factor 0.5 replaced by a quantity η_E :

$$\mathcal{E}_{y,\text{eff}} = -\frac{Q'_B + \eta_E Q'_I}{\epsilon_s} \quad (4.11.5)$$

with η_E dependent on crystal orientation, with values between 0.25 and 0.6 for electrons, and 0.33–0.4 for holes.^{57,123,137,146,147,149}

The surface mobility can then be expressed in the form

$$\mu = \mu(\mathcal{E}_{y,\text{eff}}) \quad (4.11.6)$$

where $\mu(\mathcal{E}_{y,\text{eff}})$ is a function that approximates the behavior exhibited in Fig. 4.32. Such a function can in principle be derived from the so-called Matthiessen rule for combining mobility contributions, which is^{16,65}

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} \quad (4.11.7)$$

with μ_1 the mobility due to Coulomb scattering, μ_2 the mobility due to phonon scattering, and μ_3 the mobility due to surface roughness scattering. Unfortunately, the resulting functions are complicated, and semiempirical approximations are often used. The interested reader is referred to the literature.^{113,121,141,145,159–161}

Effective Mobility and Drain Current We will now study ways to incorporate the previously discussed behavior in MOS transistor models. Recall that, in steady state, the drain current is equal to the current at any point in the channel; including drift and diffusion components, we have (4.3.4), repeated below:

$$I_{DS} = \mu W(-Q'_I) \frac{d\psi_s}{dx} + \mu W \phi_i \frac{dQ'_I}{dx} \quad (4.11.8)$$

In Sec. 4.3, we integrated the preceding equation from $x = 0$ to $x = L$, obtaining

$$I_{DS} = \frac{W}{L} \left[\int_{\psi_0}^{\psi_L} \mu(-Q'_I) d\psi_s + \phi_t \int_{Q'_{I0}}^{Q'_L} \mu dQ'_I \right] \quad (4.11.9)$$

We then made the assumption that μ was constant along the channel; thus μ could be moved outside the integral, and we obtained

$$I_{DS} = \frac{W}{L} \mu \left[\int_{\psi_0}^{\psi_L} (-Q'_I) d\psi_s + \phi_t (Q'_L - Q'_{I0}) \right] \quad (4.11.10)$$

However, we now know that μ depends on the transverse field. Since this field, in general, varies along the channel, μ will also vary. If its variation is not negligible, one *cannot* move μ outside the integral. One can then attempt to determine μ as a function of the variables of integration in (4.11.9), and try to include this function in the integration. This is very difficult analytically.²⁴ Another approach taken is to define an *effective* mobility μ_{eff} such that, when used in the expression

$$I_{DS} = \frac{W}{L} \mu_{\text{eff}} \left[\int_{\psi_0}^{\psi_L} (-Q'_I) d\psi_s + \phi_t (Q'_L - Q'_{I0}) \right] \quad (4.11.11)$$

the current is predicted correctly.[†] Since the electric field at each point in the channel and, thus, μ in (4.11.9) depend on the terminal voltages, one should expect that to make (4.11.11) give the same result as (4.11.9), μ_{eff} should be made a function of the terminal voltages. This function is often determined empirically, but sometimes a more analytical approach can be used, as will be shown. Comparing (4.11.11) to (4.11.10), we see that, when the effective-mobility approach is taken, I_{DS} will be given by the same expressions found in Secs. 4.3 and 4.4, if in them μ is replaced by μ_{eff} :

I_{DS} , including mobility dependence on normal field	$= I_{DS}$, assuming constant mobility	$\boxed{\mu = \mu_{\text{eff}}}$
---	--	----------------------------------

(4.11.12)

We now describe a way to determine μ_{eff} in principle; this is an extension⁸² of classical techniques to the all-region model. If both sides of (4.11.8) are divided by μ , and then integration is performed from $x = 0$ to $x = L$, we obtain

$$I_{DS} \int_0^L \frac{dx}{\mu} = W \left[\int_{\psi_0}^{\psi_L} (-Q'_I) d\psi_s + \phi_t (Q'_L - Q'_{I0}) \right] \quad (4.11.13)$$

Solving this for I_{DS} , we find an expression that is exactly of the form of (4.11.11), with

$\mu_{\text{eff}} = \frac{1}{\frac{1}{L} \int_0^L \frac{dx}{\mu}}$
--

(4.11.14)

As a check we note that, if μ is assumed constant, the preceding equation gives $\mu_{\text{eff}} = \mu$, as expected.

[†] A formal definition of μ_{eff} can be obtained by equating the right-hand sides of (4.11.11) and (4.11.9), and solving for μ_{eff} . However, this is not useful for evaluating μ_{eff} since, if one can evaluate the integrals in (4.11.9) with adequate simplicity, then μ_{eff} is not needed; one would, in this case, have found I_{DS} directly from (4.11.9).

For the real case where μ depends on \mathcal{E}_y , one can in principle express \mathcal{E}_y in terms of the surface potential, and relate the latter to the position x through (4.3.20). Whether this will give manageable results from (4.11.14) depends on the form assumed for $\mu(\mathcal{E}_y)$, and on the assumptions made along the way.

We will now give a simple example. Assume that we are interested in fields that are not extremely low, so that the Coulomb scattering effect is negligible. Experimental data for this case is sometimes modeled by the following relation:^{113,114}

$$\mu = \frac{\mu_0}{1 + a_\theta \mathcal{E}_{y,\text{eff}}} \quad (4.11.15)$$

where μ_0 and a_θ depend on temperature and can be considered fitting parameters.[†] For practical conditions, μ_0 may be roughly half of the bulk mobility (as evaluated for lightly doped substrates), or less for heavily doped substrates; typical values are in the range of 400–600 cm²/(V·s) for electrons and 100–200 cm²/(V·s) for holes, assuming no mechanical stress (the introduction of stress can increase these values, as discussed in Sec. 9.6). The quantity a_θ can be of the order of 10⁻⁶ cm/V at room temperature. Using (4.11.5) in the above equation we get

$$\mu = \frac{\mu_0}{1 - \frac{a_\theta}{\epsilon_s} (Q'_B + \eta_E Q'_I)} \quad (4.11.16)$$

Substituting this in (4.11.14), we have

$$\mu = \frac{\mu_0}{\frac{1}{L} \int_0^L \left[1 - \frac{a_\theta}{\epsilon_s} (Q'_B + \eta_E Q'_I) \right] dx} \quad (4.11.17)$$

Integrating the individual terms in the denominator, we can write this equation as follows:

$$\mu_{\text{eff}} = \frac{\mu_0}{1 - \frac{a_\theta}{\epsilon_s} \left[\frac{1}{L} \int_0^L Q'_B dx + \eta_E \frac{1}{L} \int_0^L Q'_I dx \right]} \quad (4.11.18)$$

which can be written as

$$\mu_{\text{eff}} = \frac{\mu_0}{1 - (a_\theta / \epsilon_s) (\bar{Q}'_B + \eta_E \bar{Q}'_I)} \quad (4.11.19)$$

with \bar{Q}'_B and \bar{Q}'_I being the average depletion region and inversion layer charges, respectively, over the length of the channel:

[†] This equation can be made plausible under certain physical assumptions.¹¹³ A better fit to experimental data is claimed to be obtained by using $\mu = \mu_0 / (1 + a_1 \mathcal{E}_{y,\text{eff}}^{b_1} + a_2 \mathcal{E}_{y,\text{eff}}^{b_2})$, where the first field term models phonon scattering and the second surface roughness scattering. The quantities a_1 and a_2 are fitting parameters, b_1 is close to 0.3, and b_2 is approximately 2 for electrons and 1 for holes.^{141,145} We note, however, that the stronger degradation of mobility at very high fields is not universally accepted.¹²¹

$$\bar{Q}'_B = \frac{1}{L} \int_0^L Q'_B dx \quad (4.11.20)$$

$$\bar{Q}'_I = \frac{1}{L} \int_0^L Q'_I dx \quad (4.11.21)$$

We have expressions for the charges in the integrands of the preceding two equations for various models, but they are functions of the surface potential, ψ_s , rather than the position x . To be able to integrate we need a change of variables involving $d\psi_s/dx$. This can be obtained from an expression of surface potential as a function of position, as in (4.3.20). Unfortunately, such a calculation can be complicated.[†] Thus, approximations are often used. One such approximation^{86,142} uses simple arithmetic means instead of the preceding averages

$$\bar{Q}'_B = \frac{Q'_{B0} + Q'_{BL}}{2} \quad (4.11.22)$$

$$\bar{Q}'_I = \frac{Q'_{I0} + Q'_{IL}}{2} \quad (4.11.23)$$

where the subscripts 0 and L indicate the values of the charge densities at the source and drain ends of the channel, respectively.

As already mentioned, (4.11.15), and thus eqs. (4.11.16)–(4.11.19) derived above, are meant for cases in which Coulomb scattering effects are not dominant. At very low fields, where such effects are pronounced, as well as at very high fields, more complicated expressions are used.^{133,137,141,145–147,160–162} These are sometimes derived starting from (4.11.7). The inclusion of Coulomb scattering effects has been modeled¹⁶¹ by adding to the denominator of (4.11.19) a term proportional to $[\bar{Q}'_B / (\bar{Q}'_B + \bar{Q}'_I)]^2$; as the electric field is raised starting from very low values, this term decreases due to the increase in the inversion charge magnitude, thus making the denominator smaller and the effective mobility larger. At higher fields, this term becomes small and it no longer affects the mobility appreciately, leaving the rest of the terms, related to phonon scattering, to dominate. Often, though, mobility modeling is limited to the previous approach for simplicity.

Surface Potential-Based Models The quantities in (4.11.22) and (4.11.23) can be easily calculated for the various all-region models we have presented. For all models, (4.3.10) with (4.11.23) gives

$$\bar{Q}'_I = -C'_{ox} \left(V_{GB} - V_{FB} - \frac{\psi_{s0} + \psi_{sL}}{2} \right) - \bar{Q}'_B \quad (4.11.24)$$

and \bar{Q}'_B can be found by the expressions for Q'_B for each model. This gives

[†] We will see examples of charge density integration in Chap. 6.

Complete all-region model (Sec. 4.3):

$$\bar{Q}'_B = -C'_{\text{ox}} \gamma \frac{\sqrt{\psi_{s0}} + \sqrt{\psi_{sL}}}{2} \quad (4.11.25)$$

Body-referenced simplified all-region model (Sec. 4.4.2):

$$\bar{Q}'_B = -C'_{\text{ox}} \gamma \sqrt{\psi_{sm}} \quad (4.11.26)$$

Source-referenced simplified all-region model (Sec. 4.4.3):

$$\bar{Q}'_B = -C'_{\text{ox}} \left[\gamma \sqrt{\psi_{s0}} + (\alpha - 1) \frac{\psi_{sL} - \psi_{s0}}{2} \right] \quad (4.11.27)$$

Corresponding expressions for charge-based models (Sec. 4.4.4) can be found in the references.^{85,163}

Strong Inversion In strong-inversion nonsaturation, the corresponding quantities can trivially be found from the preceding three equations by using (4.7.1), repeated here:

$$\psi_{s0} = \phi_0 + V_{SB} \quad (4.11.28a)$$

$$\psi_{sL} = \phi_0 + V_{DB} \quad (4.11.28b)$$

We will consider in more detail the source-referenced model as an example. Using these potentials in (4.11.27) and (4.11.24) and the result in (4.11.16), and assuming $\eta_E = 0.5$, we obtain

$$\mu_{\text{eff}} = \frac{\mu_0}{1 + \theta [V_{GS} - V_T + 2\gamma \sqrt{\phi_0 + V_{SB}} - (1 - 0.5\alpha)V_{DS}]} \quad (4.11.29)$$

where V_T is given by (4.7.20) and

$$\theta = \frac{a_\theta}{2\epsilon_s} C'_{\text{ox}} \quad (4.11.30)$$

As seen, the effect of the gate voltage is dominant. For this reason, it is sometimes said that μ_{eff} “depends on the gate field.” It is more correct, though, to say that μ_{eff} depends on the transverse field, which, in turn, depends on all terminal voltages, as has been seen. We should note here that in the denominator of the expression for μ_{eff} another term proportional to V_{DS} is often included to model “velocity saturation” effects. This is not related to the V_{DS} terms in the previous equations and will be considered in detail in Chap. 5.

One must use the approach indicated in Sec. 4.7 to extend the characteristics into the saturation region, starting by determining the value of V_{DS} at which saturation

is obtained. This value, denoted by V'_{DS} , is the value of V_{DS} at which $dI_{DSN}/dV_{DS} = 0$. Note that, because of the presence of V_{DS} in the expression for μ_{eff} , a new differentiation is required to obtain the value of V'_{DS} . The V'_{DS} expressions of Sec. 4.7 will not be valid here, since they were obtained by assuming a constant mobility. This results in additional complications that are often avoided by neglecting the dependence of μ_{eff} on V_{DS} . For example, consider (4.11.29). In that equation the last term within the brackets in the denominator is usually dropped.[†] In addition, the term dependent on V_{SB} is sometimes replaced by a term linear in V_{SB} . Thus the following form has been suggested:^{25,106‡}

$$\mu_{\text{eff}} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T) + \theta_B V_{SB}} \quad (4.11.31)$$

By now several approximations have been made, and one should expect that the values of μ_0 , θ , and θ_B used in the preceding equation may have to be determined by fitting to measurements to minimize the error. Typical values for μ_0 are 400 to 600 $\text{cm}^2/(\text{V} \cdot \text{s})$ for *n*-channel devices at room temperature. For *p*-channel devices, μ_0 is typically smaller by a factor of about 3. The parameter θ is of the form $\beta_\theta/t_{\text{ox}}$ with t_{ox} the oxide thickness, and β_θ typically 0.5 to 2 $\text{nm} \cdot \text{V}^{-1}$.[§]

The value of θ_B is usually small (for example, a few hundredths of 1 V^{-1}), and sometimes the V_{SB} -dependent term in (4.11.31) is omitted altogether. This can cause problems, as then increasing V_{SB} would imply an increase in μ_{eff} , since V_T would be increasing owing to the body effect. However, μ_{eff} should actually *decrease* with increasing V_{SB} . This is most easily seen from (4.11.29) and is also expected intuitively. Let us refer all voltages to the source. Then increasing V_{SB} means making the voltage at the body terminal more negative. This would tend to increase the normal field and “push” the electrons more toward the surface, which is the same effect an increase in the gate voltage would have; thus μ_{eff} should be expected to decrease. This is verified experimentally¹⁵⁸ and is found to be more evident in *p*-channel devices (Sec. 4.15).¹⁰⁶ Thus omitting the V_{SB} -dependent term in (4.11.31) can have a serious effect, especially in the small-signal modeling of analog circuits (Chaps. 7 and 8). If, in addition to omitting the V_{SB} -dependent term in (4.11.31) V_T is replaced by V_{T0} [as given by (4.7.20)],¹⁶⁴ μ_{eff} appears independent of V_{SB} . Although this is not done in CAD models, it is still done in some circuits books.

The effect of mobility dependence on V_{GS} is shown in Fig. 4.33. I_{DS} is plotted vs. V_{GS} for a fixed V_{SB} and very small V_{DS} . Then, as seen from (4.7.24), we will have $I_{DS} \approx (W/L)\mu C'_{\text{ox}} V_{DS}(V_{GS} - V_T - \frac{\alpha}{2}V_{DS})$. If μ is constant, a straight line is obtained in

[†] Often the effect of this term is absorbed into a similar term that results from “velocity saturation” effects; these are considered in Chap. 5.

[‡] A term containing V_{GS}^2 is sometimes added^{133,137} to the denominator of μ_{eff} , to help model the strong decrease of mobility due to surface roughness at high fields, such as those encountered in transistors with very thin oxides. Also, terms inversely proportional to V_{GS}^2 are sometimes used to model the effect of Coulomb scattering at very low fields.^{161,162}

[§] The value of θ is sometimes adjusted to help model an effect quite unrelated to the field dependence of mobility, i.e., the effective drain-source voltage reduction due to the voltage drop across the series resistance of the source and drain regions. This effect is discussed in Sec. 4.12.

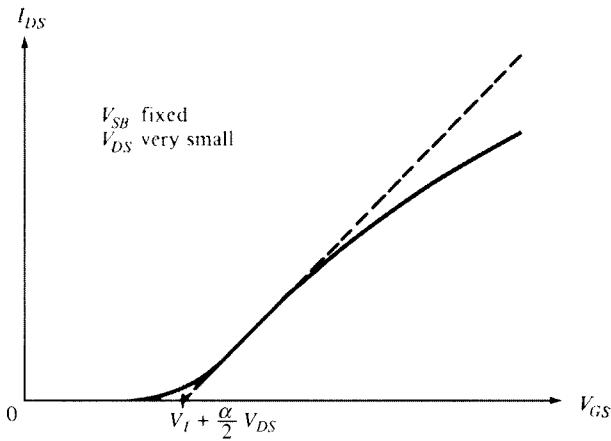


FIGURE 4.33
 I_{DS} vs. V_{GS} for fixed V_{SB} and fixed, very small V_{DS} for constant effective mobility and neglecting moderate and weak inversion (broken line), and for V_{GS} -dependent effective mobility and including moderate and weak inversion (solid line).

strong inversion, as shown by the broken line (the bottom, curved part of the solid curve is caused by moderate inversion). If μ is replaced by μ_{eff} , which varies with V_{GS} as previously suggested, the plot becomes as shown by the solid line. In some devices this effect is so strong that one never really gets to see a straight-line part in the characteristic. Note that very small V_{DS} is not a necessary condition for obtaining a straight line in *nonsaturation*. However, such plots are usually obtained for very small V_{DS} since, then, nonsaturation is observed even at small V_{GS} values, at which one can hope to see the initially straight part of the curve where mobility degradation is not strong. Also, for very small V_{DS} , the intercept of an extension of the straight-line part with the horizontal axis is approximately V_T . This is a convenient method for determining the threshold voltage experimentally. More details are given in Chap. 10.

The parameter μ_{eff} is often simply referred to as the *mobility* instead of the more complete term *effective mobility* and is denoted by μ for simplicity. For convenience, we will adopt this notation in the rest of this book. No confusion should arise because it should be understood that, every time μ is encountered in expressions that give the drain current as a function of the terminal voltages, it will be the *effective* mobility.

In saturation, it can be seen that the V_{GS} -dependent term in (4.11.31) will tend to counteract the numerator in (4.7.24b), resulting in a I_D - V_{GS} dependence, which is no longer square-law.

As already pointed out, the results obtained in the previous example are valid in the strong-inversion region, where the normal electric field is not very small and (4.11.15) is valid. At very low fields (e.g., corresponding to moderate or weak inversion) this equation is no longer claimed to be valid.¹⁵³ It has been suggested that the electron density in the inversion layer at such fields is no longer large enough to “screen” the ionized acceptor atoms, or that it can show significant fluctuations due to the presence of localized charges at the oxide-semiconductor interface.^{113,116}

Devices have been reported to exhibit a drop in effective mobility when operating in weak inversion,^{112,113,115,116,148} although not all results reported in the literature show such behavior.^{127,153}

For all-region models, expressions more complicated than (4.11.29) are used. Nevertheless, terms such as the ones in the denominator of that equation still occur in such expressions.

Mobility can be enhanced by stress; this effect is taken advantage of in modern technologies, and is described in Chap. 9.

For readers who plan to search the literature on the important topic of effective mobility, we would like to offer some warnings:

1. What is often called “effective mobility” in the literature is not meant in the general sense we have used it thus far, which allowed us to write (4.11.12) for all terminal voltage combinations, including the effect of large V_{DS} on the transverse electric field. Rather, what is often meant by this term is only the mobility at a point in the channel, and it is measured indirectly, by applying a very small V_{DS} (e.g., 10 mV) on a transistor, in an attempt to make the channel uniform.
2. Some of the data on effective mobility vs. field behavior published in the literature are distorted due to experimental errors.
3. Some studies are inconclusive, and different studies contradict each other; we gave some indication of this wherever possible.
4. Many of the detailed results on effective mobility vs. field are not usable in developing a device *I-V* model of manageable simplicity. Thus, one often ends up using simple approximations such as those we have presented. One should not be surprised, therefore, if these do not provide very accurate results, or if the parameters used must be given nonphysical values in order to help make up for the lack of accuracy.

4.12 Effect of Extrinsic Source and Drain Series Resistances

So far we have concentrated on the device part *between* source and drain, containing the inversion layer, the depletion region, the oxide, and the gate. This part is called the *intrinsic* part and is the part mainly responsible for transistor action. The rest of the device constitutes the *extrinsic* part and is responsible for parasitic effects, which can limit overall performance. For the dc situations we are considering here, the relevant elements of the extrinsic part are the series resistances of the source and drain n^+ regions.[†] These resistances depend on the fabrication and geometrical details of these regions, as well as on the size and number of contacts made on them; the evaluation of these resistances can be found in Sec. 6.8.

[†] There is also resistance associated with the gate and substrate regions. Because we are assuming dc operation and no dc leakage current through the gate and substrate, there is no voltage drop across such resistances, and thus they are not relevant here. These resistances will be considered in Sec. 6.8.2.

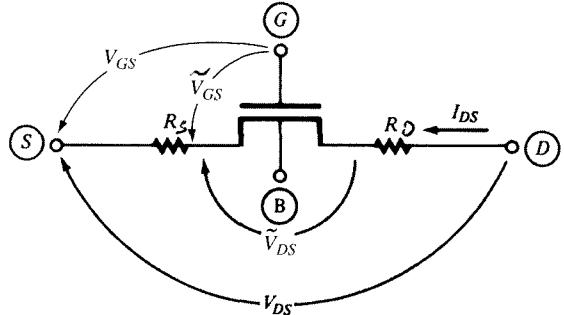


FIGURE 4.34

A MOS transistor with the associated source and drain resistances.

Let us denote the extrinsic source and drain series resistances by R_S and R_D , respectively. These resistances are traversed by the drain-source current, thus resulting in voltage drops that must be taken into account, as shown in Fig. 4.34. The effective gate-source voltage, as felt by the intrinsic part of the device, is no longer V_{GS} but rather $\tilde{V}_{GS} = V_{GS} - R_S I_{DS}$; similarly, the effective drain-source voltage is not V_{DS} , but rather $\tilde{V}_{DS} = V_{DS} - (R_S + R_D) I_{DS}$. To see the effect these have, we now perform an approximate calculation.¹⁶⁵⁻¹⁷⁰ In the simple strong-inversion nonsaturation model in (4.7.24a), we factor out V_{DS} and use the preceding \tilde{V}_{GS} and \tilde{V}_{DS} in lieu of V_{GS} and V_{DS} to obtain

$$I_{DS} = \frac{W}{L} \mu C'_{ox} \left\{ (V_{GS} - R_S I_{DS}) - V_T - \frac{\alpha}{2} [V_{DS} - (R_S + R_D) I_{DS}] \right\} [V_{DS} - (R_S + R_D) I_{DS}] \quad (4.12.1)$$

where we have neglected the change in V_T coming from the body effect though the $I_{DS} R_S$ addition to V_{SB} . If these resistive drops are small, so that we can keep only terms linear in the drops, and assuming V_{DS} is well below the saturation value given by (4.7.21), (4.12.1) can be rearranged to give

$$I_{DS} \approx \frac{W}{L} \frac{\mu}{1 + \beta_R (V_{GS} - V_T)} C'_{ox} \left((V_{GS} - V_T) V_{DS} - \frac{\alpha}{2} V_{DS}^2 \right) \quad (4.12.2)$$

where

$$\beta_R = \frac{W}{L} \mu C'_{ox} (R_S + R_D) \quad (4.12.3)$$

The $I_{DS} - V_{GS}$ characteristic obtained from the preceding equations is of the same form as that caused by a totally different effect: the effective mobility reduction with V_{GS} , illustrated in Fig. 4.33. If we assume that both effects are present, then we should replace μ in (4.12.2) by an effective mobility expression from Sec. 4.11. Let us use (4.11.31) with $\theta_B = 0$ for simplicity. Then it is easy to show that, if $\theta(V_{GS} - V_T)$ and $\beta_R(V_{GS} - V_T)$ are both sufficiently smaller than 1, the factor $\mu/[1 + \beta_R(V_{GS} - V_T)]$ in (4.12.2) should be replaced by $\mu_0/[1 + (\theta + \beta_R)(V_{GS} - V_T)]$.

This has led to the confusing usage of phrases such as “mobility reduction due to series resistance.” Such phrases do not properly describe what happens. As is clear from our analysis, the two effects are totally separate and different. The fact that both happen to contribute to a term proportional to $(V_{GS} - V_T)$ in the denominator of the current expression is because of a mathematical coincidence.

A more accurate analysis can be carried out if one does not make assumptions as to the relative magnitude of the various voltages involved. Even from such calculations, though, it is found that it is often sufficient to model the series resistance effect on the dc current I_{DS} as already discussed, i.e., by adding β_R to θ . As far as calculating this current is concerned, this approach can be efficient; because it eliminates the internal nodes in the model (which would have been required if the resistances were to be included as separate elements), it reduces computer simulation time, especially for very large circuits with hundreds of thousands of transistors. However, removing the explicit source and drain series resistances changes the equivalent network topology of the model, which can create complications when, at high frequencies, several capacitances must be added to the model. This can cause inaccuracies in large-signal and small-signal modeling. *Thus, in general, the source and drain resistances should be treated as elements external to the intrinsic transistor.*

The all-region models of Secs. 4.3 and 4.4 are in terms of surface potentials and not directly in terms of terminal voltages. Nevertheless, the influence of source and drain series resistance for such models is of the same form as shown in (4.12.2); this effect can be included in a similar way, by adding a term $(W/L)\mu_0(-\bar{Q}'_I)(R_S + R_D)$ to the denominator of (4.11.19), where \bar{Q}'_I is the inversion layer charge evaluated at the midpoint surface potential, given by (4.11.24).⁷³ Again, the reader is cautioned that this is just algebra and does not mean that the source and series resistances affect carrier mobility.

As was the case with the dependence of the effective mobility on transversal field, discussed in Sec. 4.11, source series resistance tends to linearize the idealized “square law” in saturation, with the effect being more pronounced for larger resistance values.

To keep the source/drain (and also the gate) resistances low, these regions are covered with metals (e.g., titanium, tantalum, or cobalt) that are made to react with silicon and form a “disilicide.”¹⁷⁰ This process is referred to as a *self-aligned silicide* (or *salicide*) process. It can reduce the sheet resistances by a factor of 5 to 10.

4.13 Temperature Effects

Transistor characteristics of MOS are strongly temperature-dependent.^{18,57,65,104,106,137,171–177} One of the main parameters responsible for this is the effective mobility, which is known to decrease with temperature. An often used approximation in strong-inversion is

$$\mu(T) = \mu(T_r) \left(\frac{T}{T_r} \right)^{-k_3}, \text{ strong inversion} \quad (4.13.1)$$

where T is absolute temperature, T_r is room absolute temperature, and k_3 is a constant, with various values used for it^{57,65,106,114,137} between 1.2 and 2.0.

Most of the temperature dependence is due to μ_0 in expressions such as (4.11.19) or (4.11.31). However, the various other parameters used in mobility expressions, such as α_θ , θ , and θ_B in Sec. 4.11, also depend on temperature. There are no theoretical formulations for these dependences, so in computer simulation models, empirical formulations are used. At very low fields, where Coulomb scattering is dominant (Sec. 4.11), mobility has been reported to increase with temperature.¹⁷⁸

Other key quantities that vary with temperature are the “thermal voltage” $\phi_t = kT/q$ and the intrinsic carrier concentration n_i , an approximate formula for which has been given in Sec. 1.2.1. Because of these, the Fermi potential, ϕ_F , is also temperature-dependent, and so is the quantity ϕ_0 found in our expressions for the threshold voltage. The quantity V_{FB} is also temperature-dependent through ϕ_{MS} in (2.2.6), assuming Q'_o is fixed. These effects are manifest in the value of V_T in (4.7.19), which is found to exhibit an almost straight-line decrease with temperature^{18, 57,137,171,172,174,175} and can be approximated by

$$V_T(T) = V_T(T_r) - k_4(T - T_r) \quad (4.13.2)$$

where k_4 is usually between 0.5 mV/K and 3 mV/K, with larger values in this range corresponding to heavier doped substrates, thicker oxides, and smaller values of V_{SB} .

As an example of the effect of temperature on transistor characteristics, consider a device operating in the saturation region. From (4.7.24) we have

$$\sqrt{I_{DS}} = \sqrt{\mu(T)} \sqrt{\frac{1}{2} \frac{W}{L} \frac{C'_{ox}}{\alpha} [V_{GS} - V_T(T)]} \quad (4.13.3)$$

Thus, a temperature increase should tend to increase the drain current through $V_{GS} - V_T(T)$ and to decrease it through $\mu(T)$. A set of $\sqrt{I_{DS}}$ vs. V_{GS} curves in saturation is shown in Fig. 4.35. At high currents, the decrease of $\mu(T)$ with temperature dominates; at low currents, the decrease of $V_T(T)$ with temperature dominates. In

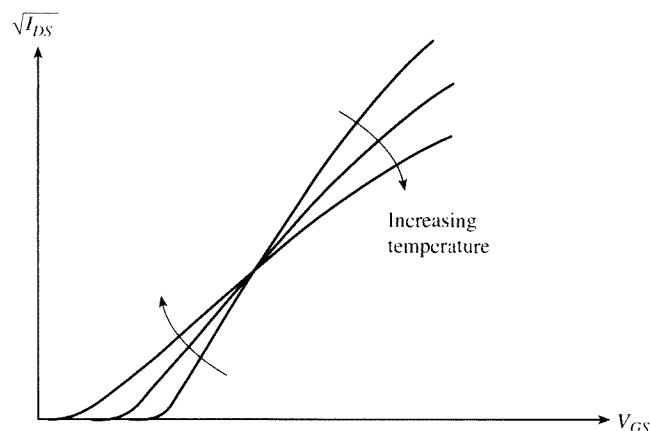


FIGURE 4.35
 $\sqrt{I_{DS}}$ in saturation vs. V_{GS} , for three different temperatures.

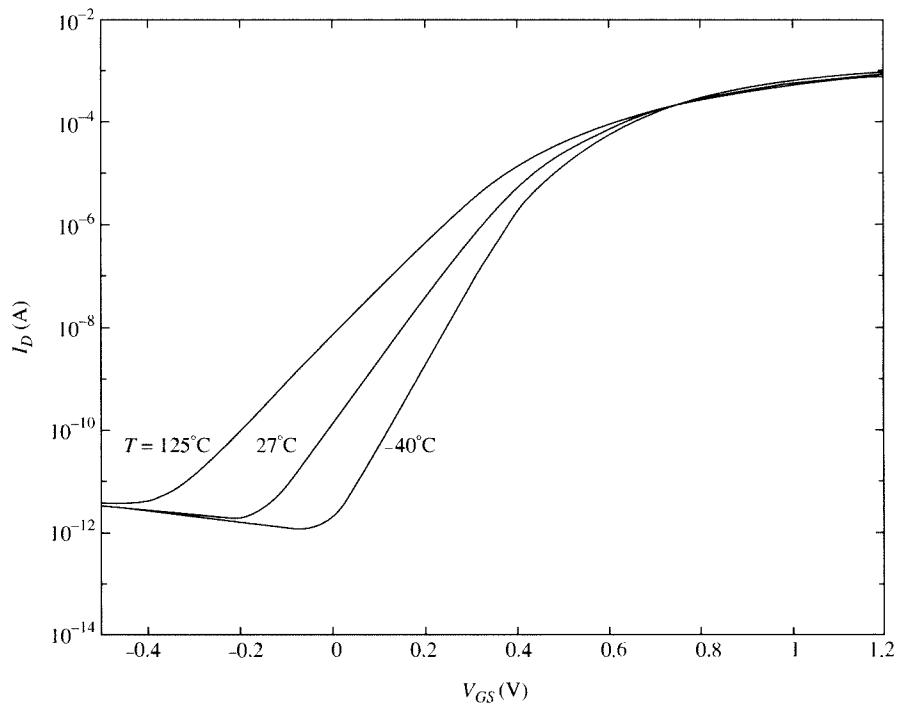


FIGURE 4.36
 I_D vs. V_{GS} for various temperatures.

certain cases, a value of V_{GS} can be found at which the current becomes practically temperature-independent over a large temperature range.^{18,171,174} This effect is evident in the figure. The bottom, curved part of the curves is due to moderate and weak inversion; the curving of the upper part is due to the dependence of effective mobility on V_{GS} (Sec. 4.11).

As can be deduced from the figure, in weak inversion and for given V_{GS} , the drain current increases with temperature. Typical plots of $\log I_D$ vs. V_{GS} are shown for various temperatures in Fig. 4.36. As seen, increasing temperature decreases the slope of the curves. At the bottom, the effect of leakage currents is evident.

The reader is referred to the literature for considerations applying to operation at very low^{176,177} and very high¹⁷⁴ temperatures.

4.14 Breakdown

The value of the various voltages that can be applied to a MOS transistor should be limited to avoid several forms of breakdown.^{18,104} One such form is *junction breakdown*. The junctions formed by the substrate and the drain or source regions will conduct a large current if the reverse bias applied to them exceeds a certain value

(because the field in the junctions near the surface is influenced by the presence of the gate, this value depends on the gate potential and can be different from that predicted by common *pn* junction theory). Junction breakdown will occur even with the device off.

When the device is on, carriers moving fast in the channel can impact on silicon atoms and ionize them, producing electron-hole pairs; this is referred to as *impact ionization*. The newly generated pairs can gain enough energy to impact on silicon atoms and produce more electron-hole pairs, etc. This is called the *avalanche effect* and is more pronounced in the pinchoff region near the drain where fields can be high. Currents larger than those predicted by common device models will then flow, and the phenomenon is referred to as *channel breakdown*. The effect of this on device characteristics is shown in Fig. 4.37. A discussion of this and other related effects can be found in Chap. 5. These forms of breakdown are nondestructive.

A *destructive breakdown* mechanism is *oxide breakdown*. It occurs when the electric field in the gate insulator exceeds a certain value, which is about 10^7 V/cm (1 V/nm) in silicon dioxide. The result is a permanent short circuit through the insulator. Static charge, such as that transferred to gates by handling devices with bare hands, is known to cause oxide breakdown. For this reason, protective devices

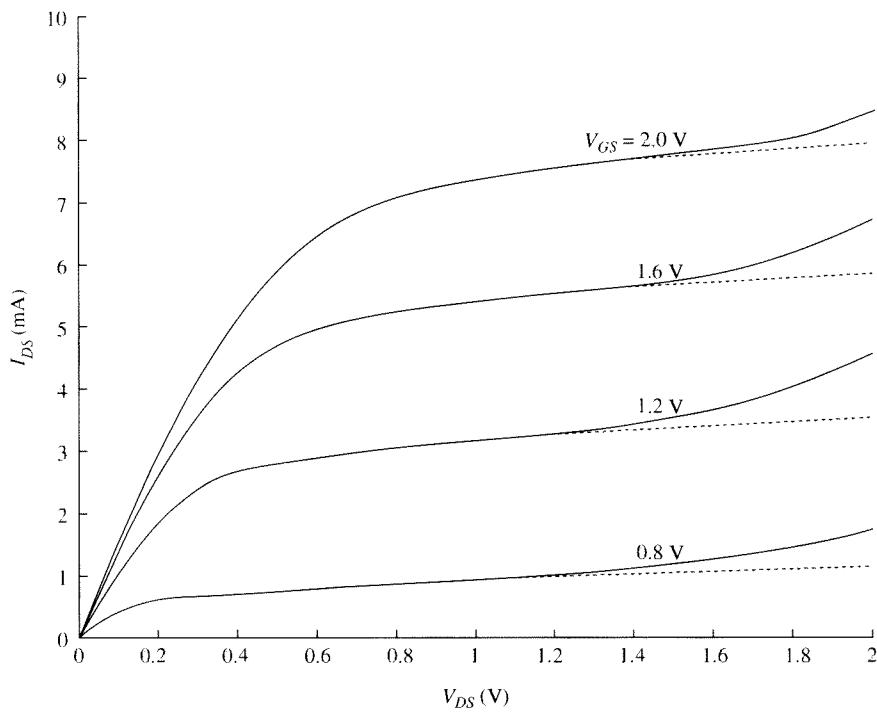


FIGURE 4.37
Effect of breakdown on I_D - V_{DS} characteristics.

are used at those input terminals of a MOS integrated circuit that are connected to transistor gates.

4.15 The *p*-Channel MOS Transistor

If the body is made of *n*-type material, and the source/drain regions of *p*⁺-type material, we have what is known as the *p*-channel MOS transistor, or *p*MOS transistor. Figure 4.38 shows such a device. An example of *p*-channel transistor characteristics is shown in Fig. 4.39.

The operation of the *p*-channel transistor is the “complement” of *n*-channel operation. The role of electrons is played by holes, and the role of ionized acceptor atoms is played by ionized donor atoms. Statements made about *n*-channel devices can be adapted to the case of *p*-channel devices with simple modifications. For example, in Fig. 4.38, the more *negative* the gate-source voltage the heavier the concentration of *holes* near the surface. The more *negative* the drain-source voltage the heavier the flow of *holes* from source to drain; hence, the more *negative* the drain current (assuming the reference direction is chosen as before, i.e., from the drain through the channel to the source). The more *negative* the source-substrate bias the larger the number of *donor* atoms that are depleted. The body effect coefficient γ is given by

$$\gamma = \frac{\sqrt{2q\epsilon_s N_D}}{C'_{ox}} \quad (4.15.1)$$

where N_D is the concentration of donor atoms in the substrate.

In describing *p*-channel devices, there are some rather obvious sign changes in the model equations; for example, instead of (4.7.18) we will have

$$I_{DSN} = -\frac{W}{L} \mu C'_{ox} \left[(V_{GS} - V_T)V_{DS} - \frac{\alpha}{2} V_{DS}^2 \right] \quad (4.15.2)$$

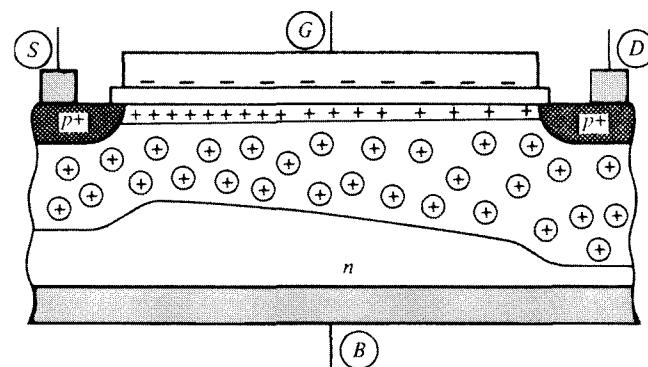


FIGURE 4.38
A *p*-channel MOS transistor.

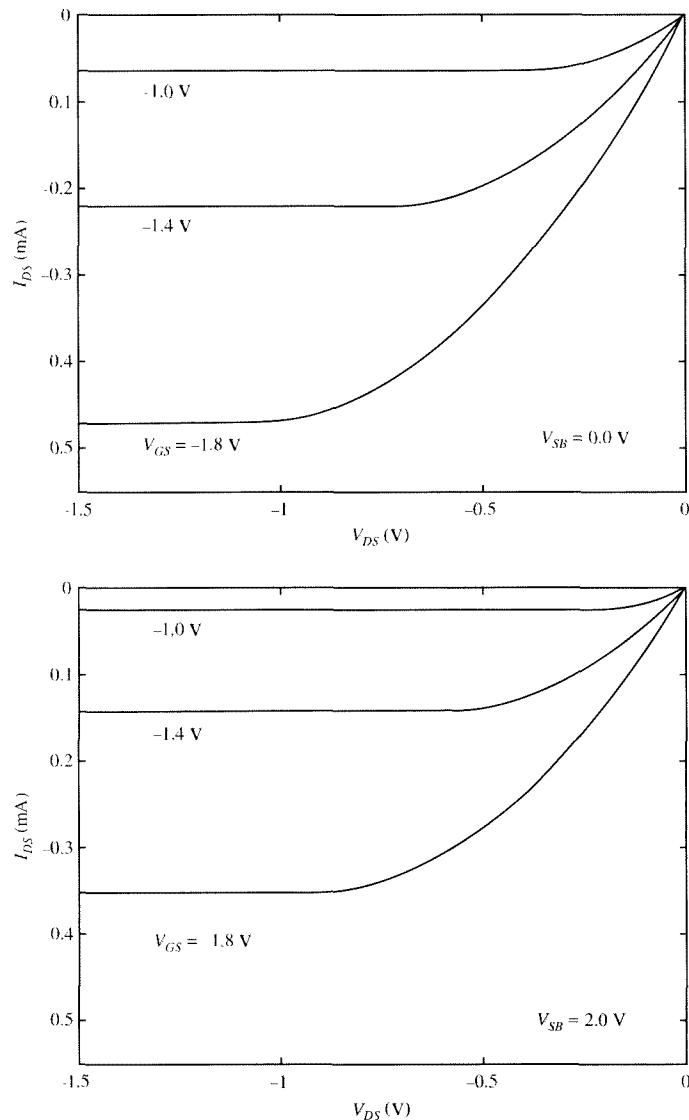


FIGURE 4.39
 I_{DS} - V_{DS} characteristics for a *p*-channel MOS transistor for two different V_{SB} values.

where instead of (4.7.20a) we will have

$$V_T(V_{SB}) = V_{T0} - \gamma \left(\sqrt{-V_{SB} - \phi_0} - \sqrt{-\phi_0} \right) \quad (4.15.3)$$

where V_{SB} and ϕ_0 are negative and, instead of (4.7.20b), we will have

$$V_{T0} = V_{FB} + \phi_0 - \gamma \sqrt{-\phi_0} \quad (4.15.4)$$

In describing *p*-channel devices, it is common to simply “flip” the polarity of the biases and then treat the analysis of the transistor as if it were an *n*-channel device. This is, in fact, how *p*MOS transistors are generally handled in circuit simulation programs. To balance this, the drain current is then also “flipped” in sign after model evaluation. For this reason, transistor characteristics for *p*MOS devices are often (but not always) presented in terms of *n*MOS equivalent bias polarities and are displayed in a form such as that in Fig. 4.24, rather than that in Fig. 4.39.

The effective mobility of *p*-channel devices behaves differently from that of *n*-channel devices. The parameter μ_0 (Sec. 4.11) is lower for *p*MOS transistors, typically by a factor of from 2 to 4. Also, values of the parameters determining the dependence of mobility on electric field are different; for example, the parameter η_E in Sec. 4.11 is lower for *p*MOS devices (typically 0.3 vs. 0.5 for *n*MOS).

4.16 Enhancement-Mode and Depletion-Mode Transistors

Consider the simple model of (4.7.24). Depending on the sign of V_{T0} , MOS transistors are separated into two categories. The *n*-channel transistors with positive V_{T0} are called *enhancement-mode* (or “normally off”) devices, whereas *n*-channel transistors with negative V_{T0} are said to be *depletion-mode* (or “normally on”) devices. These names originated in the days when weak inversion was neglected, and the *n*-channel transistor

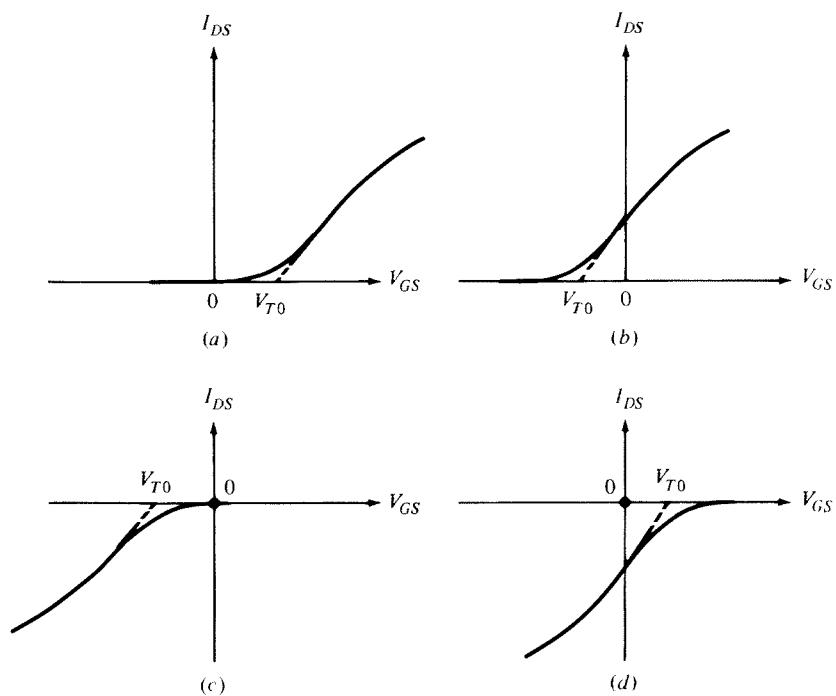


FIGURE 4.40
 I_{DS} vs. V_{GS} for $V_{SS} = 0$ and very small $|V_{DS}|$. (a) *n*-channel enhancement device; (b) *n*-channel depletion device; (c) *p*-channel enhancement device; (d) *p*-channel depletion device.

was viewed as being on for $V_{GS} > V_{T0}$ and off for $V_{GS} < V_{T0}$ (with $V_{SB} = 0$ assumed). Thus, if V_{T0} is positive, an *n*-channel device is assumed off at $V_{GS} = 0$, and it takes a positive V_{GS} to “enhance” the channel and turn the device on. If V_{T0} is negative, the *n*-channel device is already on with $V_{GS} = 0$, and it takes a negative V_{GS} to “deplete” the channel and turn the device off. For *p*-channel devices, a negative V_{T0} corresponds to the enhancement mode and a positive V_{T0} to the depletion mode. The I_D - V_{GS} characteristics for the four types of devices for a very small $|V_{DS}|$ are shown in Fig. 4.40.

The value of V_{T0} can, in principle, be set by a very shallow ion implantation during fabrication. In such a process, the device is bombarded with high-energy ions. If all such ions could end up at the oxide-semiconductor interface, their effect would be the same as that of Q'_o in (2.2.6). Hence, the value of V_{FB} could be adjusted, which, in turn, would adjust the value of V_{T0} in (4.7.20b). In practice, however, ion implantation cannot be infinitesimally shallow. The spreading of ions into the substrate can create second-order effects, and then the resulting device cannot be modeled accurately as in this chapter. Ion-implanted devices are considered in Chap. 9.

4.17 Model Parameter Values, Model Accuracy, and Model Comparison

In the models we have developed, some parameters do not have an exact theoretical value. For example, such is the case with ϕ_0 , which represents the supposedly “pinned” value of the surface potential in strong inversion (with $V_{SB} = 0$). Since the surface potential is never exactly pinned but instead varies with bias, the best single value to be used for ϕ_0 will depend on the bias range we are interested in. The same is true for V_{T0} , which contains ϕ_0 in its definition. Even parameter values that have been specified “exactly” (e.g., the body effect coefficient γ) are actually exact only for the fictitious, idealized device corresponding to the simplifying assumptions we have made (e.g., that the substrate doping is “exactly” uniform). Thus, when a model is used to represent a real device, the parameter values that will result in minimum error between model predictions and experiment are not necessarily dictated by a simple theory. For this reason, what is usually done is to assign values to these parameters in such a way as to provide “best matching” between model predictions and measurements. In this process, the “theoretical” values for the parameters might simply play the role of an initial guess. Of course, what constitutes best matching is subject to interpretation. One might desire good matching for devices with $L = 100$ nm and V_{GS} from 0.4 to 0.6 V, or, most likely, for devices of various lengths in a wide range of bias voltages. The sets of parameter values to be used in these two cases will not necessarily be the same, and the matching is likely to be better in the first case, assuming the same model is used. Also, if the same parameter appears in two different models (e.g., ϕ_0 appears in both the complete model and the simplified strong-inversion models of Sec. 4.7), different values for it might have to be used in each model for best results. For this reason, direct comparison of model equations is difficult and should be taken with a grain of salt.

Let us offer an example of problems that can arise when model predictions are compared to experimental results. Let us say that the weak-inversion model of (4.8.11) for given terminal voltages and process parameters predicts a current 4 times

larger than the measured value. Is the model bad? Not necessarily. The problem may just be that the process parameters are not accurately known. This is notably true of V_{FB} , for example, which appears in the exponential term in (4.8.11) through V_M . When n is 1.3, a 50-mV error in the value of V_{FB} is enough to give 4 times the actual current! This, of course, happens because in weak inversion the $\log I_{DS}$ vs. V_{GS} curves (e.g., Fig. 4.31) are so steep that a slight shift horizontally corresponds to a large current change. Note that, if *I-V plots* were compared rather than numbers, the above problem with V_{FB} could have been spotted immediately. One would have seen that the curves produced by the model had the same shape as the measured ones, only they were shifted somewhat horizontally. Thus, V_{FB} would be suspected and a new value would be tried for that parameter. In general, blind number comparison can be very misleading.

In the preceding example, the error in the value of V_{FB} could have been the result of inadequate knowledge of process parameters, for example, oxide thickness or substrate doping. In other cases, though, V_{FB} or some other parameters may have to be given a wrong value on purpose in order to make up for model deficiencies. This, however, can be dangerous. “Fixing” a modeling problem in one aspect in this way can cause problems to appear in other aspects (e.g., fixing the accuracy in I_{DS} can worsen the accuracy in the prediction of small-signal capacitances). What is worse, one may not be aware that such problems might be caused, because one may not be looking for them. Thus, a model based on physical considerations, in which physical values can be used for all its parameters, is the ideal situation. *If a model is good, the parameter values that provide the best fit to measurements should be close to the values expected from theory.* This point is discussed further in Chap. 10.

The determination of parameter values is done by using sophisticated and mostly automated systems devoted to data acquisition and parameter extraction. *I-V* or *C-V* characteristics are measured for many devices of a certain type. The data obtained are then processed by software that implements the algorithms for the determination of parameter values. As part of the parameter determination process, numerical optimization is performed to minimize the error between modeled and measured values. The selection of an appropriate error criterion depends on the application for which the model is intended. For example, for analog applications, the error should involve small-signal parameters. It is better to extract values for model parameters from *I-V* characteristic regions where such parameters have a strong effect. For example, if nonsaturation drain current values are plotted vs. V_{GS} for a very small V_{DS} , we obtain a curve of the form of Fig. 4.33; if, in the region of maximum slope of that curve, we draw a tangent and extrapolate to the V_{GS} axis, we obtain the value of V_T .[†] Parameter extraction is discussed in Chap. 10.

Inconsistent Usage of “Threshold Voltage” We shall take the opportunity here to emphasize that in this book V_T is the *extrapolated* threshold voltage, a term originating in Sec. 3.4.2 and also justified by the construction we have just described. Unfortunately, the term *threshold voltage* is used in the literature with at least three

[†] More accurately, as can be verified from (4.7.24a), the V_{GS} axis intercept of that tangent gives $V_T + (\alpha/2)V_{DS}$.

other meanings. Sometimes it is taken to denote the quantity $V_{FB} + 2\phi_F + \sqrt{2\phi_F + V_{SB}}$. This quantity is actually V_M , the value of V_{GS} at the lower limit of moderate inversion, as has been seen in Sec. 4.6. It is *different* from V_T in (4.7.19), simply because ϕ_0 is different from $2\phi_F$ (Sec. 2.6.2). Elsewhere, “threshold voltage” is used to imply what is sometimes called *constant current threshold voltage*. This is the value of V_{GS} needed to reach a specified value of $I_{DS}/(W/L)$, which often happens to fall somewhere in the moderate-inversion region or even in the weak-inversion region. Finally, sometimes threshold voltage is taken to mean vaguely “the value of V_{GS} at which strong inversion begins.” The name “threshold voltage” is often used indiscriminately for all these quantities, and sometimes it is attempted to match calculated values of one of these “thresholds” (e.g., $V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{SB}}$) to measured values of another threshold (e.g., the extrapolated threshold). All this originated in the early days, when V_{GS} values of, say, 15 V were not uncommon; compared to these, an ambiguity of, say, 0.2 V in the value of threshold was not a problem. Today, however, when low-voltage circuit operation is the norm, a distinction of the above quantities is important. In this book “threshold voltage” will always imply *extrapolated* threshold voltage unless stated otherwise. The extrapolation can be in terms of current, as in this section, or in terms of inversion layer charge, as in Fig. 3.6d.

References

1. J. E. Lilienfeld, U.S. Patents 1,745,175 (1930) (filed Oct. 8, 1926), 1,877,140 (1932), 1,900,018 (1933).
2. O. Heil, British Patent 439,457 (filed and granted 1935).
3. W. Shockley and G. L. Pearson, “Modulation of conductance of thin films of semiconductors by surface charges,” *Physical Review*, vol. 74, pp. 232–233, 1948.
4. D. Kahng and M. M. Atalla, “Silicon-silicon dioxide field induced devices,” *Solid-State Device Research Conference*, Pittsburgh, June 1960.
5. C.-T. Sah, “Evolution of the MOS transistor—From conception to VLSI,” *Proceedings of the IEEE*, vol. 76, no. 10, pp. 1280–1326, 1988.
6. H. K. J. Ihantola, “Design theory of surface field-effect transistor,” *Technical Report No. 1161-I*, Stanford Electronics Laboratories, Stanford University, Stanford, Calif., September 1961.
7. S. R. Hofstein and F. P. Heinman, “The silicon insulated-gate field effect transistor,” *Proceedings of the IEEE*, vol. 51, pp. 1190–1202, 1963.
8. H. K. J. Ihantola and J. L. Moll, “Design theory of a surface field-effect transistor,” *Solid-State Electronics*, vol. 7, pp. 423–430, 1964.
9. C.-T. Sah, “Characteristics of the metal-oxide-semiconductor transistors,” *IEEE Transactions on Electron Devices*, vol. ED-11, pp. 324–345, 1964.
10. H. C. Pao and C.-T. Sah, “Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors,” *Solid-State Electronics*, vol. 10, pp. 927–937, 1966.
11. J. T. Wallmark and H. Johnson, *Field Effect Transistors: Physics, Technology and Applications*, Prentice Hall, Englewood Cliffs, N.J., 1966.
12. C.-T. Sah and H. C. Pao, “The effects of fixed bulk charge on the characteristics of metal-oxide semiconductor transistors,” *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 393–409, 1966.
13. R. S. Cobbold, “MOS transistor as a four terminal device,” *Electronics Letters*, vol. 2, pp. 189–190, 1966.
14. J. A. Van Nielen and O. W. Memelink, “The influence of the substrate upon the D.C. characteristics of silicon MOS transistors,” *Philips Research Reports*, vol. 22, pp. 55–71, 1967.
15. M. B. Das, “Dependence of the characteristics of MOS transistors on the substrate resistivity,” *Solid-State Electronics*, vol. 11, pp. 305–322, 1968.
16. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.

17. M. B. Das, "Physical limitation of MOS structures," *Solid-State Electronics*, vol. 12, pp. 305–336, 1969.
18. R. S. C. Cobbold, *Theory and Applications of Field-Effect Transistors*, Wiley-Interscience, New York, 1970.
19. P. Richman, *Characteristics and Operation of MOS Field Effect Devices*, McGraw-Hill, New York, 1967.
20. R. H. Crawford, *MOSFET in Circuit Design*, McGraw-Hill, New York, 1967.
21. H. Shichman and D. A. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-3, pp. 285–289, 1968.
22. D. Frohman-Bentchkowsky and L. Vadasz, "Computer-aided design and characterization of digital MOS integrated circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-4, pp. 57–64, 1969.
23. J. E. Meyer, "MOS models and circuit simulations," *RCA Review*, vol. 32, pp. 42–63, 1971.
24. G. Merckel, J. Borel, and N. Z. Cupcea, "An accurate large-signal MOS transistor model for use in computer-aided design," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 681–690, 1972.
25. F. M. Klaassen, "A MOS model for computer-aided design," *Philips Research Reports*, vol. 31, pp. 71–83, 1976.
26. Y. Hayashi and Y. Tarui, "Exponential current in MOST-type devices and deterioration of reverse current in p-n junctions," *Technical Report SSD 67-6*, Technical Group on Semiconductors and Semiconductor Devices, Institute of Electronics and Communications Engineers of Japan, 1967.
27. A. A. Guzev, G. L. Kuryshev, and S. P. Sinitsa, "Investigation of carrier capture on the surface of silicon in a metal-insulator-semiconductor transistor," *Soviet Physics—Semiconductors*, vol. 4, pp. 1245–1249, 1971.
28. R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-7, pp. 146–153, 1972.
29. M. B. Barron, "Low-level currents in insulated-gate field-effect transistors," *Solid-State Electronics*, vol. 15, pp. 293–302, 1972.
30. R. A. Stuart and W. Eccleston, "Leakage currents of MOS devices under depletion conditions," *Electronics Letters*, vol. 8, pp. 225–227, 1972.
31. R. R. Troutman and S. N. Chakravarti, "Subthreshold characteristics of insulated-gate field-effect transistors," *IEEE Transactions on Circuit Theory*, vol. CT-20, pp. 659–665, 1973.
32. R. J. Van Overstraeten, G. Declerck, and G. L. Broux, "Inadequacy of the classical theory of the MOS transistor operation in weak inversion," *IEEE Transactions on Electron Devices*, vol. ED-20, pp. 1150–1153, 1973.
33. R. R. Troutman, "Subthreshold design consideration for insulated gate field-effect transistors," *IEEE Journal of Solid-State Circuits*, vol. SC-9, pp. 55–60, 1974.
34. T. Masuhara, J. Etoh, and M. Nagata, "A precise MOSFET model for low-voltage circuits," *IEEE Transactions on Electron Devices*, vol. ED-21, pp. 363–371, 1974.
35. R. J. Van Overstraeten, G. J. Declerck, and P. A. Muls, "Theory of MOS transistor in weak inversion—New method to determine the number of surface states," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 282–288, 1975.
36. R. R. Troutman, "Subthreshold slope for insulated gate field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 1049–1051, 1975.
37. E. Vittoz and J. Fellrath, "CMOS analog circuits based on weak inversion operation," *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 224–231, 1977.
38. G. W. Taylor, "Subthreshold conduction in MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 337–350, 1978.
39. W. Fichtner and H. W. Potzl, "MOS modelling by analytical approximations. I. Subthreshold current and threshold voltage," *International Journal of Electronics*, vol. 46, pp. 33–55, 1979.
40. I. R. M. Mansour, "On the modeling of MOS devices," *Proceedings of the Third International Symposium on Network Theory*, Yugoslavia, 1975, pp. 705–713.
41. I. R. M. Mansour, "Improved modeling of MOS devices," *Proceedings of the European Conference on Circuit Theory and Design*, Italy, 1976.
42. Y. A. El-Mansy and A. R. Boothroyd, "A new approach to the theory and modeling of insulated-gate field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-24, pp. 241–253, 1977. See also related comments by J. R. Brews, *IEEE Transactions on Electron Devices*,

- vol. ED-24, pp. 1369–1370, 1977, and reply by Y. A. El-Mansy and A. R. Boothroyd, *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 393–394, 1978.
43. G. Baccarani, M. Rudan, and G. Spadini, “Analytical i.g.f.e.t. model including drift and diffusion currents,” *IEEE Journal on Solid-State and Electron Devices*, vol. 2, pp. 62–68, 1978.
 44. J. R. Brews, “A charge sheet model for the MOSFET,” *Solid-State Electronics*, vol. 21, pp. 345–355, 1978.
 45. F. Van de Wiele, “A long-channel MOSFET model,” *Solid-State Electronics*, vol. 22, pp. 991–997, 1979.
 46. J. R. Brews, “Physics of the MOS transistor,” chap. 1, *Silicon Integrated Circuits, Part A*, D. Kahng (editor), Applied Solid-State Science Series, Academic Press, New York, 1981.
 47. R. F. Pierret and J. A. Shields, “Simplified long-channel MOSFET theory,” *Solid-State Electronics*, vol. 26, pp. 143–147, 1983.
 48. H. K. Lim and J. G. Fossum, “An analytic characterization of weak-inversion drift current in a long-channel MOSFET,” *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 713–715, 1983.
 49. C. Turchetti, “Relationships for the drift and diffusion components of the drain current in an MOS transistor,” *Electronics Letters*, vol. 19, pp. 960–962, 1983.
 50. M. Bagheri and C. Turchetti, “The need for an explicit model describing MOS transistors in moderate inversion,” *Electronics Letters*, vol. 21, pp. 873–874, 1985.
 51. P. P. Guebels and F. Van de Wiele, “A charge sheet model for small geometry MOSFETs,” *Proceedings of the IEEE International Electron Devices Meeting*, Washington, 1981, pp. 211–214.
 52. C. Turchetti and G. Mascetti, “A CAD-oriented analytical MOSFET model for high-accuracy applications,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-3, pp. 117–122, 1984.
 53. S. Yu, A. F. Franz, T. G. Mihran, “A physical parametric transistor model for CMOS circuit simulation,” *IEEE Transactions on Computer-Aided Design*, vol. 7, pp. 1038–1052, 1988.
 54. H.-J. Park, P. K. Ko, and C. Hu, “A charge sheet capacitance model of short channel MOSFET’s for SPICE,” *IEEE Transactions on Computer-Aided Design*, vol. 10, pp. 376–389, 1991.
 55. A. R. Boothroyd, S. W. Tarasewicz, and C. Slaby, “MISNAN—A physically based continuous MOSFET model for CAD applications,” *IEEE Transactions on Computer-Aided Design*, vol. 10, pp. 1512–1529, 1991.
 56. M. Miura-Mattausch, U. Feldmann, A. Rahm, M. Bollu, and D. Savignac, “Unified complete MOSFET model for analysis of digital and analog circuits,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, pp. 1–7, 1996.
 57. N. Arora, *MOSFET Models for VLSI Circuit Simulation—Theory and Practice*, Springer-Verlag, Vienna, 1993.
 58. R. Rios, S. Mudania, W.-K. Shih, and P. Packan, “An efficient surface potential solution algorithm for compact MOSFET models,” *Proceedings of the IEEE International Electron Devices Meeting*, pp. 755–758, December 2004.
 59. R. van Langevelde and F. M. Klaasssen, “Explicit surface-potential-based MOSFET model for circuit simulation,” *Solid-State Electronics*, vol. 44, pp. 409–418, 2000.
 60. T.-L. Chen and G. Gildenblat, “Analytical approximation for the MOSFET surface potential,” *Solid-State Electronics*, vol. 45, pp. 335–339, 2001.
 61. G. Gildenblat, H. Wang, T.-L. Chen, X. Gu, and X. Cai, “SP: An advanced surface-potential-based compact MOSFET model,” *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1394–1406, 2004.
 62. H. K. Gummel, private communication.
 63. G. Gildenblat, T. L. Chen, and P. Bendix, “Computationally efficient implementation of charge sheet model,” *Electronics Letters*, vol. 35, no. 10, pp. 843–844, 1999.
 64. O. Wing, private communication.
 65. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
 66. P. G. A. Jespers, C. Jusseret, and Y. Leduc, “A fast sample and hold charge-sensing circuit for photodiode arrays,” *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 232–237, 1977.
 67. T.-L. Chen and G. Gildenblat, “Symmetric bulk charge linearisation in charge-sheet MOSFET model,” *Electronics Letters*, vol. 37, no. 12, pp. 791–793, 2001.
 68. E. O. Johnson, “The insulated-gate field-effect transistor—A bipolar transistor in disguise,” *RCA Review*, vol. 34, pp. 80–94, 1973.

69. M. S. Lundstrom and J. Guo, *Nanoscale Transistors: Device Physics, Modeling and Simulation*, Springer, New York, 2006.
70. R. S. Müller and J. R. Schroeder, "Properties of a field-aided IGFET," IEEE German Physics Society—Society of Electrical Engineers European Meeting, *Semiconductor Device Research*, Munich, Germany, March 24–27, 1969, pp. S1–S7.
71. K. Vavelidis, Y. Tsividis, F. Op't Eynde, and Y. Papananos, "Six-terminal MOSFETs: Modeling and applications in highly linear, electronically tunable resistors," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 1, pp. 4–12, 1997.
72. H. Wang, T.-L. Chen, and G. Gildenblat, "Quasi-static and nonquasi-static compact MOSFET models based on symmetric linearization of the bulk and inversion charges," *IEEE Transactions on Electron Devices*, vol. 50, no. 11, pp. 2262–2272, 2003.
73. G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, 2006.
74. R. van Langevelde and G. Gildenblat, "PSP: An advanced surface-potential-based MOSFET model," W. Grabinski, B. Nauwelaerts, and D. Schreurs (editors), *Transistor Level Modeling for Analog/RF Design*, pp. 29–66, Springer-Verlag, Weinheim, 2006.
75. M. Bagheri and Y. Tsividis, "A small-signal dc-to-high-frequency nonquasistatic model for the four-terminal MOSFET valid in all regions of operation," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2383–2391, 1985.
76. B. Jñiguez and E. G. Moreno, "A physically based C_∞ -continuous model for small-geometry MOSFETs," *IEEE Transactions on Electron Devices*, vol. 42, pp. 283–287, 1995.
77. M. Bagheri, "Improving the non-quasistatic weak-to-strong-inversion four-terminal MOSFET model," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 12, pp. 2558–2560, 1987.
78. M. A. Maher and C. A. Mead, "A physical charge-controlled model for MOS transistors," pp. 211–229, *Advanced Research VLSI*, P. Lesleben (editor), The MIT Press, Cambridge, Mass., 1987.
79. C.-K. Park, C.-Y. Lee, K. Lee, B.-J. Moon, Y. H. Byun, and M. Shur, "A unified current-voltage model for long-channel nMOSFETs," *IEEE Transactions on Electron Devices*, vol. 38, pp. 399–406, 1991.
80. K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI*, Prentice-Hall, Englewood Cliffs, N.J., 1993.
81. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An explicit physical model for the long-channel MOS transistor including small-signal parameters," *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
82. A. I. A. Cunha, *Um Modelo do Transistor MOS para Projeto de Circuitos Integrados*, Ph.D. thesis, Universidade Federal de Santa Catarina, 1996 (in Portuguese).
83. C. C. Enz, *High Precision CMOS Micropower Amplifiers*, Ph.D. thesis no. 802, École Polytechnique Fédéral de Lausanne, Switzerland, 1989.
84. C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, 1995.
85. C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling*, Wiley, Hoboken, N.J., 2006.
86. C. Galup-Montoro and M. C. Schneider, *MOSFET Modeling for Circuit Analysis and Design*, World Scientific, River Edge, N.J., 2007.
87. J. Watts, C. C. McAndrew, C. Enz, C. Galup-Montoro, G. Gildenblat, C. Hu, R. van Langevelde, M. Miura-Mattausch, R. Rios, and C.-T. Sah, "Advanced compact models for MOSFETs," *Proceedings of the Workshop on Compact Modeling*, pp. 3–12, May 2005.
88. A. I. Cunha, M. C. Schneider, and C. Galup-Montoro, "Derivation of the unified charge control model and parameter extraction procedure," *Solid-State Electronics*, vol. 43, no. 3, pp. 481–485, 1999.
89. H. Oguey and S. Cserenyi, "Modèle du transistor MOS valable dans un grand domaine de courants," *Sonderdruck aus dem Bulletin des SEV/VSE*, vol. 73, pp. 113–116, 1982.
90. E. A. Vittoz, "Micropower Techniques," *Design of Analog-Digital VLSI Circuits for Telecommunications and Signal Processing*, J. E. Franca and Y. Tsividis (editors), Prentice-Hall, Englewood Cliffs, N.J., 1994.
91. A. I. A. Cunha, O. C. Gouveia-Filho, M. C. Schneider, and C. Galup-Montoro, "A current-based model for the MOS transistor," *Proceedings of the 1997 International Symposium on Circuits and Systems*, Hong Kong, June 1997, pp. 1608–1611.

92. D. M. Binkley, *Tradeoffs and Optimization in Analog CMOS Design*, Wiley, Hoboken, N.J., 2008.
93. J.-M. Sallese, M. Bucher, F. Krummenacher, and P. Fazan, "Inversion charge linearization in MOSFET modeling and rigorous derivation of the EKV compact model," *Solid-State Electronics*, vol. 47, pp. 677-683, 2003.
94. A. Vladimirescu and S. Liu, "The simulation of ICs using SPICE-2," University of California-Berkeley, Electronics Research Laboratory, Memorandum M80/7, 1980.
95. H. Wallinga and K. Bult, "Design and analysis of CMOS analog signal processing circuits by means of a graphical MOST model," *IEEE Journal of Solid-State Circuits*, vol. 24, p. 672, 1989.
96. Y. Tsividis and G. Masetti, "Problems in precision modeling of the MOS transistor for analog applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-3, pp. 72-79, 1984.
97. Y. Tsividis, M. Banu, and J. F. Khouri, "Continuous-time MOSFET-C filters in VLSI," *IEEE Journal of Solid-State Circuits*, vol. SC-21, no. 1, pp. 15-30, 1986.
98. K. v. Arnim, E. Borinski, P. Segebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha, "Efficiency of body biasing in 90-nm CMOS for low-power digital circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 7, pp. 1549-1556, 2005.
99. J. A. Geurst, "Calculation of high frequency characteristics of thin film transistors," *Solid-State Electronics*, vol. 8, pp. 88-90, 1965.
100. J. J. Paulos and D. A. Antoniadis, "Limitations of quasi-static capacitance models for the MOS transistor," *IEEE Electron Device Letters*, vol. EDL-4, pp. 221-224, 1983.
101. C. C. McAndrew, B. K. Bhattacharyya, and O. Wing, "A single-piece C_∞ -continuous MOSFET model including subthreshold conductance," *IEEE Electron Device Letters*, vol. 12, pp. 565-567, 1991.
102. J. A. Power and W. A. Lane, "An enhanced SPICE MOSFET model suitable for analog applications," *IEEE Transactions on Computer-Aided Design*, vol. 11, pp. 1418-1425, 1992.
103. N. D. Arora, R. Rios, C.-L. Huang, and K. Raol, "PCIM: A physically based continuous short-channel IGFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 988-997, 1994.
104. F. M. Klaassen, "Review of physical models for MOS transistors," *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
105. H. C. Poon, private communication, 1977.
106. F. M. Klaassen, "MOS device modeling," *Design of VLSI Circuits for Telecommunications*, Y. Tsividis and P. Antognetti (editors), Prentice Hall, Englewood Cliffs, N.J., pp. 1-25, 1985.
107. Y. Tsividis, K. Suyama, and K. Vavelidis, "A simple 'Reconciliation' MOSFET model valid in all regions," *Electronics Letters*, vol. 31, pp. 506-508, 1995.
108. P. Antognetti, D. D. Caviglia, and E. Profumo, "CAD model for threshold and subthreshold conduction in MOSFETs," *IEEE Journal of Solid-State Circuits*, vol. SC-17, pp. 454-458, 1982.
109. G. T. Wright, "Simple and continuous MOSFET models for the computer-aided design of VLSI," *IEE Proceedings*, vol. 132, part I, pp. 187-194, 1985.
110. S. Liu and L. W. Nagel, "Small-signal MOSFET models for analog circuit design," *IEEE Journal of Solid-State Circuits*, vol. SC-17, pp. 983-998, 1982.
111. A. Hiroki, S. Odanaka, K. Ohe, and H. Esaki, "A mobility model for submicrometer MOSFET device simulations," *IEEE Electron Device Letters*, vol. EDL-8, pp. 231-233, 1987.
112. P. Rossel, "Influence de la réduction de mobilité due au champ transversal sur les caractéristiques des m.o.s.t.," *Electronics Letters*, vol. 5, pp. 604-605, 1969.
113. S. A. Schwarz and S. E. Russek, "Semi-empirical equations for electron velocity in silicon, Part II: MOS inversion layer," *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 1634-1639, 1983.
114. A. G. Sabinis and J. T. Clemens, "Characterization of the electron mobility in the inverted <100> Si surface," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1979, pp. 18-21.
115. J. T. C. Chen and R. S. Muller, "Carrier mobilities at weakly inverted silicon surfaces," *Journal of Applied Physics*, vol. 45, p. 828, 1974.
116. J. R. Brews, "Carrier-density fluctuations and the IGFET mobility near threshold," *Journal of Applied Physics*, vol. 46, p. 2193, 1975.
117. C. G. Sodini, T. Ekstedt, and J. L. Moll, "Charge accumulation and mobility in thin dielectric MOS transistors," *Solid-State Electronics*, vol. 25, pp. 833-841, 1982.

118. S. C. Sun and J. D. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surface," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 1497–1508, 1980.
119. T. Ando, A. B. Fowler, and F. Stern, "Electronic properties of two-dimensional systems," *Reviews of Modern Physics*, vol. 54, pp. 437–672, 1982.
120. K. Y. Fu, "Mobility degradation due to the gate field in the inversion layer or MOSFETs," *IEEE Electron Device Letters*, vol. EDL-3, pp. 292–293, 1982.
121. M.-S. Liang, J. Y. Choi, P.-K. Ko, and C. Hu, "Inversion-layer capacitance and mobility of very thin gate-oxide MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-33, p. 409, 1986.
122. B. Majkusiak and A. Jakubowski, "The dependence of MOSFET surface carrier mobility on gate-oxide thickness," *IEEE Transactions on Electron Devices*, vol. ED-33, p. 1717, 1986.
123. N. D. Arora and G. S. Gildenblat, "A semi-empirical model of the MOSFET inversion layer mobility for low-temperature operation," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 89–93, 1987.
124. J. W. Watt and J. D. Plummer, "Universal mobility-field curves for electrons and holes in MOS inversion layers," *Proceedings of the Symposium on VLSI Technology*, pp. 81–82, 1987.
125. A. J. Walker and P. H. Woerlee, "Mobility model for silicon inversion layers," *Technical Digest, ESSDERC 1987*, pp. 667–670.
126. W. M. Soppa and H.-G. Wagemann, "Investigation and modeling of the surface mobility of MOSFETs from -25 to $+150^{\circ}\text{C}$," *IEEE Transactions on Electron Devices*, vol. 35, p. 970, 1988.
127. J. A. Wikstrom and C. R. Viswanathan, "Lateral nonuniformities and the MOSFET mobility step near threshold," *IEEE Transactions on Electron Devices*, vol. 35, pp. 2378–2383, 1988.
128. S.-W. Lee, "Universality of mobility-gate field characteristics of electrons in the inversion charge layer and its application in MOSFET modeling," *IEEE Transactions on Computer-Aided Design*, vol. 8, pp. 724–730, 1989.
129. D. T. Amm, H. Mingam, P. Delpech, and T. T. D'ouville, "Surface mobility in n^+ and p^+ doped polysilicon gate PMOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-36, pp. 963–967, 1989.
130. H. Shin, A. F. Tasch, Jr., C. M. Mazier, and S. K. Banerjee, "A new approach to verify and derive a transverse-field-dependent mobility model for electrons in MOS inversion layers," *IEEE Transactions on Electron Devices*, vol. 36, p. 1117, 1989.
131. D. S. Jeon and D. E. Burk, "MOSFET electron inversion layer mobilities—A physically based semi-empirical model for a wide temperature range," *IEEE Transactions on Electron Devices*, vol. 36, p. 1456, 1989.
132. J. S. Kang, D. K. Schroder, and A. R. Alvarez, "Effective and field-effect mobilities in Si MOSFETs," *Solid-State Electronics*, vol. 32, pp. 679–681, 1989.
133. G. M. Yeric, A. F. Tasch, and S. K. Banerjee, "A universal MOSFET mobility degradation model for circuit simulation," *IEEE Transactions on Computer-Aided Design*, vol. 9, pp. 1123–1126, 1990.
134. C.-L. Huang and G. S. Gildenblat, "Measurements and modeling of the n-channel MOSFET inversion layer mobility and device characteristics in the temperature range 60–300 K," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1289–1300, 1990.
135. D. B. M. Klaassen, "A unified mobility model for device simulation," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 357–360, 1990.
136. P. Habas, "A physics based analytical MOSFET model with accurate field dependent mobility," *Solid-State Electronics*, vol. 33, pp. 923–933, 1990.
137. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
138. V. M. Agostinelli, Jr., H. Shin, and A. F. Tasch, Jr., "A comprehensive model for inversion layer hole mobility for simulation of submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. 38, p. 151, 1991.
139. H. Shin, G. M. Yeric, A. F. Tasch, and C. M. Maziar, "Physically-based models for effective mobility of electrons in MOS inversion layers," *Solid-State Electronics*, vol. 34, pp. 545–552, 1991.
140. M. Shirahata, H. Kusano, N. Kotani, S. Kusanoki, and Y. Akasaka, "A mobility model including the screening effect in MOS inversion layer," *IEEE Transactions on Computer-Aided Design*, vol. 11, pp. 1114–1119, 1992.
141. V. M. Agostinelli, Jr., G. M. Yeric, and A. F. Tasch, Jr., "Universal MOSFET hole mobility degradation models for circuit simulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 12, pp. 439–445, 1993.

142. B. Iñiguez and E. G. Moreno, "An improved C_∞ continuous small-geometry MOSFET model for analog applications," *Journal of Analog Integrated Circuits and Signal Processing*, vol. 13, no. 3, pp. 241–259, 1997.
143. C.-L. Huang, J. V. Faricelli, and N. D. Arora, "A new technique for measuring MOSFET inversion layer mobility," *IEEE Transactions on Electron Devices*, vol. 40, pp. 1134–1139, 1993.
144. H.-S. Wong, "Universal effective mobility of empirical local mobility models for *n*- and *p*-channel silicon MOSFETs," *Solid-State Electronics*, vol. 36, pp. 179–188, 1993.
145. C. Yue, M. Agostinelli, Jr., G. M. Yeric, and A. F. Tasch, "Improved universal MOSFET electron mobility degradation models for circuit simulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 12, pp. 1542–1546, 1993.
146. S.-I. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFETs: Part I—Effects of substrate impurity concentration," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2357–2362, 1994.
147. S.-I. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFETs, Part II—Effects of surface orientation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2363–2368, 1994.
148. K. Yang, R. C. Meitzler, and A. G. Andreou, "A model for MOS effective channel mobility with emphasis in the subthreshold and transition region," *Proceedings of the IEEE 1994 International Symposium on Circuits and Systems*, vol. 1, pp. 431–434.
149. A. Emrani, G. Ghibaudo, and F. Balestra, "On the universal electric field dependence of the electron and hole effective mobility in MOS inversion layers," *Solid-State Electronics*, vol. 37, pp. 111–113, 1994.
150. F. Gámiz, J. Banqueri, J. E. Carceller, and J. A. L-Villanueva, "Effects on bulk-impurity and interface-charge on the electron mobility in MOSFETs," *Solid-State Electronics*, vol. 38, pp. 611–614, 1995.
151. F. Gámiz, J. A. L.-Villanueva, J. Banqueri, J. E. Carceller, and P. Cartujo, "Universality of electron mobility curves in MOSFETs: A Monte Carlo study," *IEEE Transactions on Electron Devices*, vol. 42, pp. 258–265, 1995.
152. F. Gámiz, J. A. L.-Villanueva, J. Banqueri, and J. E. Carceller, "Influence of the oxide-charge distribution profile on electron mobility in MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-42, pp. 999–1004, 1995.
153. J. Banqueri, J. A. L.-Villanueva, F. Gámiz, J. E. Carceller, E. L.-Tamayo, and M. Lozano, "A procedure for the determination of the effective mobility in an *n*-MOSFET in the moderate inversion region," *Solid-State Electronics*, vol. 39, pp. 875–883, 1996.
154. J. R. Hauser, "Extraction of experimental mobility data for MOS devices," *IEEE Transactions on Electron Devices*, vol. 43, pp. 1981–1988, 1996.
155. S. Villa, A. L. Lacaita, L. Perron, and R. Bez, "Effective mobility in heavily doped n-MOSFETs: Measurements and models," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 395–398, 1996.
156. S. Jallepalli, W.-K. Shih, J. D. Bude, M. R. Pinto, C. M. Mazier, and A. F. Tasch, Jr., "Understanding the differences in the effective-field dependence of electron and hole inversion layer mobilities," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 391–394, 1996.
157. B. Cheng and J. Woo, "A temperature-dependent MOSFET inversion layer carrier mobility model for device and circuit simulation," *IEEE Transactions on Electron Devices*, vol. 44, pp. 343–345, 1997.
158. M. H. White, F. Van de Wiele, and J. P. Lambot, "High-accuracy models for computer-aided design," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 899–906, 1980.
159. M. Miura-Mattausch, H. J. Mattausch, and T. Ezaki, *The Physics and Modeling of MOSFETs*, World Scientific, River Edge, N.J., 2008.
160. S. Villa, A. L. Lacaita, L. M. Perron, and R. Bez, "A physically-based model of the effective mobility in heavily-doped n-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 45, no. 1, pp. 110–115, 1998.
161. C.-L. Huang and N. D. Arora, "Characterization and modeling of the *n*- and *p*-channel MOSFETs inversion-layer mobility in the range 25–125°C," *Solid-State Electronics*, vol. 37, no. 1, pp. 97–103, 1994.
162. BSIM website. Available at <http://www-device.eecs.berkeley.edu/~bsim3/>. Accessed September 29, 2009.
163. M. Bucher, *Analytical MOS Transistor Modelling for Analog Circuit Simulation*, Ph.D. Thesis no. 2114, École Polytechnique Fédérale de Lausanne, Switzerland, 1999.

164. G. Merckel, "CAD models of MOSFETs," *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
165. P. Antognetti, C. Lombardi, and D. A. Antoniadis, "Use of 2-D MOS simulation in the study of doping profile influence on S/D resistance in short-channel MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, 1981, pp. 574–577.
166. G. Baccarani and G. A. Sai-Halasz, "Spreading resistance in submicron MOSFETs," *IEEE Electron Device Letters*, vol. EDL-4, pp. 27–29, February 1983.
167. K. K. Ng, R. J. Bayruns, and S. C. Fang, "The spreading resistance of MOSFETs," *IEEE Electron Device Letters*, vol. EDL-6, pp. 195–198, April 1985.
168. K. K. Ng and W. T. Lynch, "Analysis of the gate-voltage-dependent series resistance of MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 965–972, 1986.
169. K. K. Ng and W. T. Lynch, "The impact of series resistance on MOSFET scaling," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 503–509, 1987.
170. C. J. Koeneke and W. T. Lynch, "Lightly doped Schottky MOSFET," *Proceedings of the IEEE International Electron Devices Meeting*, San Francisco, 1982, p. 466.
171. G. Giralt, B. Andre, J. Simonne, and D. Esteve, "Influence de la température sur les dispositifs semiconducteurs du type M.O.S.," *Electronics Letters*, vol. 1, pp. 185–186, 1965.
172. R. Wong, J. Dunkley, T. A. DeMassa, and J. F. Jelsma, "Threshold voltage variations with temperature in MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-18, p. 386, 1971.
173. S. K. Tewksbury, "N-channel enhancement-mode MOSFET characteristics from 10 to 300 K," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 1519–1529, 1981.
174. F. Shoucair, W. Hwang, and P. Jain, "Electrical characteristics of large scale integration (LSI) MOSFETs at very high temperatures," *Microelectronics and Reliability*, vol. 24, part I, pp. 465–485, part II, pp. 487–510, 1984.
175. F. M. Klaassen and W. Hes, "On the temperature coefficient of the MOSFET threshold voltage," *Solid-State Electronics*, vol. 29, pp. 787–789, 1986.
176. G. S. Gildenblat and C.-L. Huang, "N-channel MOSFET model for the 60–300-K temperature range," *IEEE Transactions on Computer-Aided Design*, vol. 10, pp. 512–518, 1991.
177. F. Balestra and G. Ghibaudo, "Brief review of the MOS device physics for low temperature electronics," *Solid-State Electronics*, vol. 37, pp. 1967–1975, 1994.
178. W. Chen, P. Su, and K. Goto, "Impact of process-induced strain on Coulomb scattering of mobility in short-channel n-MOSFETs," *IEEE Electron Device Letters*, vol. 29, no. 7, pp. 768–770, 2008.

Problems

- 4.1.** Plot the total drain current, the component due to drift, and the component due to diffusion, using relations from Sec. 4.3, for an *n*-channel transistor with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2 \text{ nm}$, $V_{FB} = -0.75 \text{ V}$, $\mu = 400 \text{ cm}^2/(\text{V} \cdot \text{s})$, $W = 1 \mu\text{m}$, $L = 0.1 \mu\text{m}$, biased with $V_{SB} = 0.5 \text{ V}$ and $V_{DB} = 1.5 \text{ V}$, for V_{GB} between 0 and 1.5 V. Use a log current axis. Identify the three regions of inversion on the plot and evaluate the percentages of the current due to drift and due to diffusion at the limit points between weak and moderate inversion and between moderate and strong inversion.
- 4.2.** As explained in Sec. 4.3, in using (4.3.14b) in weak inversion a small error in each of ψ_{sL} and ψ_{s0} can cause a large relative error in I_{DS2} , since ψ_{sL} and ψ_{s0} have nearly equal values. Show that if $\psi_{sL} - \psi_{s0}$ in (4.3.14b) is substituted by the difference of the right-hand sides of (4.3.15a) and (4.3.15b), an expression much more immune to error results.⁵² Explain why the new expression obtained may not be acceptable for use with (4.3.7) in strong inversion if ψ_{s0} and ψ_{sL} are not known accurately. (See also Prob. 4.4.)
- 4.3.** Compare the simplified all-region model of (4.4.10) to the complete all-region model of Sec. 4.3 by constructing current plots for the parameters and voltage values specified in Prob. 4.1.

- 4.4.** Consider the device of Prob. 4.1 with $V_{SB} = 0.5$ V, $V_{DB} = 1.5$ V, and V_{GB} corresponding to 0.25 V below the upper limit of weak inversion. Find ψ_{s0} and ψ_{sL} from (4.3.15) to very good accuracy (e.g., to eight significant digits) by using a computer. Then, find I_{DS2} from (4.3.14b) and from the expression of Prob. 4.2, assuming ψ_{s0} and ψ_{sL} are known to only five, four, three, and two significant digits each. Which expression is more immune to limited accuracy? Explain the reasons by using specific numbers from the computation.
- 4.5.** Plot the inversion layer and depletion region charges per unit area vs. distance from the source for the device of Prob. 4.1, with $V_{SB} = 0.5$ V, $V_{DB} = 1$ V, and $V_{GB} = 1.5$ V.
- 4.6.** For the device of Prob. 4.1, indicate the regions of inversion
 (a) In the manner of Fig. 4.18, for $V_{SB} = 0$ V and 1 V.
 (b) In the manner of Fig. 4.19, for $V_{GB} = 1$ V and 1.5 V.
- 4.7.** This problem deals with various ways of defining regions of operation.
 (a) Without restricting V_{DS} to positive values, define $V = \min(V_{SB}, V_{DB})$ and give general definitions for the regions of inversion as described in Table 4.1, in terms of V and V_{GB} .
 (b) Repeat, in terms of V and V_{GA} , where $V_{GA} = \max(V_{GS}, V_{GD})$.
- 4.8.** Starting from (4.7.2a), prove (4.7.2b), as well as the following equation:

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left\{ \frac{1}{2} \left[(V_{GS} - V_{FB} - \phi_0)^2 - (V_{GD} - V_{FB} - \phi_0)^2 \right] - \frac{2}{3} \gamma \left[(V_{DB} + \phi_0)^{3/2} - (V_{SB} + \phi_0)^{3/2} \right] \right\}$$

- 4.9.** Starting from (4.7.2), show that the nonsaturation strong-inversion current can be written in terms of V_{GS} , V_{DS} , and V_{SB} as follows:

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left\{ (V_{GS} - V_{FB} - \phi_0) V_{DS} - \frac{1}{2} V_{DS}^2 - \frac{2}{3} \gamma \left[(\phi_0 + V_{SB} + V_{DS})^{3/2} - (\phi_0 + V_{SB})^{3/2} \right] \right\}$$

and find an expression for V'_{DS} , the value of V_{DS} at the boundary between nonsaturation and saturation.

- 4.10.** Extend the body-referenced simplified strong-inversion model of Sec. 4.7.2 to the saturation region. Propose a V_{DB} value for the limit between nonsaturation and saturation.
- 4.11.** For the device of Prob. 4.1 with $V_{SB} = 0$ V and $V_{GB} = 1.5$ V (strong inversion), plot I_{DS} vs. V_{DS} , from $V_{DS} = 0$ to $V_{DS} = 1.5$ V, using (a) the general model of Sec. 4.3 and (b) the complete strong inversion model of Sec. 4.7.1. Discuss the effect of the value of ϕ_0 in (b).
- 4.12.** Consider the bracketed quantity containing the 3/2 powers in Prob. 4.9. Starting from an expansion of this quantity into a Taylor series around $V_{DS} = 0$, derive (4.7.18). Prove that the value to be used for α should be less than α_1 in (4.7.27).
- 4.13.** Show that, for (4.7.17) to correspond to line a in Fig. 4.15, α_1 should have the value given in (4.7.27).
- 4.14.** For the transistor in Prob. 4.1, assume $V_{SB} = 0$ V and plot I_{DS} vs. V_{DS} for $V_{GS} = 1$ and 1.5 V, using (a) (4.7.11) with (4.7.2) and (b) (4.7.24), first using α from (4.7.27), and then modifying the value of α to improve accuracy.
- 4.15.** Show that the assumptions made in order to derive the simplified strong-inversion model of Sec. 4.7.3 lead to the conclusion that $Q'_t = 0$ when $V_{DS} = V'_{DS}$. Comment on the validity of this conclusion.

- 4.16.** Read the material on pinchoff-based models in Appendix G, and verify the drain current equations given there [except for (G.21)].
- 4.17.** Prove (4.7.33).
- 4.18.** For the simplified strong-inversion model of Sec. 4.7.3 and for the device of Prob. 4.1, $V_{SB} = 0.5$ V, $V_{GS} = 1.5$ V, and $V_{DS} = V_{DS}' - 0.3$ V, plot (a) Q'_B and Q'_I as functions of V_{CB} , and (b) V_{CB} , Q'_B , and Q'_I as functions of the distance from the source. Use α from (4.7.28) with $d_2 = 1$ and $\phi_3 = 1$ V.
- 4.19.** For the device of Prob. 4.1 in weak inversion, plot (a) I_{DS} vs. V_{DS} , with $V_{SB} = 0$, for V_{DS} between 0 and 0.5 V, and for V_{GS} values of $V_M - 50$ mV and $V_M - 100$ mV. Use (4.8.6) and (4.8.7); (b) $\log I_{DS}$ vs. V_{GS} with $V_{DS} = 1$ V, for $V_{SB} = 0$ V and $V_{SB} = 0.5$ V.
- 4.20.** Plot $\log I_{DS}$ vs. V_{GS} , for values as stated in Prob. 4.19, using (4.8.6) and (4.8.7) and (4.8.11). Compare the results. Can the agreement be improved if the values of V_M and n are slightly modified?
- 4.21.** Show that a simple approximate expression for the current in weak inversion is

$$I_{DS} = \frac{W}{L} I'_X e^{(V_{GS} - V_X)/(n\phi_t)} \left(1 - e^{-V_{DS}/\phi_t}\right)$$

where V_X is the value of V_{GS} corresponding to a surface potential of $1.5\phi_F + V_{SB}$, and

$$I'_X = \mu C'_{ox} \phi_t^2 \frac{\gamma}{2\sqrt{1.5\phi_F + V_{SB}}} e^{-0.5\phi_t/\phi_t}$$

Propose an appropriate expression for n . Comment on the accuracy of this equation in comparison with (4.8.11), and show that both equations can provide similar results if the values of the parameters in one of them are slightly modified.

- 4.22.** Show that (4.3.14) in weak inversion can be reduced to (4.8.6) and (4.8.7) as follows: Replace $\psi_{sL} - \psi_{s0}$ in (4.3.14b) by the difference of the right-hand sides of (4.3.15b) and (4.3.15a). In this difference expand the square roots of the form $\sqrt{\psi_s + \xi}$ in a Taylor series around $\xi = 0$ and keep only the first two terms (justify this); in the resulting expression for I_{DS2} , use $\psi_{s0} \approx \psi_{sL} \approx \psi_{sa}$.
- 4.23.** Plot I_{DS} in nonsaturation vs. V_{GS} using the source-referenced simplified strong-inversion model of Sec. 4.7.3 with $V_{SB} = 0$, $V_{DS} = 10$ mV, and V_{GS} between 0.6 and 1.5 V, assuming $(W/L)\mu C'_{ox} = 450$ $\mu\text{A/V}^2$, $N_A = 7 \times 10^{17}$ cm^{-3} , $t_{ox} = 2$ nm, and $V_{FB} = -1$ V. Now repeat, assuming a mobility dependence on V_{GS} as given by (4.11.31), with $\theta = 0.4$ V^{-1} . Can V_T be obtained from the second plot by extrapolating to $I_{DS} = 0$ from the point of maximum slope?
- 4.24.** For the device of Prob. 4.23, plot I_{DS} vs. V_{DS} with V_{DS} between 0 and 1.5 V, for $V_{GS} = 1$ and 1.5 V.
 (a) Assume a constant mobility.
 (b) Assume a V_{GS} -dependent mobility as given by (4.11.31) with $\theta = 0.4$ V^{-1} .
- 4.25.** Rewrite the framed equations in this chapter for *p*-channel transistors.
- 4.26.** Show that the effect of extrinsic source and drain series resistances can be modeled as in (4.12.2), starting from (4.12.1) and making approximations as discussed in the text. Also, show that the quantity β_R can be combined with the quantity θ used to characterize mobility, in the manner discussed following (4.12.3).

5

Small-Dimension Effects

5.1 Introduction

In the previous chapter, we based the development of transistor equations on a number of assumptions, e.g., that the electric field in the longitudinal direction is low, so that the velocity of the carriers is proportional to it; that the current is not affected as the drain voltage is increased above its saturation value; that the channel is sufficiently long and wide, so that “edge” effects along the four sides of the channel could be neglected and so that a one-dimensional analysis could be used; that the gate oxide is a perfect insulator; and that the gate and substrate currents are zero.

In real devices, especially ones with very small dimensions, such assumptions break down, and lead to effects we have not discussed so far. One is then faced with the task of determining how the resulting characteristics differ from those predicted by the “ideal” device equations (i.e., the ones we have presented in Chap. 4) and whether those equations can be “saved” with minor modifications in them to include the new effects. This task is the subject of the present chapter. Thus, here we will study what happens when the longitudinal field is so high that the carrier velocity saturates; we will consider the effect of the drain voltage on the saturation current; we will allow the channel to become short and/or narrow; and we will study the resulting “edge” effects and how they affect the drain current. We will discuss what happens when the magnitude of the electric field is so large that it imparts a large energy on the carriers, which are then said to be “hot” and can result in substrate current and oxide damage. We will also consider what happens when the gate oxide is very thin, in which case certain phenomena we have so far neglected must be

taken into account. These include polysilicon gate depletion, quantum effects in the inversion layer, and tunneling, which gives rise to dc gate current. Finally, we will discuss drain-body junction leakage and the influence of the gate on it. We will also describe how, traditionally, the device dimensions have shrunk from one technology generation to the next, using a process called “classical scaling,” and we will discuss why this process must now be modified.

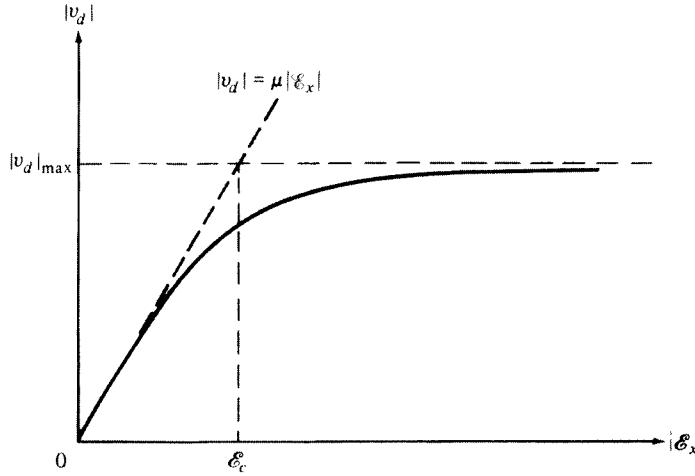
Most of the phenomena we will discuss require, in principle, complicated analyses that can be carried out numerically with the help of a computer.^{1–15} However, such analyses, albeit accurate, do not provide a simple model for efficient calculation. Thus, empirical approximations and semiempirical approaches have been used to provide useful simplification. In these, usually the complex two- or three-dimensional phenomena are broken down into simple, separate phenomena examined one at a time. A number of simplifying assumptions are then made, *which are sometimes difficult to justify rigorously*, and relatively simple relations are derived. Often such techniques are characterized by an attempt to maintain the general form of the *I-V* relations for the long- and wide-channel devices and to “stretch” these relations by modifying them somewhat so that they can be used in the case of short and/or narrow channels. What is considered to justify these empirical approaches is their success in simulating the behavior observed experimentally. Although some of these approaches necessarily lack rigor and elegance, they have often been helpful where more precise work failed to give computationally efficient models. Some representative examples of the semiempirical modeling process will be presented in this chapter.

5.2 Carrier Velocity Saturation

The phenomenon of velocity saturation is associated with drift currents, and as such it will be discussed in the context of strong inversion, where such currents are dominant. Let us first recall an assumption behind all the strong-inversion nonsaturation models we have considered so far. Let \mathcal{E}_x be the value of the *longitudinal* component of the electric field in the semiconductor, i.e., the component parallel to the semiconductor-insulator interface. We have assumed that at all points in the inversion layer $|\mathcal{E}_x|$ is small enough so that the magnitude of the carrier velocity $|v_d|$ is proportional to $|\mathcal{E}_x|$. Although this assumption was useful in deriving our basic models, it is not accurate for devices with short channels, or even for long-channel devices at or near saturation. In this section we will show how adequate modeling can be achieved in such cases.^{19–35} For simplicity in the development that follows, we will assume that $V_{DS} \geq 0$ unless stated otherwise.

Figure 5.1 illustrates the behavior of $|v_d|$ with $|\mathcal{E}_x|$. As was the case for bulk conduction (Sec. 1.3), the velocity of carriers in the inversion layer tends to saturate at high $|\mathcal{E}_x|$ values.[†] Effects due to the lack of proportionality between $|v_d|$ and $|\mathcal{E}_x|$ on device characteristics are often referred to as *velocity saturation effects*, although

[†] The velocity can attain values larger than $|v_d|_{\max}$ over very short times or very short distances. This effect, referred to as *velocity overshoot*,^{16–18} is considered in Sec. 5.9. The physics behind velocity saturation are complicated; readers with adequate background in solid-state physics can consult Ref. 19.

**FIGURE 5.1**

Magnitude of carrier velocity in the inversion layer vs. magnitude of longitudinal component of electric field, $|\mathcal{E}_x|$. μ is the surface mobility (Sec. 4.11).

the $|\mathcal{E}_x|$ values involved may be below those corresponding to the clear saturation part of the figure.

It is convenient to define a “critical” value of $|\mathcal{E}_x|$, denoted by \mathcal{E}_c , at the intersection of the $|v_d| = \mu |\mathcal{E}_x|$ line and an imaginary horizontal asymptote, as shown in the figure. We have

$$|v_d| \approx \mu |\mathcal{E}_x|, \quad |\mathcal{E}_x| \ll \mathcal{E}_c \quad (5.2.1a)$$

$$\approx |v_d|_{\max}, \quad |\mathcal{E}_x| \gg \mathcal{E}_c \quad (5.2.1b)$$

and

$$\mathcal{E}_c = \frac{|v_d|_{\max}}{\mu}$$

(5.2.2)

The quantity μ in the preceding relations is the *surface* mobility, i.e., the mobility of carriers next to the semiconductor-insulator interface; this quantity is significantly lower than the bulk mobility (e.g., by a factor of 2). As indicated in (4.11.6), this quantity depends on the *transversal* electric field, $\mathcal{E}_{y,\text{ave}}$ (also called the *normal* field).

Reported values for $|v_d|_{\max}$ vary; at room temperature, $|v_d|_{\max}$ is around 8×10^6 cm/s for electrons and 7×10^6 cm/s for holes.³⁶ The value of $|v_d|_{\max}$ is relatively independent of $\mathcal{E}_{y,\text{ave}}$.[†] Typical values of the “critical field” \mathcal{E}_c are

[†] Because μ depends on $\mathcal{E}_{y,\text{ave}}$, this means that \mathcal{E}_c in (5.2.2) is not a constant but rather depends on $\mathcal{E}_{y,\text{ave}}$.

10^4 V/cm for electrons and 3×10^4 V/cm for holes. Most MOSFET models allow the value of $|v_d|_{\max}$, and/or that of the “critical field” \mathcal{E}_c , to be chosen for best fit of model predictions to measured drain current values. The reason for this is that these parameters are difficult to measure. Also, in MOS transistor work, these theories are often applied in a simplified manner anyway, and this fact can be compensated to some degree by choosing parameter values for best fit with measurements.

In strong inversion, it is possible to include fine details of the position dependence of carrier velocity along the channel when solving for the drain current.^{23,37} However this approach has not been adopted in practical models; we will instead discuss a couple of ways that can be used to include velocity saturation in a simple, approximate manner. The following is meant to illustrate the principles, and not to lead to complete models, ready for use in CAD. In fact, proper handling of velocity saturation effects in MOSFET models for CAD can be tricky; related considerations are discussed in Chap. 10.

A simple velocity saturation model is^{30,32}

$$|v_d| = |v_d|_{\max} \frac{|\mathcal{E}_x| / \mathcal{E}_c}{1 + |\mathcal{E}_x| / \mathcal{E}_c} \quad (5.2.3)$$

It is easily seen that this relation correctly models the asymptotic behavior in (5.2.1). Although it only does an approximate job for moderate fields (fields around \mathcal{E}_c in Fig. 5.1), this relation has been widely used for MOS transistor modeling; the primary reason is that, as will be seen, it leads to simple closed form solutions. Consider operation in strong-inversion nonsaturation; as in Sec. 4.7.1, we will assume that all current is due to drift. The magnitude of the longitudinal electric field is $d\psi_s/dx$. Using (4.7.4), and assuming ϕ_0 is constant as usual, this gives

$$|\mathcal{E}_x(x)| = \frac{dV_{CB}}{dx} \quad (5.2.4)$$

Thus, the magnitude of the electron velocity at point x becomes, from (5.2.3),

$$|v_d(x)| = |v_d|_{\max} \frac{(1/\mathcal{E}_c)(dV_{CB}/dx)}{1 + (1/\mathcal{E}_c)(dV_{CB}/dx)} \quad (5.2.5a)$$

or, using (5.2.2):

$$|v_d(x)| = \mu \frac{dV_{CB}/dx}{1 + (1/\mathcal{E}_c)(dV_{CB}/dx)} \quad (5.2.5b)$$

To calculate the drain current in nonsaturation, I_{DSN} , we cannot use equations like (4.3.3a) or (4.7.5) anymore. Those equations were derived by assuming that $|v_d|$ was proportional to $|\mathcal{E}_x|$ [see eq. (1.3.8) and subsequent development]. Since this is not the case anymore, we will use the more general expression (1.3.7). For the MOS transistor, Q' in that expression is the inversion layer charge per unit area Q'_I , the value of which is negative and depends on $V_{CB}(x)$, and b is the channel width W . Thus

$$I_{DSN} = W(-Q'_I) |v_d(x)| \quad (5.2.6)$$

Using (5.2.5b) in (5.2.6) we obtain

$$I_{DSN} \left(1 + \frac{1}{\mathcal{E}_c} \frac{dV_{CB}}{dx} \right) = \mu W (-Q'_I) \frac{dV_{CB}}{dx} \quad (5.2.7)$$

Integrating from $x = 0$ (where $V_{CB} = V_{SB}$) to $x = L$ (where $V_{CB} = V_{DB}$), and assuming μ and \mathcal{E}_c are independent of the transverse field, we obtain

$$I_{DSN} \left(L + \frac{V_{DB} - V_{SB}}{\mathcal{E}_c} \right) = \mu W \int_{V_{SB}}^{V_{DB}} (-Q'_I) dV_{CB} \quad (5.2.8)$$

Using $V_{DB} - V_{SB} = V_{DS}$, we have

$$I_{DSN} = \frac{W}{L} \frac{\mu}{1 + V_{DS}/(L\mathcal{E}_c)} \int_{V_{SB}}^{V_{DB}} (-Q'_I) dV_{CB} \quad (5.2.9)$$

Comparing now this equation to (4.7.6), with μ assumed constant, we see that they are identical except for the multiplicative factor $1/[1 + V_{DS}/(L\mathcal{E}_c)]$ outside the integral. Consequently, nonsaturation region expressions developed in Chap. 4 can be extended to apply in the presence of velocity saturation effects by multiplying them with this factor. Thus

$$I_{DSN, \text{including velocity saturation}} = \frac{I_{DSN, \text{not counting velocity saturation}}}{1 + V_{DS}/(L\mathcal{E}_c)} \quad (5.2.10)$$

This is a very simple and general relationship that has been observed to fit measured device currents relatively well and so has been widely used. However, it does not keep track of the details of the transverse field dependence of μ and \mathcal{E}_c , and how they alter the surface potential and charge densities with position along the channel; so, some imprecision in the preceding relation can be expected. We remind the reader

that the derivation has assumed operation in strong inversion; in weak inversion and in the lower part of moderate inversion, the effect of velocity saturation is small as current flow is dominated by diffusion rather than drift.

Example 5.1. Using the strong-inversion nonsaturation model of (4.7.24a) in (5.2.10), we have for the current in the presence of velocity saturation effects,

$$I_{DS} = \frac{W \mu C'_{ox} [(V_{GS} - V_T)V_{DS} - 0.5\alpha V_{DS}^2]}{L \left[1 + V_{DS}/(L\mathcal{E}_c) \right]}, \quad V_{DS} \leq V'_{DS} \quad (5.2.11)$$

Because the effect of velocity saturation is to reduce the drain current and its slope with respect to V_{DS} , V'_{DS} , the value of V_{DS} at which $\partial I_{DS}/\partial V_{DS} = 0$, will no longer be given by (4.7.21), but is instead smaller and must be found directly from the new nonsaturation current expression.

Example 5.2. For the model of Example 5.1, solving for the value V'_{DS} of V_{DS} that makes $\partial I_{DS}/\partial V_{DS} = 0$ gives (Prob. 5.2)

$$V'_{DS} = \frac{V_{GS} - V_T}{\alpha} \frac{2}{1 + \sqrt{1 + \left(\frac{V_{GS} - V_T}{\alpha} \right) \frac{2}{L\mathcal{E}_c}}} \quad (5.2.12)$$

This is, as expected from our previous discussion, smaller than the value $(V_{GS} - V_T)/\alpha$ found in the absence of velocity saturation. It only approaches that limiting value as $L\mathcal{E}_c$ approaches infinity, i.e., in the absence of velocity saturation effects.[†]

For more complex models than the strong-inversion nonsaturation form (4.7.24a), direct application of (5.2.10) and determination of V'_{DS} is not as straightforward. For such models, simplifications must be used that lead to an approximate value for V'_{DS} . Beyond that point, extra corrections need to be applied, notably to take care of “channel length modulation,” which is discussed in Sec. 5.3. As will be discussed, saturation behavior happens smoothly, and in experimental data, it is not possible to identify a distinct “point” of saturation. Therefore it does not make much sense to seek an “exact” value for V'_{DS} ; it is reasonable to use an approximate value for this quantity. In CAD models, an “effective” drain-source voltage, $V_{DS,\text{eff}}$, is used in lieu of V'_{DS} ; that voltage is close to V_{DS} for low value of the latter and smoothly saturates to V'_{DS} as V_{DS} is raised. This makes possible the smooth modeling of saturation behavior. More details can be found in Chap. 10.

[†] An alternative approach to determining V'_{DS} is to equate (5.2.11) to (5.2.6) evaluated at the drain with $|v_d(L)| = |v_d|_{\max}$ and solve for $V_{DS} = V'_{DS}$. The result is the same. A more detailed calculation,²⁶ aided by numerical simulation results, leads to a complicated expression for V'_{DS} that, however, results in practically the same value as the approach suggested here.

The effect of the normal field (Sec. 4.11) can be incorporated in the preceding formulation. Assuming independence between mobility and saturation velocity,¹⁹ one can replace μ in (5.2.1) and (5.2.2) by (4.11.16), and rederive I_{DSN} using the approximations suggested in Sec. 4.11. One then obtains again the preceding expressions, only with μ replaced by the effective mobility μ_{eff} . Other approaches²³ give similar results.

A more general model than (5.2.3) is^{38,39}

$$|v_d| = |v_{d,\max}| \frac{|E_x|/\mathcal{E}_c}{\left[1 + \left(\frac{|E_x|/\mathcal{E}_c}{n}\right)^n\right]^{1/n}} \quad (5.2.13)$$

which reduces to the model of (5.2.3) for $n = 1$. The quantity n is sometimes used as a fitting parameter; for holes, $n = 1$ is often used. For electrons, usually $n = 2$ is used, which leads to desirable behavior as will be seen shortly.^{†40} Using (5.2.2), (5.2.4), and $n = 2$ in (5.2.13), we obtain

$$|v_d| = \mu \frac{dV_{CB}/dx}{\sqrt{1 + \left(\frac{dV_{CB}/dx}{\mathcal{E}_c}\right)^2}} \quad (5.2.14)$$

This can now be used in (5.2.6) and can be shown to result in the following equation, instead of (5.2.10):^{41,42}

$$I_{DSN, \text{with velocity saturation}} = \frac{I_{DSN, \text{without velocity saturation}}}{0.5 \left\{ 1 + \sqrt{1 + 2 \left[V_{DS} / (L \mathcal{E}_c) \right]^2} \right\}} \quad (5.2.15)$$

Numerical Issues Independent of accuracy comparisons, the previous relation has a definite advantage in comparison with (5.2.10). This is now discussed. In our development so far, we have for simplicity assumed that $V_{DS} \geq 0$ and have used absolute values in (5.2.3). The resulting equation, (5.2.10), is thus valid as long as $V_{DS} \geq 0$. If $V_{DS} < 0$, i.e., the usual roles of source and drain are interchanged, V_{DS} in that equation should be replaced by $-V_{DS}$. To make the equation valid for any polarity, one can replace V_{DS} in it by $|V_{DS}|$. But there is a catch. Although the equation will then be valid for the current, the derivatives of that equation with respect to V_{DS} , starting with the second-order one, will be problematic. This can be traced to the fact that the derivative of $|V_{DS}|$ with respect to V_{DS} is $+1$ for $V_{DS} > 0$ and -1 for $V_{DS} < 0$; the derivative does not exist at $V_{DS} = 0$. In general, discontinuities and singularities can wreak havoc

† A slightly different form has been claimed to lead to more accurate simulation of velocity saturation for *p*-channel devices.^{39,43}

in circuit simulations (Chap. 10); they can cause numerical problems, can result in unnatural jumps in the small-signal parameters (Chap. 7), and can cause grave errors in the evaluation of distortion. Distortion, a measure of nonlinearity, can be evaluated by considering a Taylor series expansion of a nonlinear function; the second- and higher-order derivatives are measures of such nonlinearity. For example, this issue is of central importance in the simulation of passive mixers in radio frequency circuits, in which MOSFETs are used with a nominal bias of $V_{DS} = 0$.⁴⁴ Such simulation using (5.2.10) with V_{DS} replaced by $|V_{DS}|$ will completely fail because the second and higher derivatives of that equation are discontinuous at $V_{DS} = 0$. Therefore, although this model might be satisfactory for the analysis of most digital circuits, it could create serious problems in the analysis of analog and RF circuits. In contrast to this, (5.2.15) has no such problems; this equation is smooth through $V_{DS} = 0$ and so avoids the problems with nonexistence of derivatives. This form has been demonstrated to accurately model derivatives to at least the third order.⁴³

Formulations associated with the body-referenced simplified all-region model presented in Sec. 4.4.2 are discussed elsewhere.^{41,44}

Nonsaturation-to-Saturation Transition So far we have considered velocity saturation as a perturbation to the long-channel behavior, discussed in Chap. 4, and have obtained V'_{DS} by setting $\partial I_{DS}/\partial V_{DS}$ equal to zero. The origin of this practice can be found in common strong-inversion approximations, which assume that saturation is entered when the drain end of the channel gets “pinched off,” i.e., the inversion layer charge density becomes zero at that point. However, this picture is not accurate. Consider (5.2.6). Because the current is continuous along the channel, the product of Q'_l and v_d must be constant at any position in the channel in order to satisfy that equation.[†] If Q'_l became zero near the drain, this would imply that v_d would have to be infinite at that point, which, as we know, cannot happen. A more accurate picture is the following.

Assume $V_{DS} > 0$. In strong inversion, the channel-to-body potential, V_{CB} , increases from source to drain, as has been illustrated in Fig. 4.27. Thus, as can be deduced from Fig. 3.17, the magnitude of the inversion charge density falls from the source to the drain. Because the current at any point along the channel is given by the right-hand side of (5.2.6), $|v_d|$ must be increasing from source to drain. As we have seen, $|v_d|$ cannot increase without bound, but rather saturates to a value $|v_{d,\max}|$. This means that, to maintain current continuity, the inversion charge density magnitude near the drain cannot fall to an arbitrarily low level as V_{DS} increases, but must decrease in a self-consistent manner that ensures the product $v_d Q'_l$ remains constant. It is the balance of inversion charge density reduction and saturation of the velocity to $v_{d,\max}$, to maintain continuity of I_{DS} in the vicinity of the drain, that determines saturation behavior. This is a gradual, transitional process, although for the purpose of analysis we often consider the boundary between nonsaturation and saturation regions as occurring at a specific point. Serendipitously, for long-channel

[†] We assume no impact ionization and no generation/recombination in the channel.

devices this boundary is rather distinct and is well approximated by the models we have presented so far. For short-channel devices, where velocity saturation becomes important, the boundary between nonsaturation and saturation behavior is indistinct and therefore difficult to model precisely. Measurements also bear out this gradual transition. Thus, although for short-channel devices our analyses lack precision, the boundary between nonsaturation and saturation regions is blurred anyway, so some imprecision turns out to be acceptable. The transition from nonsaturation to saturation is further complicated due to other short-channel effects, which we will discuss Secs. 5.3 and 5.4.2.

As we have already mentioned, velocity saturation is a phenomenon associated with drift currents and is thus mostly important in strong inversion. In weak inversion, current conduction is due to diffusion; the lateral electric field in the channel is, in the ideal case, zero (Sec. 4.8), and thus no drift current and velocity saturation are expected to occur. In general models, which are valid in all regions of inversion, one must make sure that the velocity saturation modeling equations do not influence weak-inversion modeling. This can be done by making the electric field dependent not on the difference between the drain and source external potentials, but rather on the difference between the corresponding surface potentials, ψ_{sL} and ψ_{s0} . In strong inversion, $\psi_{sL} - \psi_{s0}$ is equal to $V_{DB} - V_{SB} = V_{DS}$, as can be seen from (4.7.1). In weak inversion, the surface potential does not change along the channel, so $\psi_{sL} - \psi_{s0}$ goes to zero, and velocity saturation is not activated. In the moderate inversion region, this difference is not zero, so velocity saturation does have some effect, as it should, given that in that region both drift and diffusion currents are important.

Figure 5.2 compares $I_{DS}(V_{DS})$ characteristics without and with the presence of velocity saturation effects, for an nMOS transistor with a length of 90 nm (other short-channel effects are included in the models used to generate

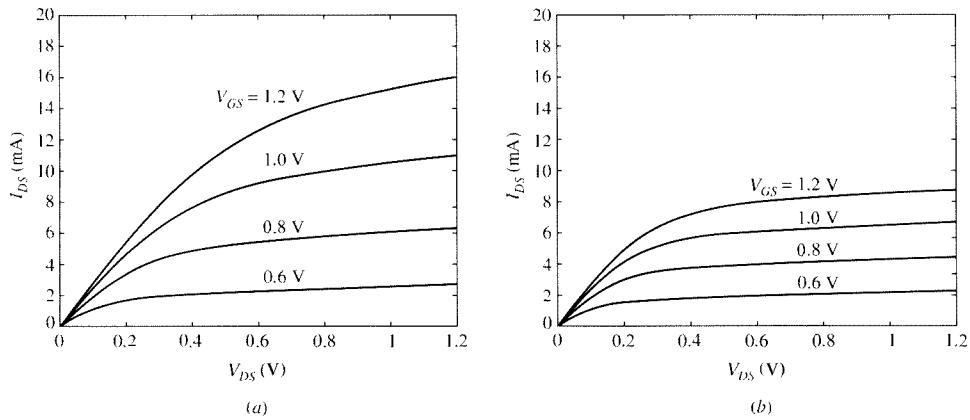


FIGURE 5.2
Characteristics (a) in the absence and (b) in the presence of velocity saturation effects for a device with a 90-nm channel length. Other small dimension effects are included (see Secs. 5.3–5.5). $t_{ox} = 1.5$ nm, $N_A = 3 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $E = 0.09 \mu\text{m}$.

these data; such effects will be discussed in the next several sections). Several features are apparent. As just discussed, the boundary between nonsaturation and saturation regions is indistinct. As noted previously, for the same V_{GS} , saturation is achieved at smaller values of V_{DS} when velocity saturation is included. Equally important, the spacing of the curves in saturation no longer follows the “square law” behavior of (4.7.24b). Notice that this would be true even without velocity saturation effects, because of the effects of mobility dependence on transversal field and source series resistance, as has been discussed in Secs. 4.11 and 4.12; but even without including such effects, it is easy to see that velocity saturation tends to make the spacing of the curves in saturation nearly proportional to the V_{DS} increment. We now have a device with a saturation current that depends almost linearly on $V_{GS} - V_T$! This can be seen from the equations in Examples 5.1 and 5.2, as follows.

In saturation, the current can be found by replacing V_{DS} in (5.2.11) by V'_{DS} . The effects of velocity saturation become stronger as L is decreased. From (5.2.12), V'_{DS} decreases with decreasing L , although not as fast as L itself. When L is very small [small V'_{DS} from (5.2.12)], we can neglect the square term in the numerator of (5.2.11), as well as the 1 in the denominator of that equation. We then have

$$I'_{DS} \approx \frac{\mu C'_{ox} (W/L)(V_{GS} - V_T)V'_{DS}}{V'_{DS}/(L\mathcal{E}_c)}, \quad \text{very small } L \quad (5.2.16a)$$

$$\approx WC'_{ox} (V_{GS} - V_T)\mu\mathcal{E}_c, \quad \text{very small } L \quad (5.2.16b)$$

The linear dependence of I'_{DS} on $V_{GS} - V_T$ is apparent. We can put this relation in an alternative form by postulating that, since V'_{DS} is small, the channel charge will be approximately uniform and independent of x , so that we have $-Q'_I \approx C'_{ox}(V_{GS} - V_T)$ at any point x . Using this fact and using (5.2.2) in (5.2.16b), we have

$$I'_{DS} \approx W(-Q'_I)|v_d|_{\max} \quad (5.2.17)$$

which could have been obtained directly from (5.2.6), assuming the carriers are traveling at maximum velocity from the moment they are injected by the source to the moment they are collected by the drain.

The previous two relations also reveal another major effect in the limit of very strong velocity saturation: *The drain current is independent of L !* A physical feeling for this effect can be obtained as follows. The transit time of carriers in the channel (the time it takes them to travel the length of the channel) is proportional to L , assuming for simplicity that they travel at constant maximum velocity. The total mobile charge in the channel is also proportional to L . Hence the current, which is the ratio of this charge to the transit time, is independent of L . To put it another way, if the distribution of mobile charge is uniform and that charge moves at constant velocity, a fixed amount of charge passes per unit time for a given W . Thus the

drain current is fixed. The length of the channel does not enter in this reasoning at all (provided that L is small enough so that the velocity saturation effect assumed is valid in the first place). An analogy can be given in terms of water coming out of the end of a pipe at constant speed. The water flow rate out of the pipe is independent of the length of the pipe.

Real devices are not fully velocity-saturated, and as such their current is not exactly proportional to $V_{GS} - V_T$; their current is sometimes modeled as being proportional to a power of this quantity between 1 and 2 (with minimum-size devices, this power is closer to 1).⁴⁵ However, equations like (5.2.10) and (5.2.11) already model such devices appropriately, through their denominator.

The approximations used in the preceding four paragraphs are actually oversimplified since, when the channel length is small, the two-dimensional field near the source and drain regions can have a significant influence on the channel charge. The previous discussion is, nevertheless, useful for providing intuition and for relating several of the results we have presented. Note that, because $|v_d|_{\max}$ has similar values for electrons and holes,^{33,36} *n*-channel and *p*-channel devices tend to perform similarly under velocity saturation, other things being equal. This is not necessarily the case in the absence of velocity saturation, i.e., for long transistors, because the surface mobility for *p*-channel devices is approximately one-third that of *n*-channel devices, and I_{DS} is proportional to mobility. However, the introduction of deliberate stressors into fabrication technologies (Sec. 9.6.) has enhanced mobility more for *p*-channel than for *n*-channel devices, so this asymmetry in performance (for comparable biases and geometries) is less pronounced in modern technologies.

5.3 Channel Length Modulation

As mentioned in Sec. 4.7, in the saturation region, the I_{DS} - V_{DS} characteristics are not exactly parallel to the horizontal axis but have, instead, a positive slope.^{20–29,45a–c,46–60} Other things being equal, this slope has been found to be larger for shorter channels and may be easily noticeable, as shown in Fig. 5.2a. Historically, this phenomenon was the first “short-channel effect” to be studied. It was not originally classified as such, partly because its investigation began well before various other short-channel effects were recognized and named that way, and partly because it can play an important role in circuit work even in devices with long channels.[†] One of the reasons for the observed positive slope of I_{DS} - V_{DS} characteristics in long-channel devices is the phenomenon of channel length modulation, which we will discuss in this section. This phenomenon is often a major contributor to the positive I_{DS} - V_{DS} slope in long-channel devices, especially in strong inversion; unless indicated otherwise, we will assume operation in this region in the following descriptions.

[†] An example is the analog “CMOS inverter.” The small-signal voltage gain of such circuits turns out to be inversely proportional to the sum of the I_{DS} - V_{DS} slopes of two devices in saturation. The models of Sec. 4.7 would thus predict infinite gain!

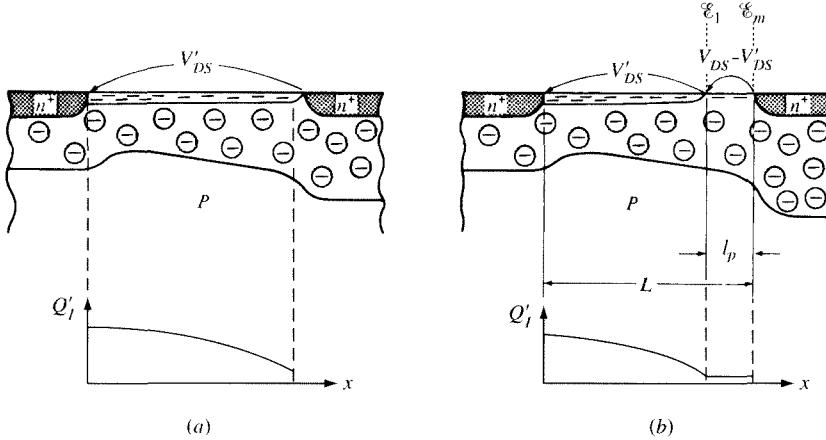
As we discussed in Sec 5.2, the transition from nonsaturation to saturation behavior is smooth rather than abrupt. As saturation is approached, the inversion charge density near the drain falls to a small value, and the velocity of the carriers increases to support the current, as implied by (5.2.6). We will assume for simplicity that saturation velocity is reached for a finite value of the electric field, at $V_{DS} = V'_{DS}$; the inversion layer charge density near the drain then reaches its minimum value, consistent with that equation.^{21,22} The potential difference across the channel will be V'_{DS} ; this quantity can thus be viewed as the potential needed across the channel in order to make the carrier velocity near the drain saturate, and the corresponding inversion charge density there minimum. This situation is pictured in Fig. 5.3a.

If now V_{DS} is increased beyond V'_{DS} , the excess voltage $V_{DS} - V'_{DS}$ must be dropped across a portion of the channel on the right; this portion is shown in Fig. 5.3b, and its length is denoted by l_p . Throughout that portion, the carrier velocity is saturated. This can be seen as follows. In Fig. 5.3a, a voltage V'_{DS} was enough to cause velocity saturation for a transistor with channel length L . From Sec. 5.2, we know that velocity saturation effects become stronger if the channel length is shortened. In Fig. 5.3b, the part of the channel that must support V'_{DS} is shorter than in Fig. 5.3a; thus, the right end of that part must already be in velocity saturation. In fact, velocity saturation will occur slightly to the left of that point; we will ignore this detail, and we will continue to assume that velocity saturation occurs at the point where the channel voltage is V'_{DS} .[†] This approximation is reasonable if the channel is not too short.

If V_{DS} is raised still further, more excess voltage must be dropped across the region where the electrons travel at maximum speed. Because the inversion charge density there is low, the total space charge density in that region will be assumed to be the acceptor density N_A . Thus, although there is inversion layer charge in this region, the total charge in it is practically the same as would be found in a depletion region. From Appendix A, one-dimensional electrostatics show that in a region of constant charge density the potential increases (quadratically) with distance. Therefore if V_{DS} is raised, the above region must expand to support the increased potential drop $V_{DS} - V'_{DS}$. Thus the rest of the channel will correspondingly shrink somewhat. If we think of $L - l_p$ in Fig. 5.3b as the *effective* channel length, we see that this length is “modulated” by the drain-source voltage. This is why the phenomenon we are discussing is referred to as *channel length modulation* (CLM).

Effect on Drain Current Let us now consider the effect of channel length modulation on the drain current. At $V_{DS} = V'_{DS}$, the current I'_{DS} can be estimated from the non-saturation equations, which are assumed valid up to this point. Based on our analysis of long-channel devices (e.g., Sec. 4.7), we expect this current to be inversely proportional to the channel length L . Consider now some value of V_{DS} larger than V'_{DS} and let the corresponding current be I_{DS} . This current can be estimated by considering

[†] This point is classically referred to as the “pinchoff” point, and the region between it and the drain is referred to as the “pinchoff region.” This terminology originates with the assumption that the inversion layer charge can reach zero density next to the drain, commonly made in the context of strong-inversion models (Sec. 4.7). We know that zero inversion layer charge is impossible at that point, because that would require infinite velocity to carry a nonzero current.

**FIGURE 5.3**

Channel (a) at pinchoff; (b) above pinchoff. \mathcal{E}_1 is the magnitude of the horizontal electric field at the end of the inversion layer (i.e., the beginning of the pinchoff region) and \mathcal{E}_m is the maximum electric field magnitude which develops near the n^+ junction.

the part of the channel that is *not* velocity-saturated in Fig. 5.3b. For that part, the situation is the same as in Fig. 5.3a, save for the fact that the role of L is now played by $L - l_p$; thus, because currents are assumed to be inversely proportional to lengths, we will have

$$I_{DS} = I'_{DS} \frac{L}{L - l_p} = \frac{I'_{DS}}{1 - l_p/L}, \quad V_{DS} \geq V'_{DS} \quad (5.3.1)$$

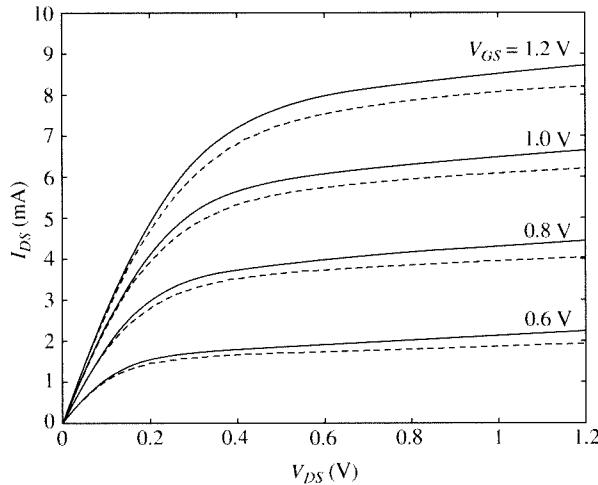
Because the effective channel length decreases with increasing V_{DS} , one expects that the drain current will increase, imparting a positive slope on the I_{DS} - V_{DS} characteristic. This is indeed what is observed experimentally.

If $l_p/L \ll 1$, (5.3.1) can be approximated by

$$I_{DS} = I'_{DS}(1 + l_p/L), \quad V_{DS} \geq V'_{DS} \quad (5.3.2)$$

This equation is sometimes preferred in computer simulation models. [Note that, for some combinations of device parameters and terminal voltages, the denominator of (5.3.1) could become zero, especially because a numerical routine, in an attempt to converge to a solution, may try “unrealistic” values for the terminal voltages; this problem is absent from (5.3.2).] From both of the preceding equations it can be expected that the effect described is more severe for shorter-channel devices.

Figure 5.4 shows simulated I_{DS} - V_{DS} characteristics for a 90-nm long transistor, with the effects of velocity saturation and other effects, notably drain-induced barrier lowering (Sec. 5.5) included, and with CLM both included and excluded. It is apparent that CLM affects the device characteristics, but also, as we have mentioned, that it is

**FIGURE 5.4**

Effect of channel length modulation (CLM) on I_{DS} - V_{DS} characteristics. Solid lines are with CLM included. Dashed lines are without CLM. Other effects that affect the slope in saturation are included in both cases (Secs. 5.2, 5.4, and 5.5). $t_{ox} = 1.5 \text{ nm}$, $N_A = 3 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 0.09 \mu\text{m}$.

not the only factor that contributes to the positive slope of I_{DS} - V_{DS} in saturation. One point to note from this figure is that all of the phenomena described occur together, not separately; therefore, although we model each effect separately, the observed experimental behavior of real devices includes all of the effects working in concert, and balance, together. For example, with velocity saturation coming into play, V'_{DS} depends on channel length [see (5.2.12)], and therefore as the length of the “effective” channel, $L - l_p$, decreases due to CLM, V'_{DS} also decreases. These secondary interactions do not affect the basic qualitative behavior we have described. In a CAD model, the transition from nonsaturation to saturation must be gradual, without kinks in the current or its derivatives; this is achieved by using smoothing functions (Chap. 10), which are seen to have an effect even in the upper part of nonsaturation in Fig. 5.4.

Evaluation of l_p One can use Poisson’s equation (1.2.29) to relate the length of the velocity saturation region, l_p , to the potential across the region, $V_{DS} - V'_{DS}$, assuming that the electric field in the velocity saturation region is longitudinal. If \mathcal{E}_1 denotes the magnitude of the electric field at the left end of this region (Fig. 5.3b), we obtain (Appendix A and Prob. 5.6)

$$l_p = \sqrt{\frac{2\epsilon_s}{qN_A}} \left[\sqrt{\phi_D + (V_{DS} - V'_{DS})} - \sqrt{\phi_D} \right] \quad (5.3.3)$$

where

$$\phi_D = \frac{\epsilon_s \mathcal{E}_1^2}{2qN_A} \quad (5.3.4)$$

This confirms that l_p increases, starting from an initial value of zero, as V_{DS} is raised above V'_{DS} . According to our previous discussion, \mathcal{E}_1 can be taken to be a value of the field above which the velocity is practically saturated; this value is often taken to be the critical field \mathcal{E}_c in Fig. 5.1 or somewhat higher.

This simple picture maintains the one-dimensional field picture we have been assuming so far. However, a detailed numerical analysis of the semiconductor equations shows that the situation near the drain is two-dimensional.^{23,24,29,49–51,54–57,61,62} Such analysis predicts that the gate field “pulls” the inversion layer toward the interface to the left of a certain point near the drain and “pushes” it below the interface to the right of this point. Thus the inversion layer dips below the surface,^{49,57} as indicated in Fig. 5.5. To accurately model CLM, we therefore need to accurately model l_p as a function of drain bias by using a two-dimensional analysis of the pinchoff region. Various approaches have been proposed; the most commonly used one is based on a pseudo-two-dimensional analysis⁵⁷ that has been extended to include the effect of the junction depth d_j .^{61,62} This approach has been found to give satisfactory agreement with experimental data. The derivation of such results is fairly involved and is not given here; details can be found in these original sources or in the summary in Ref. 41. The final result is^{57,62}

$$l_p = l_a \ln \left(\frac{\frac{V_{DS} - V'_{DS}}{l_a} + \mathcal{E}_m}{\mathcal{E}_1} \right) \quad (5.3.5)$$

where l_a is a characteristic length given by

$$l_a = \sqrt{\frac{\epsilon_s}{\epsilon_{ox}} t_{ox} d_j} \approx \sqrt{3 t_{ox} d_j}, \quad (5.3.6)$$

\mathcal{E}_m is the maximum magnitude of the x -directed electric field in the pinchoff region, which occurs at the drain boundary (Fig. 5.3b), given by

$$\mathcal{E}_m = \sqrt{\left(\frac{V_{DS} - V'_{DS}}{l_a} \right)^2 + \mathcal{E}_1^2} \quad (5.3.7)$$

and \mathcal{E}_1 is the magnitude of the field at $x = L - l_p$. From our previous discussion, \mathcal{E}_1 can be expected to be near the critical field for velocity saturation, \mathcal{E}_c .

A depiction of field variation is shown in Fig. 5.5; note that a one-dimensional analysis, as was used to derive (5.3.3), predicts that the magnitude of the field in the x direction, \mathcal{E}_x , would vary linearly with x , which the pseudo-two-dimensional analyses shows is not the case. Nevertheless, it is sometimes assumed that \mathcal{E}_x does vary linearly with x , starting at value \mathcal{E}_1 at the beginning of the pinchoff point. Using this assumption results in^{63,64}

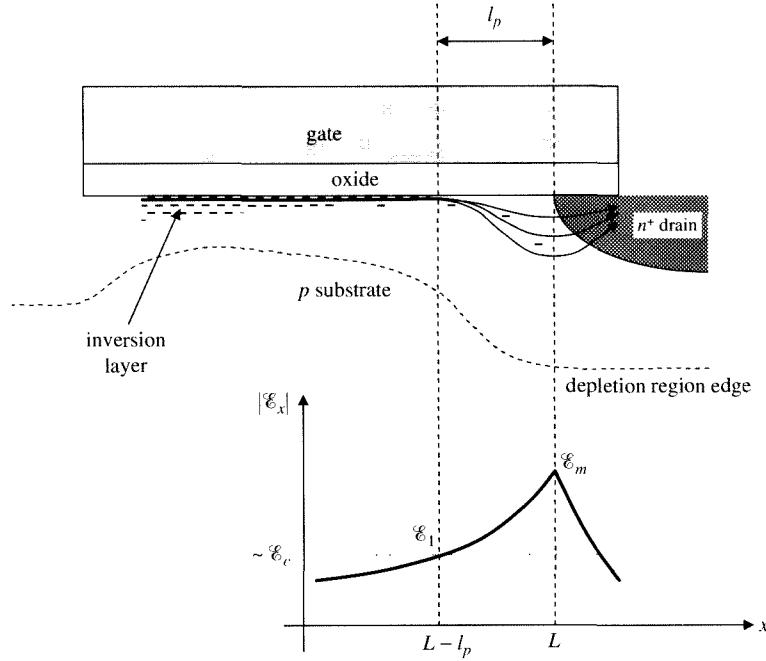


FIGURE 5.5
Channel current flow and longitudinal electric field in the vicinity of the pinchoff region.

$$l_p = l_a \ln \left(1 + \frac{V_{DS} - V'_{DS}}{V_E} \right) \quad (5.3.8)$$

where V_E , although it can in principle be expressed in terms of quantities given so far, is treated as a fitting parameter (typical values are less than 1 V). This approximation has been shown to be reasonable in some situations.^{65,66}

If $V_{DS} - V'_{DS}$ is small compared with V_E , (5.3.8) can be simplified by noting that $\ln(1 + \epsilon) \approx \epsilon$ for small ϵ , and thus $l_p \approx l_a(V_{DS} - V'_{DS})/V_E$. Substituting this result in (5.3.2) gives

$$I_{DS} \approx I'_{DS} \left(1 + \frac{V_{DS} - V'_{DS}}{V_A} \right) \quad (5.3.9)$$

where

$$V_A = V_E \frac{L}{l_a} \quad (5.3.10)$$

This is a relatively crude model, and it predicts that the slope $\partial I_{DS}/\partial V_{DS}$ will be constant in saturation, which is not what is observed experimentally; the assumption of constant

slope is especially inadequate in analog circuit design. Nevertheless for some hand analysis and rough evaluation of device behavior, this model is used because of its simplicity. Note that V_A decreases for decreasing channel length; hence the slope of the I_{DS} - V_{DS} characteristics increases, as observed in practice. (The quantity V_A is loosely called the “Early voltage,” a name used for an analogous quantity in bipolar transistor models; however, the physical mechanisms involved in the two types of devices are different.)

Transition from Nonsaturation to Saturation and the Value of V'_{DS} The core MOS transistor models we developed in Chap. 4 predict a zero slope for $\partial I_{DS} / \partial V_{DS}$ in saturation (whether approached asymptotically, as for the all-region models, or at V'_{DS} for the simplified strong-inversion models). In contrast, (5.3.9) predicts non-zero slope. Thus, *simply joining together the nonsaturation and saturation models at $V_{DS} = V'_{DS}$ causes a slope discontinuity at that point.* This is not only physically incorrect, but also it can cause problems for the numerical algorithms used in circuit simulation programs and can lead to inaccurate predictions of small-signal parameters and distortion. Some older models can exhibit very unsmooth and unphysical behavior around saturation. This problem is sometimes avoided by slightly redefining V'_{DS} as the V_{DS} value that gives a nonsaturation slope equal to that predicted by (5.3.9) (Prob. 5.8), but this is artificial and not appropriate for computer simulation. In addition, as we have seen, velocity saturation effects also modify the saturation behavior of a transistor and therefore need to be taken into account to ensure that a model smoothly transitions from nonsaturation to saturation. As we have mentioned in Sec. 5.2, many models do not use V_{DS} directly, but rather use an “effective” drain-source voltage $V_{DS,\text{eff}}$ that approaches V_{DS} for $V_{DS} < V'_{DS}$ and V'_{DS} for $V_{DS} > V'_{DS}$ in a continuous manner. This naturally smoothes the transition around V'_{DS} . Replacing $V_{DS} - V'_{DS}$ in the preceding expressions developed above by $V_{DS} - V_{DS,\text{eff}}$ avoids the abrupt transition at V'_{DS} . Modern CAD models, using the technique described, have a well-behaved and smooth transition to saturation. Ensuring that a model behaves well in this respect over all geometries, biases, temperatures, and parameter values is not a trivial exercise. The model used to plot Fig. 5.4 uses this approach, and it can be seen that CLM is gradually turned off as one enters nonsaturation. This effect does not cause a serious numerical error.

We conclude this section by noting that channel length modulation is only one among several effects that cause a nonzero slope in saturation. Another effect is “drain-induced barrier lowering,” which can play a dominant role in weak and moderate inversion; this effect is discussed in Sec. 5.5. Another effect is body current, which we have so far neglected; if this current is combined with body parasitic resistance, internal voltage drops develop that can increase the slope of the current in saturation. This effect is considered in Sec. 5.8.

5.4 Charge Sharing

5.4.1 Introduction

When the channel is short and/or narrow, edge effects along its periphery cannot be neglected, and the results of our one-dimensional analysis in Chap. 4 are no longer adequate. Two- and three-dimensional structural details can strongly affect the

overall electric field distribution within the channel, and the corresponding charge distribution in it, especially if the channel dimensions are not much larger than the combined widths of the depletion regions around the source and drain. One view of such effects is termed *charge sharing*, a name suggesting that the gate must share control of the depletion region charge under the channel with other nearby parts of the structure. Proper modeling of charge-sharing effects requires numerical solutions of two- and three-dimensional semiconductor equations. However, such an approach does not lead to simple models that can be used efficiently for circuit simulation. Therefore simplified, albeit less rigorous, approaches are used. A common approach is to continue to use the equations derived for long and wide channels and make simple modifications to them to “save” them in the case where the channel is short and/or narrow. Such modifications will be discussed in this and the following sections and notably involve the threshold voltage; as will be seen, a corresponding modification can then be easily applied to models that do not involve this quantity. The body will be assumed to be uniformly doped until further notice.

To introduce such modifications, let us first review the value of the threshold voltage for devices with long and wide channels. The origin of this quantity can be found in (3.4.11). The threshold is commonly measured with a very small V_{DS} value, as shown in Fig. 4.33. Thus, for simplicity in the following discussion, we assume that $V_{DS} = 0$. Neglecting edge effects, the depletion region can be assumed to be uniform under this assumption. If we multiply both numerator and denominator in the last term of (3.4.11) by the area of the gate, we get

$$V_T = V_{FB} + \phi_0 - \frac{Q_B}{C_{ox}} \quad (5.4.1)$$

where Q_B and C_{ox} are the *total* depletion region charge and the *total* oxide capacitance. Another expression for the threshold is found by using in (3.4.11) Q'_B from (3.4.8b) and evaluating the result at the source end of the channel; this gives

$$V_T = V_{FB} + \phi_0 + \gamma \sqrt{\phi_0 + V_{SB}} \quad (5.4.2)$$

which agrees with (4.7.19).

Effective Threshold If edge effects are not very strong, the general shape of transistor characteristics is approximately preserved. However, one finds experimentally that the I_{DS} - V_{GS} characteristics shift horizontally to lower or higher V_{GS} values; this is observed, for example, in characteristics such as those in Figs. 4.31 or 4.33. This phenomenon is often attributed to charge sharing (to be discussed in detail shortly) and can be taken care of in a rather simple way. Consider, for example, the strong-inversion model (4.7.24). It is seen that V_{GS} appears only in the difference $V_{GS} - V_T$. Thus the previous horizontal shift (always assuming the general shape is otherwise preserved) can be implemented by replacing V_T by an “effective” threshold voltage, \hat{V}_T .^{63,67-123} Much work has been devoted to characterizing and measuring the effective threshold, and circuit designers feel comfortable using this concept. Among other things, the value of the effective threshold (often called just “threshold”) is

decisive in determining whether a transistor can be turned adequately off in digital operation. This focus on threshold does not mean that one totally ignores more general models based on surface potential and inversion charge density, which are not threshold-based. For example, in the surface potential model of (4.3.14) and (4.3.15), one can effect a horizontal shift in I_{DS} - V_{GS} characteristics by changing the flatband voltage V_{FB} .

In this section, we will discuss how the effective threshold can be estimated using charge-sharing concepts. However, we warn the reader that charge-sharing models have been developed with the aim of obtaining simple semiempirical expressions for describing very complex two-dimensional or even three-dimensional phenomena; in such developments, *it has not been possible to justify rigorously all steps*. This will be brought up at several points in the discussion that follows.

5.4.2 Short-Channel Devices

Let us review some of our assumptions for a *long-channel* device, with $V_{DS} = 0$, shown in Fig. 5.6a. Neglecting the edge effects, as we have done so far, is equivalent to assuming that the situation between source and drain is the same as one would have with the source and drain removed, as shown in Fig. 5.6b, but with the channel somehow still communicating with the external world. This assumption provides accurate results as long as the value of L is large. Consider now a short-channel device, as shown in Fig. 5.6c, made with the same process and biased with the same terminal voltages as the long-channel device. Here the edge effects practically extend over all of the channel. Neglecting these effects amounts to viewing the device as shown in Fig. 5.6d, which can hardly be expected to provide credible results. Indeed, assume a very small but nonzero V_{DS} , so that drain current can flow without appreciably disturbing the picture in the channel in Fig. 5.6c. It has been found experimentally that, for a given V_{GS} , I_{DS} is larger than what would have been expected from the picture in Fig. 5.6d. Equivalently, the V_{GS} value required to produce a given I_{DS} value is smaller, which can be viewed as due to a decrease in the effective threshold

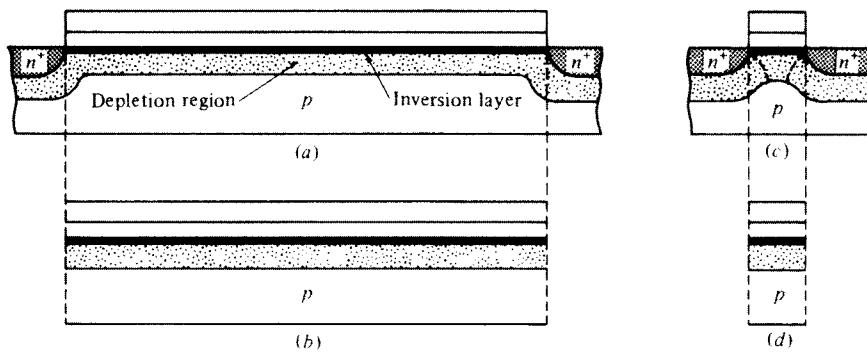


FIGURE 5.6

(a) A long-channel transistor; (b) the channel of (a) with edge effects neglected; (c) a short-channel resistor; (d) the channel of (c) with edge effects neglected.

voltage. An intuitive explanation of the reasons for this effect is now provided. In Fig. 5.6c, the charge in the channel region is influenced by field lines emanating from all nearby structures. Two such structures considered in the analysis of long-channel devices are the gate and the substrate (“back gate”). In short-channel devices the source and drain are so close to all points in the channel that they can affect the latter through their proximity, just as the gate does. In effect, the source and drain now play something of a gate role in addition to their normal function. Field lines emanating from all four structures (gate, bulk, source, and drain) and terminating on charges in the channel must be considered for an accurate description of the device. Essentially, charge control in the channel is shared by all four structures. Bringing the source and drain regions closer to all points in the channel is similar to bringing the gate closer to the channel. Since the source and drain potentials help support the channel charge, their effect on that charge increases as L is decreased, for given V_{GS} and V_{SB} . The corresponding increase in drain current, however, is not predicted by long-channel theory and is modeled instead by using an effective threshold as already mentioned.

This “charge-sharing” viewpoint has provided the basis for most of the analytical/empirical models for short-channel devices.^{63,67-71,75-77,79-83,85,88,92-96,98,101,103-105} We now expand on this idea. Let us assume that the equivalent interface charge Q'_o is zero for simplicity. In Fig. 5.6c and d, it can be imagined that one field line emanates from each positive charge q on the gate. Such a field line terminates either on an electron in the inversion layer or on an ionized acceptor in the depletion region, neglecting the “fringing” field lines terminating on the n^+ regions. In Fig. 5.6d, all of the depletion region charges are “imaged” on gate charges through connecting field lines. In the more realistic picture of Fig. 5.6c, however, this is not the case. Some of the field lines terminating on ionized acceptor atoms near the n^+ regions can be originating from ionized donor atoms in the n^+ regions (inside a thin depletion layer there). Thus, only part of the depletion region charge is imaged on the gate charge in this case. If we can assume that the gate charge in Fig. 5.6c and d is the same, and since some gate charges in c cannot be imaged on depletion charges, more of them are available to be imaged on inversion layer charges, which must, thus, increase in magnitude to accept the extra field lines. The extra I_{DS} observed is attributed to this extra inversion layer charge, although, strictly speaking, one would have to know the spatial distribution of that charge before such a conclusion could be reached. A number of arbitrary assumptions are obviously involved in the preceding arguments (Prob. 5.10). Some of these assumptions are critically considered elsewhere.^{93,100}

Next, a large empirical step is taken: The short-channel device in Fig. 5.6c is supposed to behave as a device with a uniform depletion region, only with a smaller total depletion charge \hat{Q}_{BL} , corresponding to the part of the depletion charge controlled by the gate (the center part between the curved broken lines in Fig. 5.6c). Thus, in lieu of (5.4.1), we will have

$$\hat{V}_{T1} = V_{FB} + \phi_0 - \frac{\hat{Q}_{BL}}{C_{ox}} \quad (5.4.3)$$

where the subscript 1 simply reflects the fact that this is the first charge-sharing phenomenon we are discussing (there will be more below). Comparing the preceding equation to (5.4.1) and (5.4.2), we see that we can write it as follows:

$$\hat{V}_{T1} = V_{FB} + \phi_0 + \hat{\gamma}_1 \sqrt{\phi_0 + V_{SB}} \quad (5.4.4)$$

where $\hat{\gamma}_1$ is an *effective body effect coefficient*, given by

$$\hat{\gamma}_1 = \gamma \frac{\hat{Q}_{B1}}{Q_B} \quad (5.4.5)$$

In this formula, \hat{Q}_{B1}/Q_B is the ratio between the gate-controlled depletion region charge in Fig. 5.6c, to the depletion region charge in Fig. 5.6d, which is less than 1. Thus, instead of the long-channel body effect γ , we have an *effective body effect coefficient* $\hat{\gamma}_1 < \gamma$, and instead of the long-channel threshold V_T , we have an *effective threshold* $\hat{V}_{T1} < V_T$. As might be expected, the value of $\hat{\gamma}_1$ depends on V_{SB} , because that voltage affects the width of the depletion region around the source (recall that in our present discussion we are assuming $V_{DS} = 0$, which implies that the drain-body voltage is equal to V_{SB}). Needless to say, $\hat{\gamma}_1$ also depends on L ; this will be considered further.

There is a large amount of literature on how \hat{V}_{T1} should be calculated, but very little on why it is justifiable to use this value in lieu of V_T in long-channel equations for the drain current (derived by using the gradual-channel approximation) and expect the result to be correct for short-channel devices. The issue is not just what value of threshold voltage should be used in such equations when modeling short-channel devices, but rather why the very *form* of such equations should be valid for short-channel devices in the first place. For example, for a short-channel device, one cannot get an equation of the form of (5.4.1) by multiplying the numerator and denominator in (3.4.11) by the gate area, because the depletion region is not uniform, in contrast to what was assumed for devices with long and wide channels. Even (3.4.11) itself is not valid, because it has been obtained assuming that the field is vertical, which is not the case here. Thus, convincing arguments as to why one can use (5.4.4) and (5.4.5) in long-channel equations are not apparent. However, satisfactory agreement with measurement has led to wide use of the aforementioned approach.

Of the many techniques proposed for determining the quantity \hat{Q}_{B1}/Q_B , let us consider one as an example.⁷¹ The n^+ -region edge will be assumed cylindrical with radius equal to the junction depth d_j , as shown in Fig. 5.7. Next, the depletion regions around each junction and below the inversion layer are drawn as if each existed by itself (i.e., no interaction is assumed), and then they are joined together. The width of all three regions is considered equal by assuming that the junction built-in potential ϕ_{bi} is equal to ϕ_0 ; thus, this width is, from (3.4.7),

$$d_B = \sqrt{\frac{2\epsilon_s}{qN_A}} \sqrt{\phi_0 + V_{SB}} \quad (5.4.6)$$

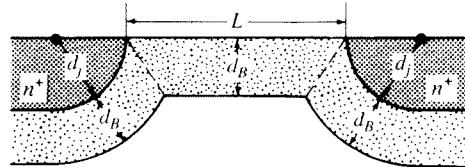


FIGURE 5.7
Trapezoid approximation used for deriving the effective bulk charge.

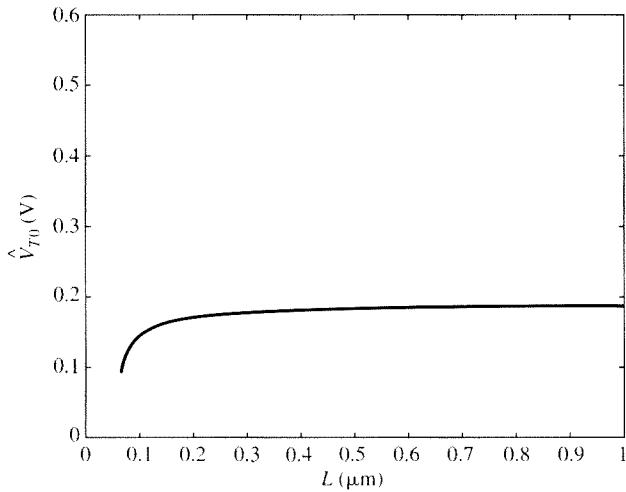
In this picture, L is assumed to be at least large enough so that a region of trapezoidal cross section can be defined as shown in Fig. 5.7. \hat{Q}_{B1} is taken to be equal to the charge in that region, and Q_B is assumed to be the charge corresponding to a rectangle of the same depth and length L . Simple geometry then gives (Prob. 5.11)

$$\frac{\hat{Q}_{B1}}{Q_B} = 1 - \frac{d_j}{L} \left(\sqrt{1 + \frac{2d_B}{d_j}} - 1 \right) \quad (5.4.7)$$

Note that, as expected, this ratio depends on V_{SB} through (5.4.6). Clearly, as V_{SB} is increased, d_B increases and thus \hat{y}_1 , from (5.4.7) and (5.4.5), decreases. This makes intuitive sense: in Fig. 5.7, an increasing d_B implies a decrease in the bottom base of the trapezoid. Thus a change in body voltage will only have a limited effect on the depletion charge, which implies a decrease in the body effect.

The effective threshold voltage, \hat{V}_{T1} , can now be found from (5.4.4). A plot of \hat{V}_{T1} vs. L , assuming $V_{SB} = 0$, is shown in Fig. 5.8. The threshold is seen to decrease as L is decreased. However, in most modern devices a threshold variation in the opposite direction is observed, as is now explained.

Reducing Short-Channel Effects; Halo Implants As may be expected from the previous discussion, if a channel is made very short, the source and drain depletion regions in Fig. 5.6c tend to approach each other, and charge sharing can become severe. In addition, drain-induced barrier lowering, a phenomenon discussed in Sec. 5.5, also worsens. When such phenomena become very strong, long-channel behavior disappears. One then can no longer talk about a threshold shift; the very shape of I - V characteristics changes, in an undesirable way. To avoid this, the depletion region widths d_B must be limited. From (5.4.6), it follows that this can be done by using a higher substrate doping. But it is not desirable to increase substrate doping everywhere, as this is not optimal for other aspects of device operation; for one, this would increase the body effect coefficient. Thus, what is done is to increase the substrate doping locally, around the source and drain, where the depletion regions occur; this is done through a process called ion implantation, discussed in Chap. 9. The result is shown in Fig. 5.9a. Although clear demarcations are indicated for the implant regions (called “halo” regions), in practice the doping concentrations in these regions follow a smooth distribution with distance (Chap. 9). The length of the halo regions is assumed to be fixed, independent of channel length. If now the channel

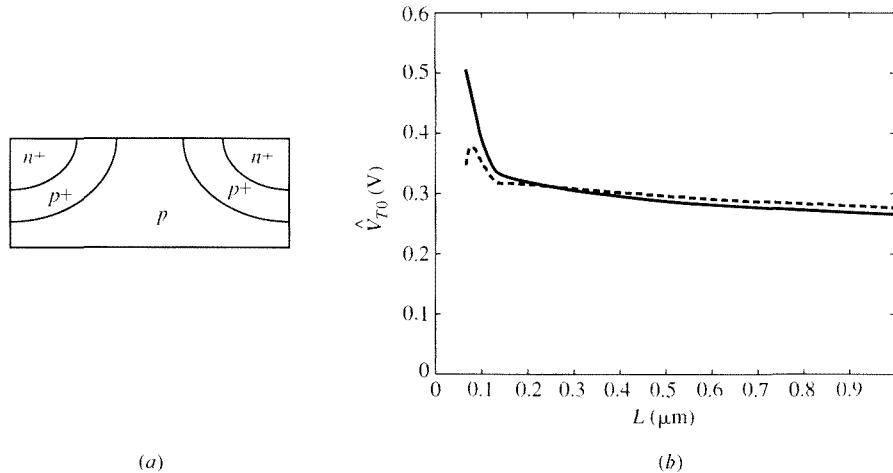
**FIGURE 5.8**

Effective threshold voltage vs. channel length in the presence of charge-sharing effects. $t_{ox} = 1.5$ nm, $N_A = 3 \times 10^{17}$ cm $^{-3}$, $W = 10$ μ m.

length is reduced, the two halo regions approach each other, and the average channel doping increases. As might be expected from (3.4.3), this makes the effective body effect coefficient *larger*; thus, the effective threshold voltage also becomes larger, as shown by the solid line in Fig. 5.9b. If the channel is made shorter than the sum of the two halo region lengths, the doping concentrations in the halo regions get superimposed, resulting in a drastic increase in the effective doping concentration; this gives rise to a rather sudden increase in the magnitude of the slope, as shown toward the left in the solid line in Fig. 5.9b. Finally, if the channel becomes really short (i.e., not much longer than the sum of the source and drain depletion regions in the halo regions), it is possible that charge-sharing effects once again dominate, as shown by the broken line in the figure.

Before the advent of halo implants, devices generally followed the behavior depicted in Fig. 5.8. Nevertheless, behavior as in Fig. 5.9 had also been observed, due to unintended doping nonuniformity effects caused by ion implantation;¹¹⁷ when this was first observed, the effect was called the “reverse short-channel effect.” The behavior in Fig. 5.9b is now the dominant one for threshold voltage variation with length for most modern fabrication processes and will be analyzed in more detail in Chap. 9, where substrate nonuniformity is considered in detail. As already mentioned, in some cases the charge-sharing effect may still be apparent at very small values of L , as shown by the broken part of the upper line; however, such L values are often below the minimum allowed for a given technology. Modeling charge sharing is thus mostly important for higher voltage and older technologies, which do not have halo implants and have thicker oxides, lower body doping, and longer channels.

The preceding analysis has shown qualitatively how charge sharing affects the effective threshold voltage and body effect for short-channel devices and has given approximate techniques for quantifying this effect. Because of the approximations and simplifications involved, these are mostly adequate to convey the principles,

**FIGURE 5.9**

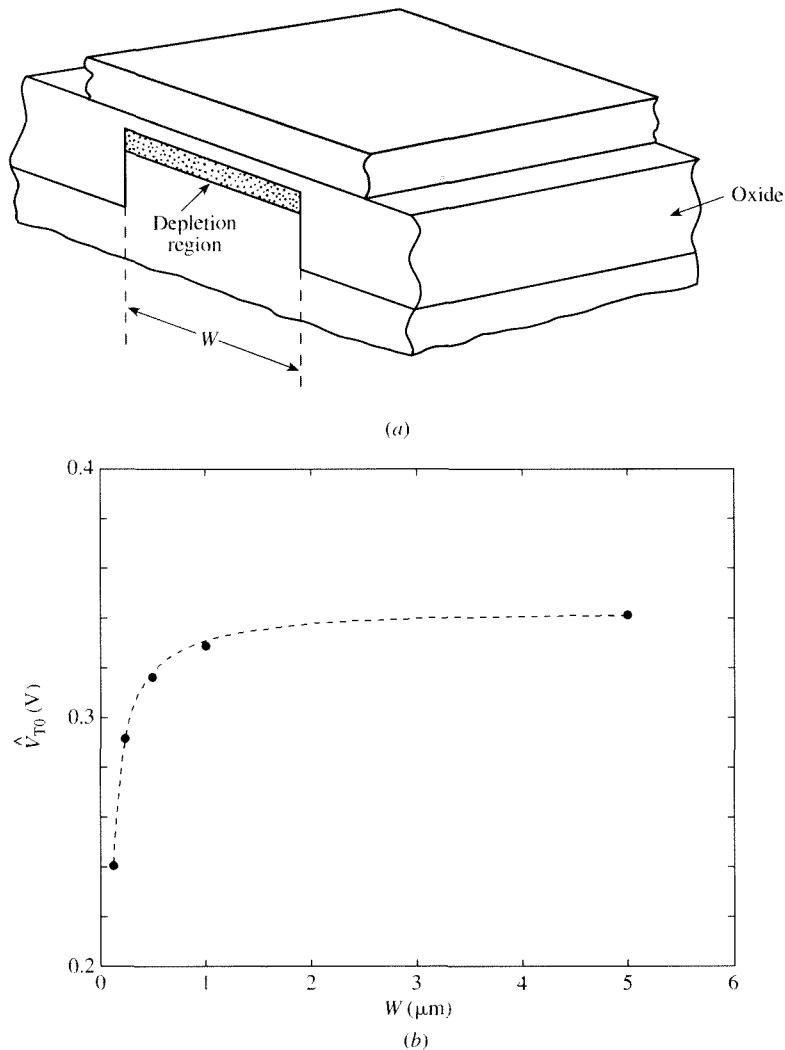
(a) Idealized doping structure of a device with halo implants; (b) the resulting effective threshold voltage vs. channel length if charge sharing is not included (solid line) and included (broken line). $t_{ox} = 1.5$ nm, $N_A = 3 \times 10^{17}$ cm⁻³, $W = 10$ μm.

but cannot be relied upon to provide accurate values. Thus, CAD models usually incorporate empirical corrections to model the effects involved.

5.4.3 Narrow-Channel Devices

Deviations in charge distribution from what is predicted by the gradual channel approximation are also observed along the *width* direction of the channel. The details depend on device construction. We consider two cases.

STI Process A cross section along the channel *width* of a transistor fabricated using STI (shallow-trench isolation) is shown in Fig. 5.10a; a long channel is assumed. The bulk charge is confined to the region directly below the gate, due to the presence of the STI walls. There are the normally expected field lines, emanating from gate charges and terminating on acceptor atoms in the depletion region immediately below the gate. These correspond to the thin oxide capacitance. However, there are additional field lines from the gate, terminating on the sides of the channel; these constitute what is called a *fringing field*. If W is large, these lines constitute a small percentage of the total and can be neglected as has been done in Chap. 4. However, with small W values, the fringing field cannot be ignored. This field helps the vertical field to deplete the region below the thin oxide and make it deeper, thus increasing the surface potential and helping start the inversion layer. Thus, it takes a lower V_{GS} to deplete the charge required before an inversion layer can be formed; this is manifest as a decrease in the effective threshold. The total capacitance of the gate to the substrate has now increased by the fringing capacitance of the sides, from its wide-channel value C_{ox} to a larger value, \hat{C}_{ox2} ; the decrease in threshold is empirically attributed to this capacitance increase and is modeled by assuming that in lieu of (5.4.1), we have^{120–123}

**FIGURE 5.10**

(a) Cross section of a transistor made in a STI process, along the *width* direction. (b) Effective threshold vs. channel width. Dots: measured data; line: model. $t_{\text{ox}} = 3.6 \text{ nm}$, $N_A = 2.5 \times 10^{17} \text{ cm}^{-3}$, $L = 10 \mu\text{m}$.

$$\hat{V}_{T2} = V_{FB} + \phi_0 - \frac{Q_B}{C_{\text{ox}2}} \quad (5.4.8)$$

Comparing this to (5.4.1) and (5.4.2), we see that it can be written as

$$\hat{V}_{T2} = V_{FB} + \phi_0 + \hat{\gamma}_2 \sqrt{\phi_0 + V_{SB}} \quad (5.4.9)$$

with

$$\hat{\gamma}_2 = \gamma \frac{C_{\text{ox}}}{\hat{C}_{\text{ox}2}} \quad (5.4.10)$$

Because $\hat{C}_{\text{ox}2} > C_{\text{ox}}$, we see that $\hat{\gamma}_2 < \gamma$, and thus $\hat{V}_{T2} < V_T$, where V_T is the threshold of a wide-channel device.

To be more quantitative, let the fringing capacitance per unit of *length* (i.e., in a direction perpendicular to the width direction) be C_F'' on each side. This quantity can be estimated,^{120–123} but it is typically determined for a given fabrication process through measurements or parameter extraction. The total fringing capacitance, on both sides, will be $2C_F''L$. Thus,

$$\begin{aligned} \frac{C_{\text{ox}}}{\hat{C}_{\text{ox}2}} &= \frac{C'_{\text{ox}}WL}{C'_{\text{ox}}WL + 2C_F''L} \\ &= \frac{1}{1 + \frac{2C_F''/C'_{\text{ox}}}{W}} \end{aligned} \quad (5.4.11)$$

It is thus seen that for wide devices, $\hat{\gamma}_2$ approaches γ , and thus the effective threshold approaches the value in (5.4.1). This makes sense; as W increases, the capacitance under the oxide increases, but the fringe capacitance does not. The latter becomes a negligible part of the total, and we are back to the ideal, wide-channel case assumed in Chap. 4. Figure 5.10b shows the effective threshold of a narrow-channel device with $V_{SB} = 0$, vs. device width.

LOCOS Process We now consider devices made using LOCOS (local oxidation of silicon). This process is still in use for high-voltage integrated circuits. The situation now is very different from the previous one. This is shown in Fig. 5.11a, which shows a cross section along the width direction of such a device, which shows the characteristic “bird’s beak” shape of the oxide in this process. The depletion region in this case is not limited to just the area below the thin oxide. This is because some of the fringing field lines from the gate charges terminate on ionized acceptor atoms to the *sides* of this region. The gate is now responsible for inducing depletion over a region that is *larger* than what we considered in our analyses in Chap. 4.^{72–74,77,80,88,92,94–99,101,103–105} Therefore to achieve the same level of induced bulk and inversion charge in the “core” transistor, it takes a higher gate bias; this is because some of the field from the gate charge is wasted on inducing bulk charge on the sides, rather than directly under the gate, where the inversion layer forms. It is thus now more difficult to form an inversion layer. This is manifest in an effective threshold voltage that is larger than in the wide-channel case. Note that this is the opposite for the STI case in Fig. 5.10a; there, all fringing field goes to *help* deplete the region right underneath the channel, effectively increasing the surface potential

there and helping inversion; in contrast, in the LOCOS case in Fig. 5.11a, the fringing field is *wasted* outside the channel region.

An increase in source-body bias, while keeping the gate-source voltage fixed, will deplete more charges both under the oxide and in the side regions, which will require additional gate charge to balance them; thus, the body effect will be more pronounced than in the wide-channel case.

Similar to our empirical analysis for short devices, if we define \hat{Q}_{B3} to be the bulk charge in a narrow channel LOCOS device, we will have, in lieu of (5.4.1):

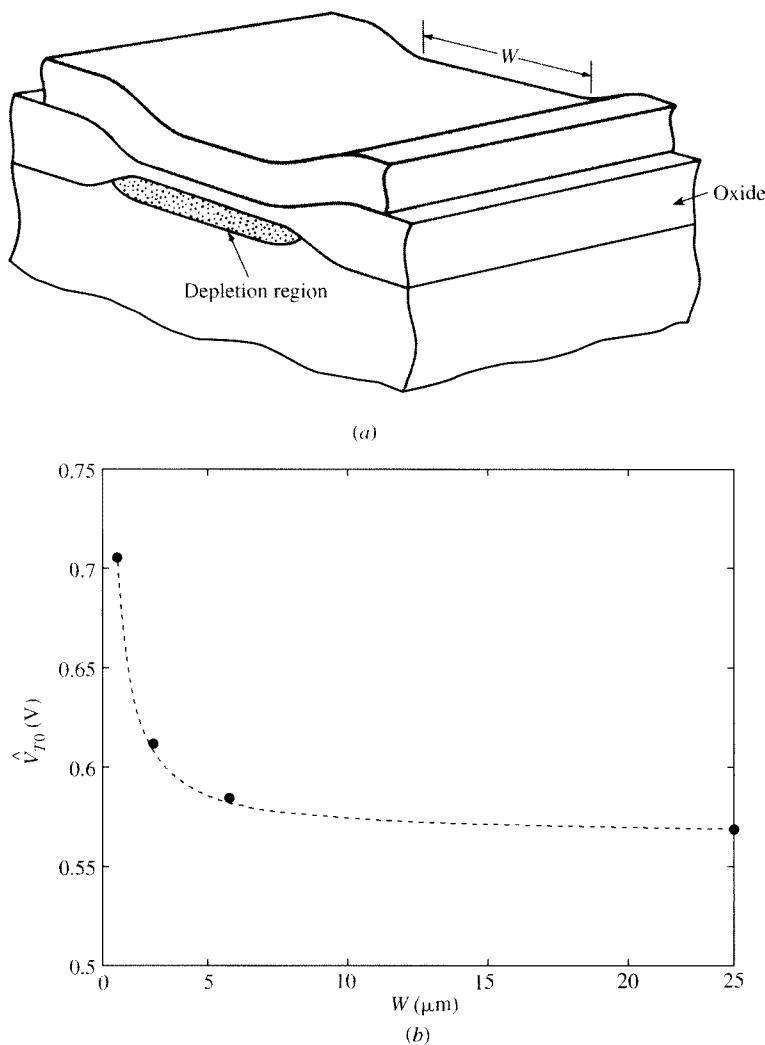


FIGURE 5.11

(a) Cross section of a transistor made in a LOCOS process, along the *width* direction. (b) Effective threshold vs. channel width. Dots: measured data; line: model. $t_{\text{ox}} = 10.8 \text{ nm}$, $N_A = 1.8 \times 10^{17} \text{ cm}^{-3}$, $L = 25 \mu\text{m}$.

$$\hat{V}_{T3} = V_{FB} + \phi_0 - \frac{\hat{Q}_{B3}}{C_{ox}} \quad (5.4.12)$$

By comparing this to (5.4.1) and (5.4.2), we see that it can be written in the form

$$\hat{V}_{T3} = V_{FB} + \phi_0 + \hat{\gamma}_3 \sqrt{\phi_0 + V_{SB}} \quad (5.4.13)$$

where

$$\hat{\gamma}_3 = \gamma \frac{\hat{Q}_{B3}}{Q_B} \quad (5.4.14)$$

Thus, in this case we have $\hat{\gamma}_3 > \gamma_3$ and $\hat{V}_{T3} > V_T$.

Several different approaches to quantitatively estimate the charge-sharing effect for LOCOS processes have been proposed.^{63,77} Here we sketch one such approach. Assume that there is no bird's beak effect and thus the silicon surface is flat; denote the depletion region depth by d_B under the thin oxide region and assume that at the edges of the device the depletion region edge is circular with radius d_B .⁷⁷ Then the charge ratio required in (5.4.14) can be found from geometrical arguments to be (Prob. 5.15)

$$\frac{\hat{Q}_{B3}}{Q_B} = 1 + \frac{\pi}{2} \frac{d_B}{W} \quad (5.4.15)$$

Figure 5.11b shows the influence of width on effective threshold voltage for a LOCOS device, assuming $V_{SB} = 0$.

As a final observation on LOCOS-isolated devices, note that in the "bird's beak" region the oxide thickness changes continuously and not abruptly. As the gate voltage increases, the extent of the region in the width direction that becomes inverted will increase and the conducting channel will become wider. The effective width of the device is therefore no longer simply defined by W , but will vary with bias, increasing with increased gate voltage and decreasing with increased V_{SB} .¹⁰⁵ In fact, for $V_{DS} > 0$ the level of inversion will decrease from the source to the drain, and the "extra" inversion level at the sides will also decrease; therefore the effective width of the channel will depend on position along the channel. This effect is ignored in most analyses and models.

5.4.4 Limitations of Charge-Sharing Models

It should be kept in mind that, although a horizontal shift "saves" the classical equations to a significant extent, it cannot do a perfect job; the reason is that charge-sharing effects do not just shift the I_{DS} - V_{GS} characteristics, but to some extent change their shape as well. Also, even if a pure shift is the only effect assumed, the common calculations of the effective threshold are not rigorous, for reasons we have already pointed out. Thus, the analyses we have presented cannot be expected to accurately

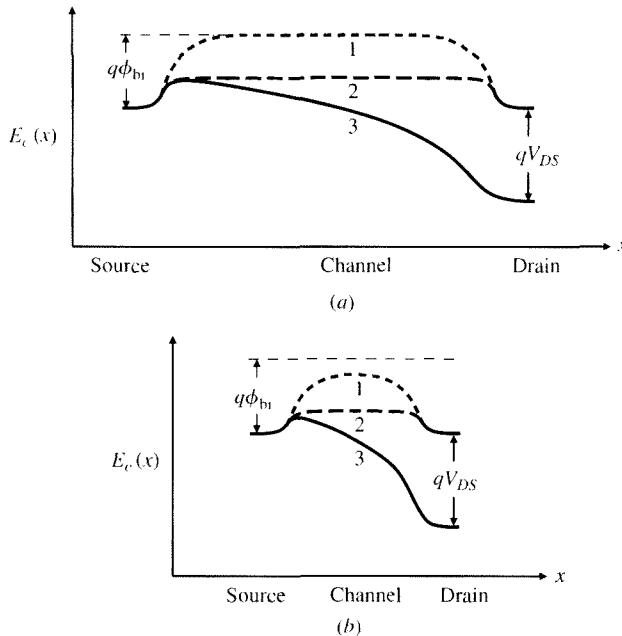
model the influences of two-dimensional details of charge-sharing phenomena. Therefore, to increase the modeling accuracy of the equations produced previously, it is sometimes necessary to empirically modify the values of the parameters in them. Despite these limitations, these analyses do provide an indication of important physical sources of effective threshold voltage shifts for devices with a short or a narrow channel. For devices with channels that are both short *and* narrow, interactions between the edge effects require three-dimensional modeling, so, as far as is known, accurate yet simple modeling is not feasible.

These difficulties are compounded by the fast evolution of CMOS technologies. The fundamental geometrical aspects of devices change rapidly as technology evolves. This has meant that physical modeling of charge-sharing effects has generally lagged the evolution of technologies. With more and more influences to consider, most modern MOS transistor models have at best a tenuous theoretical underpinning for charge-sharing models. It is recognized that the relative severity of the departures from ideal behavior is likely to keep increasing as device dimensions decrease, because the relative portion of a device whose charge is affected by nonidealities increases. Lacking detailed analytical results, most models today include empirical parameters that control the geometry dependence of the threshold voltage and body effect coefficient. As may be expected from the preceding analyses and discussions, such models include terms in $1/L$, $1/W$, and $1/(WL)$, with the latter being important for modeling devices that are both short and narrow. Extraction of the empirical parameters involved is done from experimental data. However, this practice does not provide a physical basis for understanding the cause of the effects or, in the absence of experimental data, of how to estimate what the effects will be for a different device or technology.

5.5 Drain-Induced Barrier Lowering

An important effect encountered in short-channel devices is drain-induced barrier lowering, or DIBL.^{78,84} In order to discuss this effect, we will use the barrier-lowering concept discussed in Secs. 3.2 and 4.1. The reader is encouraged to review that material before proceeding.

Let us first review the situation for a long-channel device,^{207,222} already discussed in Sec. 4.1 and depicted in Fig. 5.12a, which shows how the conduction band edge E_c at the surface varies with position along the channel. We assume a uniform substrate for simplicity, and $V_{SB} = 0$. Curve 1 shows the situation when the gate voltage is equal to the flatband voltage and $V_D = V_S$. The *p*-type substrate away from the source and drain is neutral, and the total potential barrier encountered by electrons attempting to cross into the channel is $q\phi_{bi}$, where ϕ_{bi} is the built-in potential of the source-body and drain-body junctions. This barrier is high, and practically no electrons in the n^+ regions have sufficient energy to cross it. When the gate voltage is increased (still maintaining $V_D = V_S$), the surface potential becomes positive and the corresponding energy in the channel is lowered, as shown by curve 2 in Fig. 5.12a. This lowers the energy barrier, and many electrons enter the channel. Finally, if we keep the gate voltage at the same value as for curve *b*, but increase the drain voltage, the potential rises toward the right by V_{DS} . This corresponds to a lowering of E_c by an amount qV_{DS} , as

**FIGURE 5.12**

Conduction energy band edge at the surface along the longitudinal direction, for (a) a long-channel and (b) a short-channel device with $V_{SB} = 0$. For both parts, curve 1 is for V_{GS} equal to the flatband voltage and $V_{DS} = 0$; curve 2 is for V_{GS} in strong inversion and $V_{DS} = 0$; and curve 3 is for the same V_{GS} as for curve 2, but for a large value of V_{DS} .

shown by curve 3. Source electrons with enough energy to enter the channel are now accelerated by the field toward the drain, and a drain current is observed.

We come now to the case of short-channel devices. The situation is as illustrated in Fig. 5.12b. The three curves are drawn for the same gate voltages as for the corresponding curves in the upper part of the figure. Consider first curve 1. Because the source and drain are both close to the entire channel, their depletion regions approach each other, and the field situation becomes two-dimensional. Field lines now emanate from the source and drain and end up throughout the channel. As we have seen in Sec. 5.4.2, this raises the surface potential there. This corresponds to lowering the energy barrier, as shown by curve 1 in Fig. 5.12b, in comparison with the corresponding curve in Fig. 5.12a. The electrons now find it easier to cross the barrier, and their population in the channel increases. The same is true for curve 2 in Fig. 5.12b, in comparison with curve 2 in Fig. 5.12a, for the case of strong inversion. The barrier is lowered again, and the level of inversion becomes stronger; equivalently, the threshold is lowered, as is also expected from the two-dimensional charge-sharing approach in Sec. 5.4.2.

If now the drain voltage is raised, the drain field, penetrating all the way to the source, decreases the barrier there even further, as can be seen by comparing the peak of curve 3 with that of curve 2 in Fig. 5.12b. This is what is referred to as “drain-induced barrier lowering.” The result is that the device is driven into stronger inversion, i.e., its threshold is lowered further and becomes a function of V_{DS} . A similar conclusion can be reached qualitatively from charge sharing; thus, in Fig. 5.6c, if

the drain voltage is raised, the depletion region around the drain widens, and thus the charge-sharing effect becomes more severe, resulting in a decrease in threshold. However, the quantitative dependence of V_T on V_{DS} cannot be determined from such a simple charge-sharing picture. Two-dimensional numerical calculations,^{95,106} or pseudo-two-dimensional analyses are used for this purpose.^{57,107–114} A representative result from such analyses is that the threshold voltage is lowered by an amount¹¹²

$$\Delta V_{T,\text{DIBL}} \approx -[3(\phi_{\text{bi}} - \phi_0) + V_{DS}]e^{-L/\lambda} \quad (5.5.1)$$

where ϕ_{bi} is the source (or drain) to channel junction built-in potential, given by (1.5.1), and λ is a characteristic length, or “gauge” for the channel length given by

$$\lambda = \sqrt{\frac{\epsilon_s t_{\text{ox}} d_B}{\epsilon_{\text{ox}} \beta_3}} \quad (5.5.2)$$

with d_B being the depth of the depletion region below the channel near the source, given by (5.4.6), and β_3 is a fitting parameter close to 1. Equation (5.5.1) is derived under the assumption that $L \gg d_B$ ($L \geq 4d_B$ is sufficient).

Equations (5.5.1) and (5.5.2) predict that $|\Delta V_{T,\text{DIBL}}|$ decreases as the channel doping increases, because then d_B and therefore λ decrease, in agreement with what we expect from Sec. 5.4.2. Finally, these equations predict an increase in the DIBL effect with increasing V_{SB} , because d_B and therefore λ increase. This agrees with measurements. On the other hand, it is well known that $|\Delta V_{T,\text{DIBL}}|$ decreases with decreasing source/drain junction depth d_j ; yet the preceding equations do not predict a dependence on d_j , although it has been suggested that they may be modified empirically to include it.¹¹² In lieu of the exponential dependence on L predicted by (5.5.1), other analyses predict a dependence proportional to $1/L$,¹²⁴ or to $1/L^3$.⁸⁷ Because of the difficulty in obtaining accurate analytical results, often empirical relations are used, based on experimentally observed behavior over length and bias; for example, one such relation is¹²⁵

$$\Delta V_{T,\text{DIBL}} = -\frac{(\beta_4 + \beta_5 V_{SB})}{L^{\beta_6}} V_{DS} \quad (5.5.3)$$

where β_4 , β_5 , and β_6 are empirical constants chosen to match experimental results.

For general models, which do not have threshold voltage as a parameter, the observed threshold shift due to DIBL can be applied instead to the flatband voltage.¹²⁶ With considerable computational complexity, a more general inclusion of DIBL and other short-channel effects in general models is possible.^{63,92,98,114,127}

The onsets of the moderate- and strong-inversion regions are strongly tied to the effective threshold and follow qualitatively a similar dependence on V_{DS} (as well as on V_{SB} and L). For example, let us briefly look at the current in weak inversion. Assuming that (4.8.11) holds, but with V_M shifted by $\Delta V_{T,\text{DIBL}}$, we see that

$$I_{DS,\text{short channel}} = I_{DS,\text{long channel}} e^{-\Delta V_T/(n\phi_i)} \quad (5.5.4)$$

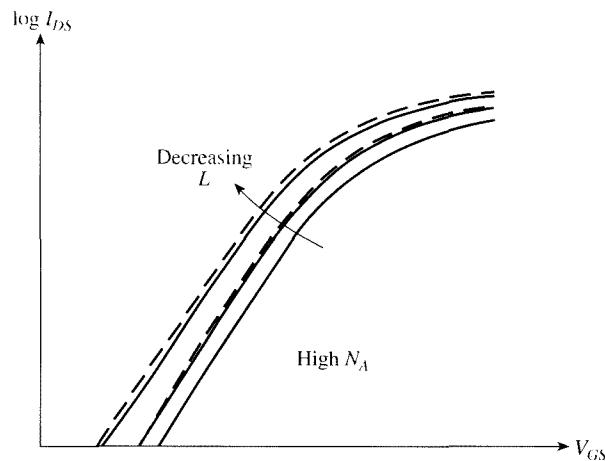


FIGURE 5.13

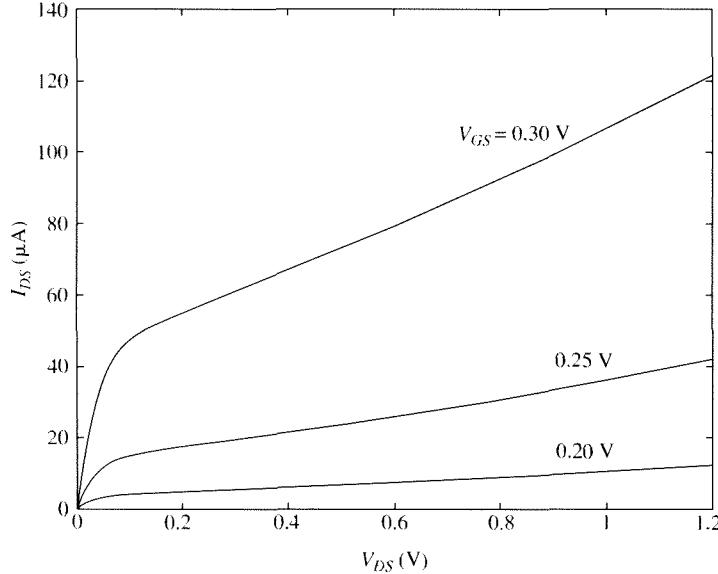
$\log I_{DS}$ vs. V_{GS} for three different channel lengths, for well-behaved devices (modified from Ref. 104). Solid lines are for one value of V_{DS} ; broken lines are for another, larger V_{DS} value.

The effect of the threshold shift is equivalent to a shift of the I_{DS} - V_{DS} characteristics, as illustrated in Fig. 5.13;¹⁰⁴ the larger V_{DS} value lowers the threshold and thus shifts the curves to lower V_{GS} values. A more complicated picture emerges when the current is plotted vs. V_{DS} . It is important to note that the DIBL effect continues even beyond pinchoff. Thus, even if channel length modulation is neglected, I_{DS} in the “saturation” region will not saturate, because the drain voltage will still electrostatically affect the electric field and therefore also the charge distribution and potentials in the transistor. The current will thus continue to increase as V_{DS} is increased,^{27,79,82,84} because the effective threshold will keep decreasing.[†] The effect on the I_{DS} - V_{DS} characteristics is qualitatively similar to that of channel length modulation.[‡] DIBL is a main cause for the large slope of saturation I_{DS} vs. V_{DS} in weak inversion; an example is shown in Fig. 5.14.

Up to now we have considered electrostatic effects that become important as the channel length decreases, but that are small enough so that they do not compromise the device behavior very much. Perhaps the most important criterion of such “electrostatic integrity” at a given small L is whether the weak-inversion slope of $\log I_{DS}$ vs. V_{GS} stays nearly constant with V_{DS} , remaining almost equal to that of a long-channel device built with the same process technology. The qualifiers “nearly” and “almost” are used here on purpose because it is common that in any given MOSFET generation, the device with the minimum acceptable channel length at the highest

[†] Note, also, that an incorrectly designed device that has been turned off by reducing V_{GS} sufficiently below V_t may turn on just by increasing V_{DS} if V_t is lowered enough by the above mentioned effect.

[‡] However, the effect of DIBL is significantly increased with V_{SB} , whereas that of channel-length modulation is not.

**FIGURE 5.14**

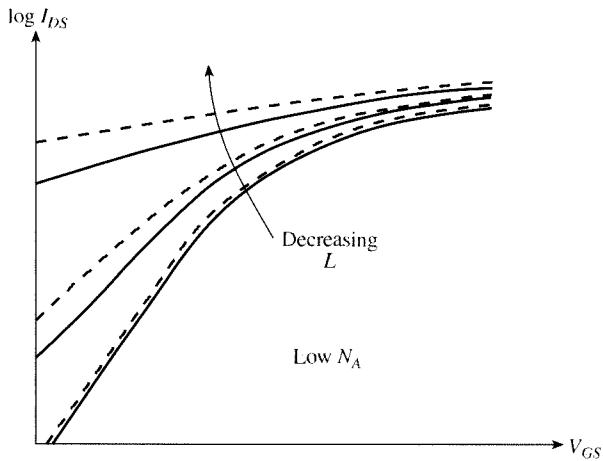
Effect of DIBL on the current in weak inversion. $t_{ox} = 1.5$ nm, $N_A = 3 \times 10^{17}$ cm $^{-3}$, $W = 10$ μm, $L = 0.09$ μm, $V_{SB} = 0$ V.

expected V_{DS} will generally show a subthreshold slope decrease of a few percent relative to the long-channel case. Subthreshold slope is usually indicated by the S factor, defined in Sec. 4.8 as $S = dV_{GS}/d(\log I_{DS})$. This is numerically equal to the swing of V_{GS} needed to change I_{DS} by one decade. This quantity is used in the design of digital circuits and is a sensitive indicator of the presence or absence of short-channel effects. Typical S factors in modern processes are around 80 mV/decade, so the maximum acceptable change, due to short-channel and high V_{DS} effects, would be about 5 mV/decade; e.g., if the long-channel device exhibits $S = 80$ mV/decade, then the shortest-channel devices should have $S \leq 85$ mV/decade. A good example of device characteristics with good electrostatic integrity has been shown in Fig. 5.13.

The small shift in the characteristics due to DIBL is typically characterized by $\Delta V_{GS}/\Delta V_{DS}$, in mV/V, for a constant I_{DS} in weak inversion. Typical values of DIBL for devices with good electrostatic integrity are less than 100 mV/V.

5.6 Punchthrough

The concept of a simple shift of the I_{DS} - V_{GS} (and $\log I_{DS}$ - V_{GS}) characteristics by an amount $\Delta V_T(L, V_{DS}, V_{SB})$, covered in the preceding two sections, is a very convenient way to capture short-channel effects to first order. However, as the channel becomes very short, the above simple picture fails. Most noticeable is the failure of the $\log I_{DS}$ vs. V_{GS} characteristics to maintain their slope if L is too small and V_{DS} is increased. In such cases, the slope of $\log I_{DS}$ becomes very small, and the device cannot be turned

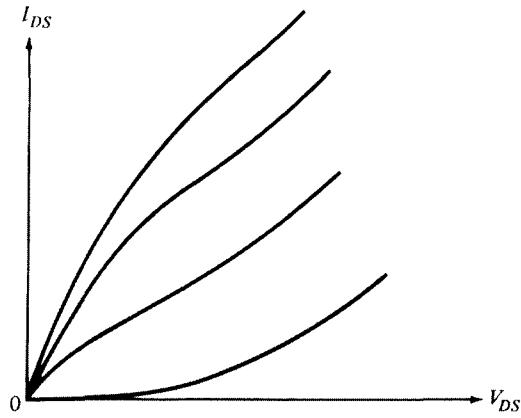
**FIGURE 5.15**

Log I_{DS} vs. V_{GS} for three different channel lengths, revealing poor electrostatic integrity (modified from Ref. 104). Solid lines are for one value of V_{DS} ; broken lines are for another, larger V_{DS} value.

off adequately even if V_{DS} is decreased significantly. An example¹⁰⁴ is shown in Fig. 5.15. The device characteristics for the shorter channels fail the criteria of good electrostatic integrity because they exhibit strong V_{DS} dependence of their S factor (see Sec. 5.5). As already mentioned in Sec. 5.4, these devices suffer from the “punchthrough” effect.^{23,28,60,77,84–87,104,128–132} Punchthrough is a severe case of barrier lowering causing electron flow from source to drain either along the surface, in which case it is referred to as *surface punchthrough*, or in the bulk, referred to as *bulk punchthrough*. An example of what the I_{DS} - V_{GS} characteristics of a punchthrough-suffering device look like is shown in Fig. 5.16. These effects are easily modeled by two-dimensional device simulation but are often not included in compact models, as they are not encountered in practical devices under normal operation.

In a rather simplified way, surface punchthrough is said to occur when the depletion regions of the source and drain, in the absence of any depletion from the gate, reach each other. Figure 5.17 shows two such cases: (a) surface punchthrough for the case of a uniformly doped substrate, as was the case for the device in Fig. 5.6c, and (b) bulk punchthrough for the case of ion-implanted channel with higher concentration at the surface, which limits the spread of the depletion regions there. It is worth pointing out that the signature of bulk punchthrough is primarily a “bottoming out” of I_{DS} with decreasing V_{GS} , at a value which fails to decrease when V_{GS} is decreased.

Although the notion of touching depletion regions is simplistic (for example, it ignores the fact that after all a depletion region always exists along the channel due to the gate), it does convey the idea that with a diminished neutral region between source and drain, field lines from the latter can reach the vicinity of the former and modulate the surface potential and hence Q'_L . Clearly, as pointed out in Sec. 5.4.2 the remedy for punchthrough is to shrink the depletion regions by adding doping, usually through

**FIGURE 5.16**

Typical characteristics of a device with punchthrough problems.

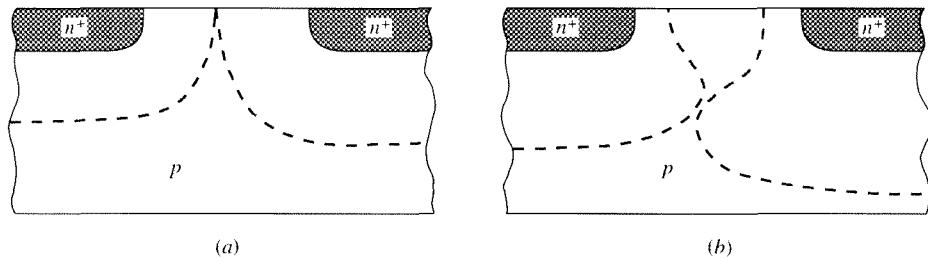
**FIGURE 5.17**

Illustration of source and drain depletion edges at the onset of punchthrough. (a) Surface punchthrough; (b) bulk punchthrough.

ion implantation. For example, in the case depicted in Fig. 5.17b, a “punchthrough-control” implant can be introduced, in addition to the surface V_T -control implant.

5.7 Combining Several Small-Dimension Effects into One Model—A Strong-Inversion Example

In the previous sections, we have discussed many different small-dimension effects that take place in the MOS transistor and have considered ways to model them. For the purposes of presentation, our approach was to focus on one effect at a time. A complete device model, though, must incorporate several such effects simultaneously. Doing so properly would require considering the interaction between these effects, but then, unfortunately, the expressions obtained become very complicated. An empirical approach often taken, if each effect by itself is small, is to assume that the effects are noninteracting. For example, consider a device with a channel that is

both short *and* narrow. Here three-dimensional analysis would be needed to obtain accurate results. However, for simple estimates such a device is sometimes modeled by using an effective-threshold formulation, with the effective threshold \hat{V}_T being given by $V_T + \Delta V_{TL} + \Delta V_{TW}$, where V_T is the long-channel threshold, ΔV_{TL} is the change assuming only short-channel effects are present, and ΔV_{TW} is the change assuming only narrow-channel effects are present. For small ΔV_{TL} and ΔV_{TW} , this approach can be justified to some extent (Prob. 5.19).

In developing certain models, one is faced not only with combining several effects that may be active in a given region of operation, but also with combining different expressions that have each been developed for particular regions only (e.g., combining nonsaturation and saturation expressions). In such cases, one must make sure that, as a boundary between two regions is crossed, continuity is maintained not only for I_{DS} but also for its derivatives with respect to the terminal voltages (and also for charges and their derivatives, to be discussed in Chaps. 6–8).

We now illustrate the preceding considerations with a simple example.

Example 5.3. We will sketch the development of strong-inversion drain current expressions that include the following effects:

1. Effect of L on effective threshold
2. DIBL
3. Effect of W on effective threshold
4. Velocity saturation effects
5. Effective mobility dependence on normal field
6. Channel-length modulation in saturation

The approach used will be a variation of one used elsewhere.²⁶ Effects 1, 2, and 3 will be modeled by using the following effective threshold voltage:

$$\hat{V}_T(L, W, V_{DS}, V_{SB}) = V_T(V_{SB}) + \Delta V_{TL}(L, V_{DS}, V_{SB}) + \Delta V_{TW}(W, V_{SB}) \quad (5.7.1)$$

where $V_T(V_{SB})$ can be calculated from (4.7.19), and ΔV_{TL} and ΔV_{TW} can be calculated as described in Secs. 5.4. and 5.5.

Effect 4 will be modeled as in Example 5.1 (Sec. 5.2), and effect 5 will be included by using (4.11.31). Thus, in the nonsaturation region, we have

$$I_{DS} = \frac{\mu_0 C'_\text{ox} \frac{W}{L} \left\{ [V_{GS} - \hat{V}_T(L, W, V_{DS}, V_{SB})] V_{DS} - 0.5 \alpha V_{DS}^2 \right\}}{\left\{ 1 + \theta [V_{GS} - \hat{V}_T(L, W, V_{DS}, V_{SB})] + \theta V_{SB} \right\} [1 + V_{DS}/(L \mathcal{E}_c)]}, \quad V_{DS} \leq V'_{DS} \quad (5.7.2)$$

where \hat{V}_T is given by (5.7.1).

The value of the saturation drain-source voltage V'_{DS} will be found as shown in Example 5.2 by setting dI_{DS}/dV_{DS} equal to zero from (5.7.2) and neglecting the dependence of V_T on V_{DS} for this calculation. This results in (5.2.12) again. A more exact calculation results in a much more complicated expression that gives practically the same value as (5.2.12).²⁶

We will now determine I_{DS} in saturation. This requires some caution. If \hat{V}_T were independent of V_{DS} , as implicitly assumed in Sec. 5.3, we could use (5.3.1), where I'_{DS} would be given from (5.7.2) after replacing V_{DS} by V'_{DS} . In the resulting expression for I'_{DS} , V_{DS} itself would not appear. This would be so because the channel end, being considered to be pinched off, would be assumed to be at potential V'_{DS} with respect to the source, no matter what the actual value of V_{DS} . However, here we want to include the effect of V_{DS} on \hat{V}_T due to DIBL as explicitly indicated in (5.7.1). As already pointed out, this effect is assumed to be unrelated to pinchoff and is present whether V_{DS} is smaller or larger than V'_{DS} , as explained in Sec. 5.5. Hence, even in saturation, \hat{V}_T will continue to be a function of V_{DS} , not V'_{DS} . Accordingly, we have

$$I_{DS} = \frac{\mu_0 C'_\text{ox} \frac{W}{L} \left\{ [V_{GS} - \hat{V}_T(L, W, V_{DS}, V_{SB})] V'_{DS} - 0.5\alpha V'^2_{DS} \right\}}{\left\{ 1 + \theta [V_{GS} - \hat{V}_T(L, W, V_{DS}, V_{SB})] + \theta_B V_{SB} \right\} \left[1 - \frac{l_p}{L} + V'_{DS} / (L \mathcal{E}_c) \right]}, \quad V_{DS} > V'_{DS} \quad (5.7.3)$$

with l_p as in Sec. 5.3. Note that, with the exception of the argument of \hat{V}_T , V_{DS} has been replaced by V'_{DS} in the preceding equation.

We now need to make sure that the nonsaturation and saturation expressions give not only the same value of I_{DS} at $V_{DS} = V'_{DS}$, but also the same slope dI_{DS}/dV_{DS} . This can be achieved by a slight modification of the value of V'_{DS} , as explained in Sec. 5.3. However, there are still two problems. First, although the redefinition of V'_{DS} can guarantee continuity in dI_{DS}/dV_{DS} , it does not guarantee continuity in d^2I_{DS}/dV_{DS}^2 . This has repercussions on small-signal modeling and distortion prediction, as discussed in Chap. 10. Second, the use of (5.7.2) or (5.7.3) depending on whether or not $V_{DS} \leq V'_{DS}$, requires an "IF" statement in the computer code implementing the model. This is undesirable (Chap. 10).

The above problems would not exist if a better approach had been followed, using the all-region models of Sec. 4.3. These make possible a single expression for I_{DS} , which saturates smoothly, with its slope with respect to V_{DS} also varying smoothly, as V_{DS} is increased. If such a model cannot be used, an approach sometimes taken is the use of the "effective" drain-source voltage introduced in Sec. 5.3. According to this approach, we can replace (5.7.2) and (5.7.3) by the following single-piece expression:

$$I_{DS} = \frac{\mu_0 C'_\text{ox} \frac{W}{L} \left\{ [V_{GS} - \hat{V}_T(L, W, V_{DS}, V_{SB})] V_{DS,\text{eff}} - 0.5\alpha V_{DS,\text{eff}}^2 \right\}}{\left\{ 1 + \theta [V_{GS} - \hat{V}_T(L, W, V_{DS}, V_{SB})] + \theta_B V_{SB} \right\} \left[1 - \frac{l_p}{L} + V_{DS,\text{eff}} / (L \mathcal{E}_c) \right]} \quad (5.7.4)$$

where $V_{DS,\text{eff}}$ is a quantity that approaches V_{DS} if $V_{DS} < V'_{DS}$, and V'_{DS} if $V_{DS} > V'_{DS}$. Of course, the transition from $V_{DS,\text{eff}} = V_{DS}$ to $V_{DS,\text{eff}} = V'_{DS}$ should be appropriate for providing realistic values for the current and its slope. This practice is discussed in Chap. 10 where much more can be found on compact modeling.

5.8 Hot Carrier Effects; Impact Ionization

This section is expanded from a corresponding section in the 2nd edition of this book, written by D. A. Antoniadis.

Consider a transistor in strong-inversion saturation. We have seen in the previous sections that the longitudinal electric field in the channel increases from source to drain. For an abrupt body-drain junction, this field is maximum at the junction boundary (Fig. 5.5) and is higher for shorter-channel devices and for higher drain-source voltages. The carriers move faster as they approach the drain in such a way that the product of drift velocity and inversion layer charge density is kept constant (5.2.6). Eventually, velocity saturation occurs. For long-channel devices, this happens in the so-called pinchoff region; for short-channel devices, it can occur over much of the channel. When carriers move at the saturation velocity, they continue to acquire kinetic energy from the field but their velocity is randomized by excessive collisions such that their average velocity along the field direction no longer increases but their *random* kinetic energy does. Depending on the statistics of scattering, a small fraction of the overall carrier population acquires a significant amount of energy, and these are called *hot carriers*. Clearly, the higher the field, the higher the proportion of hot carriers in the overall population. Generally, in MOSFETs, these high fields are encountered in saturation in the pinchoff region.

Figure 5.18 illustrates the *hot carrier effects*. “Cool” electrons flow into the pinchoff region and are *heated* by the field. Some of them acquire enough energy to create *impact ionization* of silicon lattice atoms, whereby new electrons and holes are created; this effect is also referred to as *weak avalanche*. The new electrons join the stream of channel electrons and move on toward the drain; the new holes are swept by the normal field into the substrate, where they give rise to *drain-to-substrate current*, I_{DB} .^{64,133–167} A much smaller fraction of electrons acquire even higher energy, which allows them to overcome the silicon-oxide barrier, get injected into the oxide, and be collected by the gate as *gate current*. Such energetic carriers also create damage at the silicon-oxide interface that manifests itself as an increase in interface states density (Sec. 2.2), and yet another small fraction become trapped in the oxide giving rise to a localized change of Q'_o .^{145,146,148,150–153} (see also Sec. 2.2). The ensuing corruption of the oxide results in device degradation with operating time, or “aging.”

Note that the electrons moving in the channel have a statistical distribution of speeds and do not all just move at the average velocity; the higher the average electron speed, the more statistically likely it is that some electrons will have sufficient energy to generate an additional electron-hole pair. The ionization process therefore does not abruptly start at some specific current level, but rather increases smoothly as the drain voltage is increased.

Drain-Body Current Let us consider the drain-to-substrate current I_{DB} . This current is proportional to the number of electrons susceptible to ionization per unit time, which in turn is proportional to I_{DS} . Also, according to the previous discussion, I_{DB} is an increasing function of the maximum field at the drain; this field is, in turn, a function of the “excess” drain voltage, $V_{DS} - V'_{DS}$.

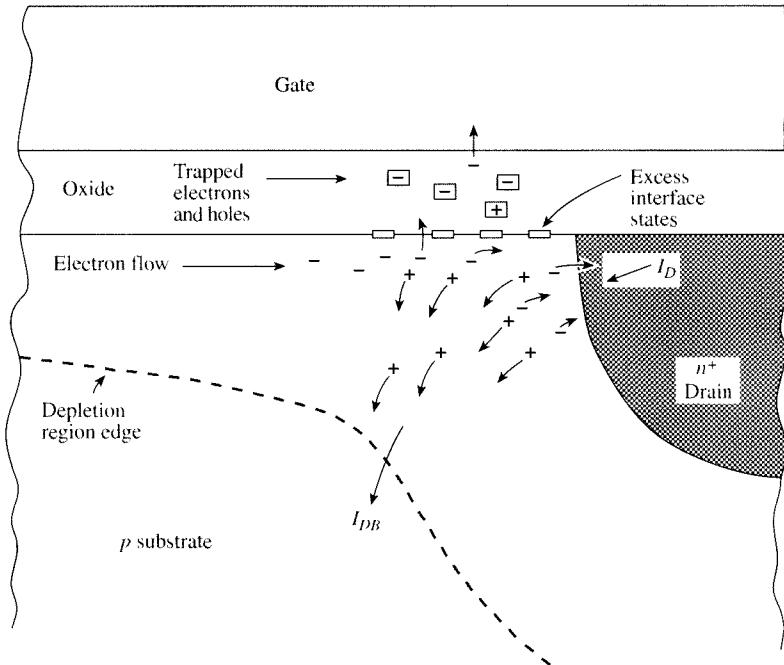


FIGURE 5.18
Schematic representation of hot carrier effects in a region of high longitudinal electric field in the channel of an nMOSFET.

Plots of I_{DB} vs. V_{GS} , for various values of V_{DS} , behave as shown in Fig. 5.19. For a given V_{DS} , when V_{GS} is increased starting from low values, I_{DS} increases, and thus I_{DB} increases too, according to the previous discussion. Further increases of V_{GS} increase V'_{DS} significantly, causing a strong decrease in $V_{DS} - V'_{DS}$ and thus in the maximum field at the drain. This causes I_{DB} to decrease at large V_{GS} values, as seen toward the right in the figure.

Developing an analytical expression for I_{DB} is a lengthy process, the main steps of which will now be sketched. We will assume for simplicity that carriers carrying the channel current travel approximately horizontally, even within the pinchoff region. For low-level impact ionization, we can assume that all impact ionization events are primary, i.e., are produced by the electrons carrying the drain current. This means that the electron-hole pairs generated by these events do not, themselves, cause further (secondary) impact ionization. Consider now a small element of length Δx at position x in the pinchoff region. In it, each electron-hole pair results from the impact of one of the electrons carrying the drain-source current I_{DS} . These pairs contribute to the bulk current, I_{DB} , an amount ΔI_{DB} , which will be proportional to I_{DS} and to the length of the region Δx :

$$\Delta I_{DB}(x) = \alpha_n(x) I_{DS} \Delta x \quad (5.8.1)$$

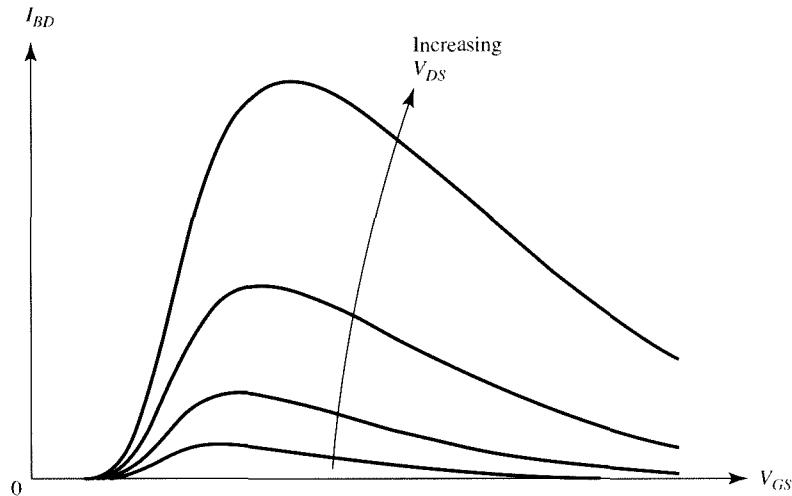


FIGURE 5.19
Substrate current resulting from impact ionization.

where the constant of proportionality, $\alpha_n(x)$, is called the ionization coefficient; the units of this coefficient are cm^{-1} . The combined effect of all such contributions can be obtained by letting the finite differences in become differentials and integrating over the length of the pinchoff region, where impact ionization is assumed to take place. Thus,^{104,168}

$$I_{DB} = I_{DS} \int_{L-l_p}^L \alpha_n(x) dx \quad (5.8.2)$$

where l_p is the length of the pinchoff region; to include the possibility that impact ionization also takes place somewhat outside this region, l_p can be replaced by a somewhat different “effective” length. The ionization coefficient can be estimated from the probability of occurrence of an impact ionization event and, as might be expected, turns out to be a strong function of the longitudinal field magnitude, $\mathcal{E}_x(x)$:¹⁶⁹

$$\alpha_n(x) = A_n e^{-B_n/\mathcal{E}_x(x)} \quad (5.8.3)$$

where A_n and B_n are of the order of 10^6 cm^{-1} and 10^6 V/cm , respectively, but are commonly left to be chosen empirically as fitting parameters.[†] Finally, the field $\mathcal{E}_x(x)$ is estimated using a pseudo-two-dimensional approximation and is expressed as a

[†] In reality, the carriers take a finite distance to absorb enough energy from the electric field to be able to cause ionization, so impact ionization is a nonlocal effect and the point of peak ionization rate can be downstream of the point of peak electric field; however, the local ionization model discussed here, with appropriate selection of parameters, fits experimental data with reasonable accuracy.

function of the potential across the pinchoff region, $(V_{DS} - V'_{DS})$ (Sec. 5.3). Using the result in (5.8.2) and (5.8.3) and following further approximations, it can be shown that the resulting I_{DB} is of the form^{63,150,167}

$$I_{DB} = I_{DS} K_i (V_{DS} - V'_{DS}) \exp\left(-\frac{V_i}{V_{DS} - V'_{DS}}\right) \quad (5.8.4)$$

where K_i and V_i are parameters that must be fitted empirically to data.[†] Typical values for K_i are 1 to 3 V⁻¹, and for V_i , 10 to 30 V. The *total* drain current, I_D , is given by

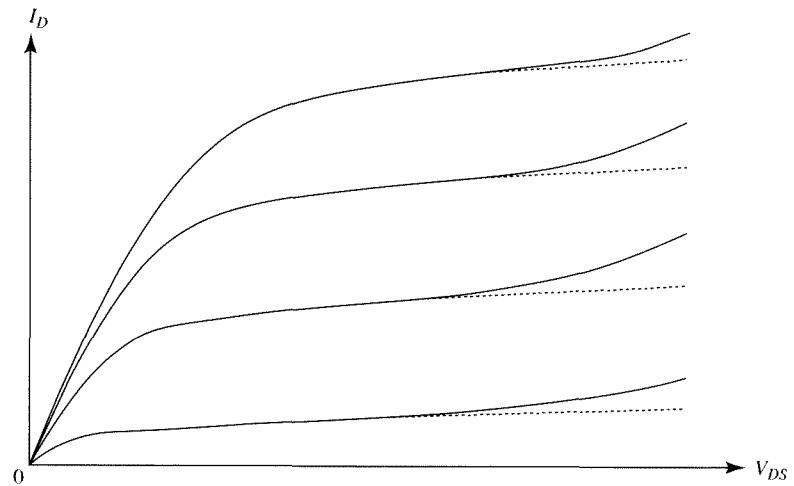
$$I_D = I_{DS} + I_{DB} \quad (5.8.5)$$

Channel Breakdown The discussion so far about substrate current has been based on the assumption that the latter is small compared with drain current, for example, more than 2 to 3 orders of magnitude smaller. However, as V_{DS} increases to well above V'_{DS} , depending on device design, a point is reached at which the substrate current becomes comparable to the drain current. At that point, the *total* drain current I_D is seen to increase rapidly with V_{DS} . This condition marks the onset of considerable avalanche multiplication, in which the newly created carriers can participate in significant numbers in generating more pairs as they gather energy from the field; the device is then said to be operating in the *channel breakdown* region. Figure 5.20a shows I_D - V_{DS} characteristics of a device that is affected by impact ionization; the onset of breakdown can be clearly seen.

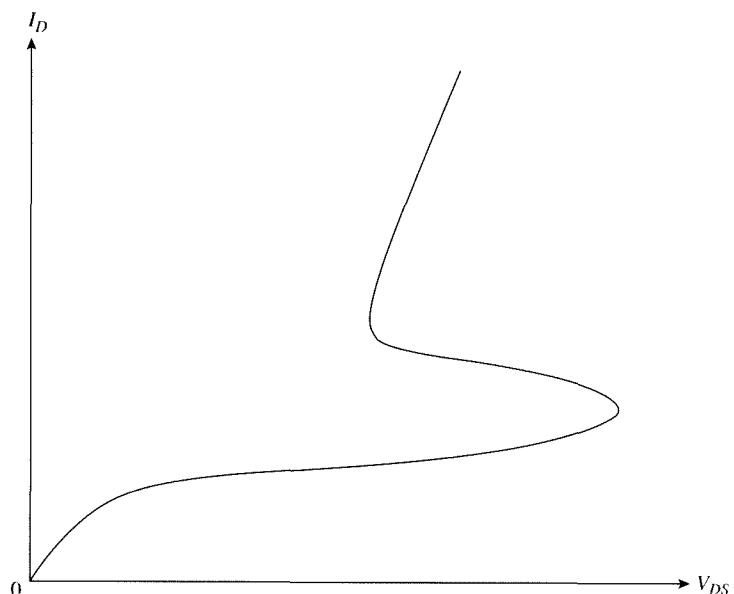
A significant hole current in the substrate (Fig. 5.18) can cause enough ohmic voltage drop across the resistance of the substrate material to forward-bias the substrate-source *pn* junction. For sufficiently high forward bias, electrons can then be injected by the source into the substrate, and a parasitic bipolar transistor action can be initiated, contributing to breakdown. The effect of this on the I_D - V_{DS} characteristics is called “snapback” and can be seen if one sweeps I_D and measures V_{DS} . This results in the type of I_D - V_{DS} characteristic shown in Fig. 5.20b. Snapback is not a pure MOS effect, but is rather due to the interaction between the MOS transistor and the parasitic bipolar transistor consisting of the drain, body, and source regions. It can be modeled by adding a bipolar device, along with resistors, to the MOS transistor.¹⁷⁶ The snapback region is not a region where transistors are normally operated, but it can be important in circuits used at the inputs/output of chips to provide electrostatic discharge protection (ESD).

Channel breakdown is a reversible effect; i.e., the device is not destroyed immediately when such breakdown occurs. However, with excessive V_{DS} , MOSFETs degrade (age) rapidly because of the large population of hot carriers impinging on the silicon-oxide interface, as discussed previously. Therefore, MOSFETs are not operated intentionally in this region.

[†] In other models⁶⁴ for the substrate current, the multiplicative factor $(V_{DS} - V'_{DS})$ is replaced by a constant, and the factor 0.8 is introduced in front of V'_{DS} in the exponent.



(a)



(b)

FIGURE 5.20

(a) Effect of impact ionization on I_D - V_{DS} characteristics. Solid lines are with impact ionization included. Dashed lines are without impact ionization. (b) Snapback.

Models also exist in the literature for the gate current and device aging as a result of hot carriers.^{63,160,162,163,167} As gate oxide thickness decreases, hot-carrier-generated gate current can become nonnegligible. Nevertheless, the limit of oxide thickness is set by what is called *direct tunneling*, as discussed in Sec. 5.12, rather than by hot carriers. One of the reasons for this is that, as oxide thickness is scaled down, so is the power supply voltage, and as a result hot carrier effects rapidly diminish, because hot carrier currents decrease rapidly with decreasing V_{DS} , as seen from (5.8.4). The same is true for the device aging rate, which is rapidly reduced as power supply voltage is scaled down.

In our discussions of various other phenomena in this book, the substrate and gate currents will be assumed negligible unless specifically stated otherwise.

To limit hot electron effects, it is advantageous to have a lightly doped drain region.^{170–175} This is because a significant part of the depletion is then inside the drain, absorbing some of the potential that otherwise would exist in the pinchoff region and lowering the maximum electric field. However, a lightly doped drain in standard processes would mean a lightly doped source, since both regions are made simultaneously, in the same way. This would result in a large source resistance across which a large voltage drop could develop, which is undesirable, as has been discussed in Sec. 4.12. A compromise can be reached by using light doping only for part of the drain adjacent to the channel. The resulting structure, known as *lightly doped drain* (LDD),^{170,172} used to be popular in the past. It is still used for older digital technologies (gate lengths of 130 nm and greater), and for power technologies. In modern fine-line technologies source/drain extensions are used instead, as discussed in Sec. 1.7 (Fig. 1.38).

5.9 Velocity Overshoot and Ballistic Operation

At the microscopic level, moving carriers contribute to transport in a complicated fashion. They accelerate, gaining energy from the electric field, then transfer energy to the semiconductor lattice via interactions called “scattering” (Sec. 4.11), then accelerate and scatter again, and so on. The scattering interactions are statistical in nature, and so they can be averaged over time scales larger than the typical time interval between scattering events, or distances larger than the typical distance, called the “mean free path,” separating two such interactions. The drift-diffusion model we have used so far is based on average measures of the behavior of carriers. The drift velocity is the *average* velocity a carrier attains as it is scattered by the mechanisms we have discussed in Sec. 4.11; the concept of mobility relates average carrier velocity to the electric field. Such averaging makes sense if the channel length is much larger than the mean free path, which is of the order of 10 nm. For distances smaller than the mean free path, or for time scales smaller than it takes for a carrier to traverse a mean free path distance, carriers do not undergo significant scattering. Thus, before carriers are scattered, their velocity can attain values above the drift velocity. This is called *velocity overshoot*.^{†177–182}

[†] The use of this term here should not be confused with a different use of the same term in discussions of GaAs devices.

In modern devices, the channel length is smaller than 35 nm, and carriers may under some conditions go through the channel without being scattered. This mode of operation goes by the name of *ballistic transport*.^{183–187} It is nicely illustrated in Fig. 5.21,¹⁸² which shows results from a Monte Carlo simulation—a simulation in which the physics of a large number of individual carriers are included, and those carriers are individually tracked.[†] The horizontal axis is horizontal position along the channel, and the vertical axis is total carrier energy. The continuous line is the edge of the conduction band E_c , which represents electron potential energy and whose shape has been discussed in conjunction with Fig. 5.12. The flat parts are inside the source and drain, where electrons are in equilibrium with the lattice. A barrier exists near the source, as already discussed in Sec. 5.5. In the source, there are plenty of electrons, and some have energies that are above E_c by more than the barrier height; when these diffuse thermally toward the right, they can pass the barrier and then find themselves in the electric field caused by the drain voltage. As they move toward the right, the distance between their position along the energy axis and the E_c plot represents their kinetic energy. Some electrons lose kinetic energy due to scattering; their total energy drops, and they are thus found close to E_c . However, some electrons are seen to have moved horizontally all the way to the drain, at the energy level they had when they passed the barrier at the source. These have moved ballistically, without scattering, with kinetic energy increasing throughout their trip from source to drain; thus, as they approach the drain, their kinetic energy, and thus their velocity, is maximum. This velocity can exceed the thermal velocity (i.e., the average velocity they acquire due to their thermal motion at the silicon lattice temperature, about 1.8×10^7 cm/s), and these electrons are called “hot electrons” for this reason (Sec. 5.8). Their velocity also exceeds the saturation velocity we discussed in Sec. 5.2 (about 8×10^6 cm/s) by a significant factor (e.g., a factor of 2 or even more).

It is interesting to observe that energetic electrons exist not only at the source, but also at the drain, as seen in Fig. 5.21. However, with a significant V_{DS} applied, practically none of the electrons at the drain can enter the channel and reach the source, as they have to overcome a huge energy barrier (see Fig. 5.12), and the field does not help them. If V_{DS} were zero, the energy barriers at the source and drain would be equal, and electrons could then be expected to enter the channel from the drain, just like they do from the source. This creates two opposite channel currents, which cancel each other out. Thus I_{DS} is zero when $V_{DS} = 0$, as expected.

Let us assume $V_{DS} > V'_{DS}$. In order to describe the operation of both classical and ballistic devices in a unified manner, one can use an expression for current in terms of carrier velocity, similar to (5.2.6), but using, instead of drift velocity, an effective velocity, v_{eff} . Further, because such an equation is valid for any x in the channel, we can choose to apply it for $x = 0$, which is defined as the point where the barrier is maximum (i.e., at the point where E_c reaches its peak). This point is very close to the edge of the physical source and is sometimes called the “virtual source.” We thus have

$$I_{DS} = W[-Q'_I(0)]v_{\text{eff}}(0) \quad (5.9.1)$$

[†] This figure is for a double-gate device, but the physics it illustrates is valid for single-gate devices as well.

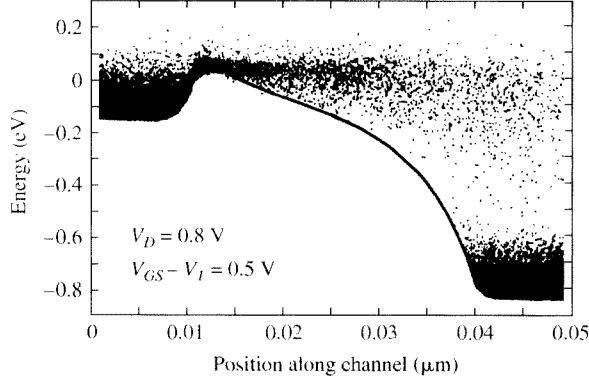


FIGURE 5.21
Results of Monte Carlo simulation of carriers in the channel. Shown is the bottom of the conduction energy band, and individual electrons in the source, channel, and drain (Ref. 182, © 1992 by IEEE).

The effective velocity is given by

$$v_{\text{eff}}(0) = B v_{\text{inj}}(0) \quad (5.9.2)$$

In this equation, $v_{\text{inj}}(0)$ is the injection velocity at $x = 0$, a quantity that is approximately equal to the thermal velocity[†] at low charge densities, but can significantly exceed this value at high charge densities, due to quantum-mechanical effects.¹⁸⁵ The factor B is the ballistic efficiency factor; it is equal to 1 for purely ballistic operation and is less than 1 otherwise. In nonballistic operation, this nonideality factor takes into account “backscattering,” a phenomenon in which an electron, having just left the source, is scattered, changes direction, and re-enters the source. Obviously, such electrons do not contribute to the drain current. At the time of this writing, minimum-size devices exhibit a B value of about 0.6.¹⁸⁷

Note that (5.9.1) and (5.9.2) can be used to describe even the operation of long-channel devices in nonsaturation by making B a function of V_{DS} . Thus, consider a long-channel device operating in nonsaturation. We can consider two flows of electrons, one from the source and one from the drain; the net drain current corresponds to the difference of the two and becomes zero when $V_{DS} = 0$, as discussed earlier. This corresponds to $B = 0$ in (5.9.2). When $V_{DS} > 0$, the dependence of I_{DS} on V_{DS} in nonsaturation can be explained using the same equations, as follows: As the drain potential rises, the downward-going part of the E_c plot becomes steeper, and the hump of the energy barrier near the source becomes narrower. Thus, the field at points to the right of the peak increases. Thus, more of the carriers that find themselves in that vicinity due to their thermal motion are likely to be carried to the right by the field,

[†] A measure of the speed involved in the thermal motion of carriers and thus a measure of the temperature of the material. Ref. 185 discusses which of several existing definitions for this quantity should be used in this context.

and the backscattering effect is reduced. B thus increases; it becomes maximum in saturation.

This viewpoint is sometimes used as the basis of modeling MOS transistors.¹⁸⁸

5.10 Polysilicon Depletion

The standard material for gate formation in modern MOSFET technologies is polycrystalline silicon. So far, we have implicitly assumed that this material is heavily doped to behave almost like a metal. However, in many technologies it is not possible to dope the polysilicon gate to arbitrarily high concentrations, and a careful analysis requires treating it as what it is—a semiconductor, which can exhibit accumulation, depletion, and inversion behavior, just as the body does. It is typical to dope the gate *n*-type in *n*MOS devices and *p*-type in *p*MOS devices. The gate then depletes when an inversion layer is formed in the channel beneath it.^{189–197} For example, in an *n*MOS device, the negative charges in the body repel the electrons in the gate polysilicon, leaving a depleted region in that material adjacent to the oxide, as shown for a two-terminal MOS structure in Fig. 5.22. This phenomenon is generally referred to as “polysilicon depletion” or “poly depletion.” It alters the details of charge balancing in a transistor and so affects device characteristics. Its effect, as will be seen, is to increase the distance between the inversion layer charges and the controlling mobile gate charges; i.e., it increases the effective oxide thickness. It is thus mostly of importance when the oxide is thin, which is why it is treated in this chapter on small-dimension effects. Although it is expected that the use of metal gates will obviate this phenomenon, most technologies at the time of this writing use polysilicon gates, and so the effect is still of importance.

Before providing an analytical description of the poly depletion effect, let us consider its consequences qualitatively. As indicated in Fig. 5.22, we now have a potential drop across the polysilicon depletion region, analogous to the surface potential between the SiO_2 interface and the bulk. Just as the surface potential ψ_s from the oxide interface to the bulk in the *p*-type substrate is positive when the latter is depleted or inverted, the corresponding potential from the upper oxide interface to deep in the *n*-type poly gate is negative; we define ψ_{poly} in the opposite direction as shown in the figure, and thus ψ_{poly} is a positive quantity when the poly depletes. Because of this potential drop, there is less potential available to induce inversion in the silicon for a given V_{GB} , so the magnitude of the inversion charge Q'_I is reduced. The potential drop ψ_{poly} increases with increasing V_{GB} ; therefore the reduction in the magnitude of Q'_I (compared with the situation without the effects of polysilicon depletion) depends on V_{GB} . For a complete transistor, this results in a reduction in the drain current. Also, note that the gate depletion region capacitance is now in series with the oxide capacitance, just like the body capacitance is in series with it [this fact, which is rather obvious intuitively, can be deduced analytically in a manner analogous to that used in Sec. 2.7 to derive (2.7.8)]. This reduces the total gate capacitance, C'_{gb} , to values lower than those expected in the absence of poly depletion, as shown in Fig. 5.23. Note that as V_{GB} is increased, the polysilicon depletes more, and the depletion region capacitance is decreased; this effect counteracts the increase of the inversion layer capacitance, so a maximum is observed, as

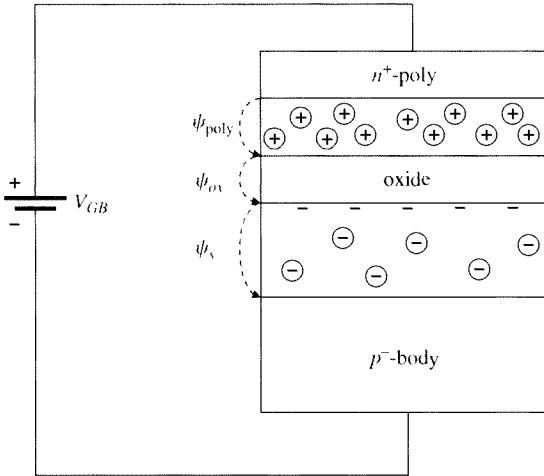


FIGURE 5.22
Polysilicon depletion.

shown. Finally, at large values of V_{GB} , the poly depletion region becomes so deep, and the potential across it so large, that the gate region adjacent to the oxide eventually inverts; then plenty of holes provide an effective plate just above the oxide (effective, that is, at low frequencies), and the capacitance goes up toward the value expected in the absence of polysilicon depletion (not shown in Fig. 5.23).

We will now give the fundamental equations for the two-terminal structure of Fig. 5.22, for the case of poly depletion. The potential drop ψ_{poly} , and the charge per unit area, Q'_{poly} (the latter now playing the role of the gate charge per unit area Q'_G), must be included in the potential balance and charge balance equations that we developed in (2.3.1) and (2.3.3). Thus, we have

$$V_{GB} = \psi_{poly} + \psi_{ox} + \psi_s + \phi_{MS} \quad (5.10.1)$$

and

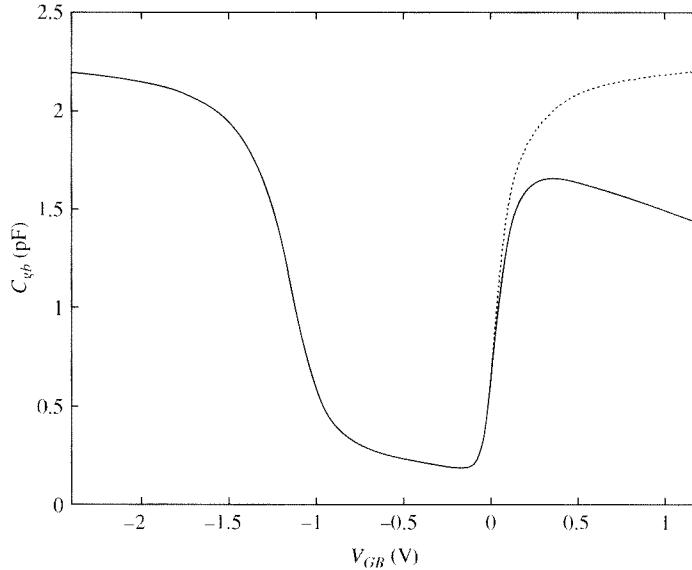
$$Q'_{poly} + Q'_o + Q'_I + Q'_B = 0 \quad (5.10.2)$$

The charge-potential relation for the oxide capacitor becomes

$$Q'_{poly} = C'_{ox} \psi_{ox} \quad (5.10.3)$$

Finally, Q'_{poly} and ψ_{poly} are related by an equation analogous to (2.6.7):

$$Q'_{poly} = \sqrt{2q\epsilon_s N_{poly}} \sqrt{\psi_{poly}} \quad (5.10.4)$$

**FIGURE 5.23**

Capacitance-voltage characteristics for a two-terminal MOS structure with polysilicon depletion not included (broken line) and included (solid line). $t_{\text{ox}} = 1.5 \text{ nm}$, $N_A = 3 \times 10^{17} \text{ cm}^{-3}$, $N_{\text{poly}} = 5 \times 10^{19} \text{ cm}^{-3}$, $W = L = 10 \mu\text{m}$.

Thus, together with the equations we have developed for $Q'_I(\psi_s)$ and $Q'_B(\psi_s)$ in Chaps. 2 and 3, we have a system of six equations in six unknowns: Q'_{poly} , Q'_I , Q'_B , ψ_{poly} , ψ_{ox} and ψ_s . Various results can thus be derived for the MOS structure in the case of poly depletion.^{192–194}

Weak Inversion As an example, let us estimate several quantities of interest in weak inversion, where Q'_I can be neglected in the charge balance equation, for the structure of Fig. 5.22. We begin with the relative values of ψ_{poly} and ψ_s ; assuming Q'_o is negligible, from (5.10.2) we get $Q'_{\text{poly}} = -Q'_B$. Using in this (5.10.4) and (2.6.7), we obtain¹⁹⁵

$$\psi_{\text{poly}} = \frac{N_A}{N_{\text{poly}}} \psi_s, \quad \text{Weak inversion, } Q'_o = 0 \quad (5.10.5)$$

The poly depletion region capacitance per unit area can be found by differentiating (5.10.4) with respect to ψ_{poly} :

$$C'_{\text{poly}} = \frac{\sqrt{2q\epsilon_s N_{\text{poly}}}}{2\sqrt{\psi_{\text{poly}}}} \quad (5.10.6)$$

Using the preceding two equations and the formula for the body capacitance, C'_b [(2.7.16), with the exponential neglected in depletion], it is easy to see that

$$C'_{\text{poly}} = \frac{N_{\text{poly}}}{N_A} C'_b \quad (5.10.7)$$

The gate controls the channel no longer through the oxide capacitance C'_{ox} directly, but rather through the series combination of C'_{poly} and C'_{ox} , which we will denote by $C'_{\text{poly-ox}}$, and which satisfies

$$\frac{1}{C'_{\text{poly-ox}}} = \frac{1}{C'_{\text{poly}}} + \frac{1}{C'_{\text{ox}}} \quad (5.10.8)$$

It is sometimes convenient to express the preceding effect in terms of the effective oxide thickness, \hat{t}_{ox} , one would need to have in order to obtain an oxide capacitance equal to $C'_{\text{poly-ox}}$, assuming an ideal gate. Using in the above equation $C'_{\text{poly-ox}} = \epsilon_{\text{ox}} \hat{t}_{\text{ox}}$, $C'_{\text{poly}} = \epsilon_s / d_{\text{poly}}$ [in analogy to (1.5.25)] in (5.10.8) and $C'_{\text{ox}} = \epsilon_{\text{ox}} / t_{\text{ox}}$, we easily obtain $\hat{t}_{\text{ox}} = t_{\text{ox}} + (\epsilon_{\text{ox}} / \epsilon_s) d_{\text{poly}}$; this means that the oxide thickness effectively increases by the amount:

$$\Delta t_{\text{ox}} = \frac{\epsilon_{\text{ox}}}{\epsilon_s} d_{\text{poly}} \quad (5.10.9)$$

where, in analogy with (1.5.16), we have

$$d_{\text{poly}} = \sqrt{\frac{2\epsilon_s}{qN_{\text{poly}}}} \sqrt{\psi_{\text{poly}}} \quad (5.10.10)$$

Let us now calculate the weak inversion slope factor, n . Substituting $C'_{\text{poly-ox}}$ for C'_{ox} in (2.7.25) and using (5.10.8) and (5.10.7), we can easily show that this quantity needs to be increased above the value given by (2.7.25) by¹⁹⁵

$$\Delta n = \frac{N_A}{N_{\text{poly}}} \quad (5.10.11)$$

This has an effect on complete transistor characteristic. Thus, the slope of the plot of $\log I_D$ vs. V_{GB} in weak inversion, which is proportional to n (Sec. 4.8), is reduced, which is an undesirable effect.

Equations (5.10.5), (5.10.7), and (5.10.11) make it obvious that as the poly doping concentration is increased, the effect of poly depletion gets reduced, as expected.

Strong Inversion The additional potential drop ψ_{poly} in the polysilicon depletion region shows up directly in series with the gate and will manifest itself as a shift in the effective gate voltage. To see this, recall that in strong inversion the surface

potential is assumed pinned at the value $\psi_s = \phi_0$, given by (2.6.17). Going through a development analogous to that in Sec. 2.6.2, we obtain a relation similar to (2.6.20):

$$Q'_I = -C'_{\text{ox}}(V_{GB} - \hat{V}_{T0}) \quad (5.10.12)$$

where

$$\hat{V}_{T0} = V_{T0} + \psi_{\text{poly}} \quad (5.10.13)$$

with V_{T0} given by the usual equation (2.6.22). Thus, the threshold is effectively shifted by the potential across the poly depletion region. As V_{GB} is varied, this potential will vary slightly, and so will V_{T0} . For low values of V_{GB} , where Q'_I is small, one can use $\psi_s = \phi_0$ in (5.10.5) and the result in (5.10.13), obtaining^{193,194}

$$\hat{V}_{T0} \approx V_{T0} + \frac{N_A}{N_{\text{poly}}} \phi_0 \quad (5.10.14)$$

This is seen to reduce to V_{T0} when the polysilicon is very heavily doped, i.e., when the poly acts as a metal, which is what we have assumed before this section.

General Models In the general case, the equations given have no analytic solution and must be solved iteratively to accurately take into account poly depletion.¹⁹⁵ The effect of finite polysilicon doping does generate an observable shift in device characteristics, but the shift does not drastically change device behavior, and not all possible conditions of operation (accumulation through inversion) are encountered in the polysilicon in normal operation. So it is not necessary in practice to implement the full complexity of a coupled solution for ψ_s and ψ_{poly} . Rather, poly depletion can be treated as a perturbation to the analyses we have presented so far, which leads to simpler yet still sufficiently accurate models.[†]

[†] An efficient and practical procedure for taking polysilicon depletion into account in surface potential-based models involves three steps: Compute ψ_s assuming there is no polysilicon depletion; compute ψ_{poly} based on that value of ψ_s , assuming that the polysilicon is depleted (but not inverted or accumulated); and then iterate once, generating an update value for ψ_s . This is substantially more efficient in terms of computation time than doing a detailed coupled solution for both potentials. Then, $Q'_I(\psi_s)$ and $Q'_B(\psi_s)$ can be determined as usual.

For a transistor, ψ_s varies with position along the channel and so will ψ_{poly} . This can be taken into account at the cost of complexity. However, in strong inversion, it is simpler to just calculate the potential drop across the polysilicon region at the source end of the channel and then reduce the effective ψ_s used in the standard calculations, without poly doping, by that amount; this has proven effective in BSIM4.¹⁹⁷ For this approach, ψ_s at the source is approximated as $\phi_0 + V_{SB}$, and this is used to calculate ψ_{poly} ; this approximately captures the bias dependence of ψ_{poly} .

5.11 Quantum Mechanical Effects

We have so far assumed that the inversion layer charge forms a charge sheet of small thickness, with peak charge density immediately next to the semiconductor-oxide interface. In the energy band diagram of Fig. 2.7d, this means that the inversion layer is concentrated over a negligible distance corresponding to the minimum point of conduction band edge E_c ; the charges are then said to exist in a “potential well,” determined by the oxide interface on one side and E_c on the other. In transistors with very thin oxides and large substrate doping, the electric field magnitude and the band bending near the surface become large, and the potential well becomes very narrow. The inversion electrons then follow the physics of tightly confined particles,¹⁹⁸ which require quantum mechanical treatment. There are two main effects that are caused by this.

One is that the inversion layer peak is found not at the interface, but away from it,¹⁹⁹ as shown in Fig. 5.24,²⁰⁰ by a distance d_m . The inversion layer is now further away from the gate, and the control exerted by the latter is somewhat smaller. To model this effect in a simple way, we will assume that the inversion layer is still a charge sheet of zero thickness as before, but is now located where the peak of the distribution is, i.e., a distance d_m from the interface.[†] Effectively, this is as if we had a thicker oxide; the situation is similar to that encountered for poly depletion in Sec. 5.10. Following the reasoning there, we can write an equation analogous to (5.10.9) for the effective increase in the oxide thickness:¹⁹⁹

$$\Delta t_{\text{ox,qm}} = \frac{\epsilon_{\text{ox}}}{\epsilon_s} d_m \quad (5.11.1)$$

If both poly depletion and quantum mechanical effects are present, the two corrections shown in (5.10.9) and (5.11.1) are added together.

The second quantum mechanical effect is that not all of the energy levels above E_c can be occupied; the allowable energy levels split into discrete “sub-bands,” with energy levels *above* that of the classical conduction band edge. As a result, the semiconductor behaves as if it has an increased energy gap, by an amount that depends on the transverse field and increases with it.^{199,201} The increased energy gap makes it more difficult for electrons to jump from the valence band to the conduction band and results in a decreased intrinsic carrier concentration. This then increases the effective Fermi potential, according to (1.2.13a). It now takes a larger amount of band bending (Sec. 2.4) to reach a certain level of inversion;²⁰² this corresponds to an increase in the surface potential by an amount which varies with the transverse electric field. This affects the device characteristics; it leads, for example, to an increased value for the threshold voltage in *n*MOS devices. If the pinned value for the surface potential in strong inversion is ϕ_0 when no quantum mechanical effects are present (assuming $V_{SB} = 0$), this quantity will increase to a value $\phi_0 + \Delta\phi_0$ when such effects are present.

[†] Other approaches use other depths for the location of this effective charge sheet, e.g., the “centroid” (analogous to the center of gravity) of the distribution. Such approaches are sometimes used even in the absence of quantum-mechanical effects to model the effect of finite thickness for the inversion layer.

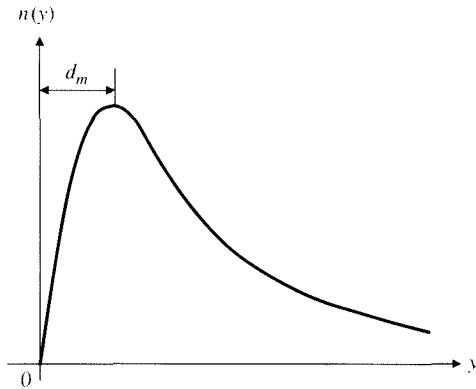


FIGURE 5.24
Electron concentration per unit volume vs. distance from the surface in the presence of quantum effects.

Using the latter quantity in lieu of ϕ_0 in (2.6.22), leads to an increased value for V_{T0} , and it is easily seen that the threshold increase will be given by

$$\Delta V_{T0} = \Delta\phi_0 + \gamma \left(\sqrt{\phi_0 + \Delta\phi_0} - \sqrt{\phi_0} \right) \quad (5.11.2)$$

The determination of d_m and $\Delta\phi_0$ in (5.11.1) and (5.11.2) requires quantum mechanics; one has to solve a coupled system consisting of Poisson's equation, which governs electrostatics, and Schrödinger's equation, which governs the behavior of tightly confined particles, and make certain approximations. Such solutions are beyond the scope of this book, so here we will only quote the result and attempt to make it qualitatively plausible. The quantity d_m depends on the transverse field that, as we know from Sec. 4.11, depends on Q_B and Q'_I ; indeed, the result turns out to contain these quantities^{199,203}

$$d_m = B_1 \left| Q'_B + \frac{11}{32} Q'_I \right|^{-1/3} \quad (5.11.3)$$

where $B_1 \approx 10^{-9} (\text{C} \cdot \text{cm})^{1/3}$, but this value may have to be adjusted empirically due to the simplifications made in obtaining this equation.¹⁹⁶ When used in (5.11.1), the preceding equation results in a $\Delta t_{\text{ox,qm}}$ of typically a few tenths of 1 nm.

Similarly, an approximation has been developed for $\Delta\phi_0$:^{204,205}

$$\Delta\phi_0 = B_2 |Q'_B|^{2/3} + \frac{|Q'_B|}{\epsilon_s} d_m \quad (5.11.4)$$

where $B_2 \approx 500 \text{ V}/(\text{C} \cdot \text{cm}^{-2})^{2/3}$. The value of $\Delta\psi_s$ is typically one or two tenths of 1 V.

Note that, strictly speaking, in order to determine Q'_B and Q'_I one needs to know d_m and $\Delta\phi_0$. Thus, the required equations are solved iteratively.¹⁹⁶

In a transistor, in principle, one would have to account for the preceding effects all along the channel, if $V_{DS} > 0$. However, the issue of carrier confinement very close to the surface is mostly present near the source, where the transverse field is highest, and it turns out that it can be ignored further away from the source.²⁰⁶

Figure 5.25 shows how quantum mechanical effects alter the small-signal gate capacitance and the I_{DS} - V_{DS} characteristics of a transistor. The onset of strong inversion, where the gate capacitance increases significantly with increasing gate voltage, occurs at a higher gate voltage when quantum effects are taken into account than when they are not; this is the threshold voltage increase just discussed. The gate capacitance in strong inversion is smaller when quantum effects are included, as expected from the increased effective oxide thickness in (5.11.1). Both effects are rather negligible in weak inversion.

For future generations of devices, other quantum mechanical effects will need to be considered as well; for example, quantum mechanical tunneling may occur directly between source and drain for channels shorter than 10 nm.²⁰⁷

5.12 DC Gate Current

For many years, one of the classical advantages of MOSFETs compared with bipolar transistors was that in dc operation, MOS transistors do not have current flowing in the control terminal (the gate, in the case of a MOSFET); this is one reason that MOSFETs were called IGFETs (insulated gate field effect transistors) in the early days. This property has been very important for certain types of circuits. The reason

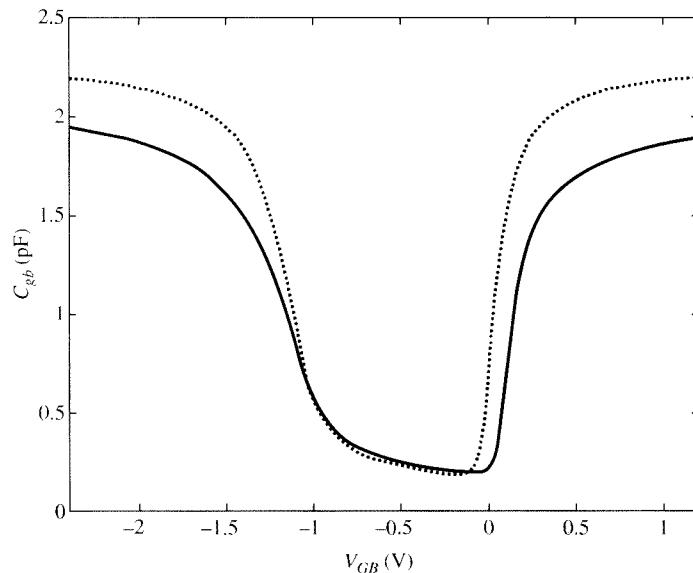


FIGURE 5.25
Capacitance-voltage characteristics for a two-terminal MOS structure without (dashed line) and with (solid line) quantum effects. Polysilicon depletion is not included. $t_{ox} = 1.5$ nm, $N_A = 3 \times 10^{17}$ cm $^{-3}$, $W = L = 10$ μ m.

there was no dc gate current flow was that the dielectric, which isolates the gate from the silicon surface where the conducting channel is formed, behaved as an electrical insulator. In terms of energy bands, the conduction band edge for SiO_2 is about 3.1 eV higher than that of silicon, and thus the electrons in the channel face a high energy barrier. This is shown in Fig. 5.26a, which is similar to figures we have already seen (e.g., Fig. 2.7d) and assumes a rather thick oxide and gate voltage sufficiently above flatband to cause inversion at the surface. The electrons are thus confined to “a potential energy well,” with the oxide conduction band on one side and the silicon conduction band on the other. One such confined electron is shown schematically in Fig. 5.26a.

With the continued evolution (scaling) of CMOS technologies, the gate dielectric has become thinner in order to maintain the historic trend of increased MOS transistor “drive” currents, which in turn leads to increased switching speeds of digital CMOS circuits. Until the 0.18- μm technology generation, the assumption that there was no gate current flow was reasonable. For technology nodes since then, the gate dielectric has become so thin that the oxide can no longer be relied upon to be a good insulator. As we have already seen with other effects (Sec. 5.11), when dimensions become very small, it is necessary to evoke quantum mechanical considerations. Such considerations dictate that even if electrons do not have the required kinetic energy to go over the energy barrier, there is a nonzero probability that they can find themselves beyond that barrier.^{208,209} This phenomenon is called “tunneling.” Even for oxides that are not very thin (e.g., 4 nm or thicker), tunneling allows a nonnegligible number of electrons in the channel to “pierce the energy barrier” near the top, where that barrier has a triangular shape, and find themselves in the oxide’s conduction band; this is referred to as “Fowler-Nordheim tunneling.” Such electrons

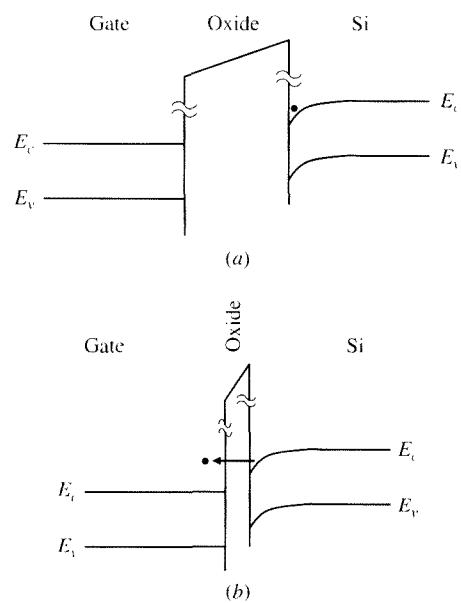


FIGURE 5.26
Energy band diagram (a) for a thick oxide and (b) for a thin oxide, schematically indicating tunneling.

can thus find their way to the gate. However, the currents that result are generally extremely small, and for most practical cases, they can be neglected for common voltage ranges.

For oxides thinner than about 3 to 4 nm, a different form of tunneling dominates: Electrons can now pierce the energy barrier at the level of the silicon conduction band and find themselves on the other side of the oxide, i.e., in the gate, as indicated schematically in Fig. 5.26b. This is called *direct tunneling*, and it is a phenomenon of major importance in modern technologies, as it can result in significant de gate current. This current, which is a strong function of gate and drain voltage, depends drastically on oxide thickness; in fact, *a reduction of oxide thickness by 0.5 nm can increase this current by two orders of magnitude!* The oxide thus has to be maintained sufficiently thick so that, under worst-case tolerances, it is still thick enough to keep this current acceptably small.

As already explained, with a sufficiently positive gate potential, electrons tunnel from the channel to the gate. Because they carry a negative charge, this is equivalent to a positive current entering the gate. If, instead, the gate potential is low, the opposite happens: Electrons now tunnel from the polysilicon gate, through the oxide, to the substrate, thus giving rise to a negative gate current.

Figure 5.27 shows schematically gate current flow for a MOSFET. A main component of gate current is from tunneling through the gate dielectric between the gate and the channel. There are also components that flow between the gate and the overlap regions at the source and drain ends of the channel, and for large negative biases on the gate, when there is no channel, there is a gate-bulk tunneling component.

There are several different formulations of gate current direct tunneling. Unfortunately, they do not lead to simple analytical expressions. In one approach, the gate current density $J_G(x)$ (gate current per unit of gate area) at horizontal position x is taken to be proportional to the inversion layer charge magnitude $-Q'_I$, the field in the oxide $\mathcal{E}_{ox}(x)$, and the “tunneling probability” $p_t(x)$:²¹⁰

$$J(x) = B_t(-Q'_I)\mathcal{E}_{ox}(x)p_t(x) \quad (5.12.1)$$

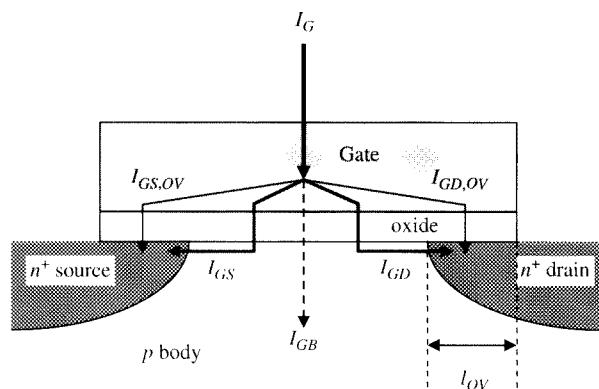


FIGURE 5.27
Gate current components.

where B_t is a constant of proportionality. The tunneling probability is then expressed as a function of $\mathcal{E}_{ox}(x)$ which, along with Q'_l can be found for a given x once the external bias voltages are known. The gate current contributed by an element of area $W\Delta x$ around horizontal position x is $J_G(x)W\Delta x$. To calculate the total gate-to-channel current, I_{GC} , we let Δx become a differential and integrate over the channel length:

$$I_{GC} = W \int_0^L J_G(x) dx \quad (5.12.2)$$

The gate-to-channel current eventually makes its way out of the source and drain terminals. It is thus partitioned into two components, one flowing from gate to source, I_{GS} , and another flowing from gate to drain, I_{GD} . Detailed analysis suggests a rather intuitive result: The closer a point x in the channel is to the source, the more the current density there contributes to I_{GS} ; and the closer it is to the drain, the more it contributes to I_{GD} . Thus, it has been shown that to determine the two components, $J_G(x)$ should be weighted by normalized distances as follows^{210,211} (we will see in Chap. 6 that this same partitioning is also applicable for source and drain charge calculations):

$$I_{GS} = W \int_0^L J_G(x) \left(1 - \frac{x}{L}\right) dx \quad (5.12.3)$$

and

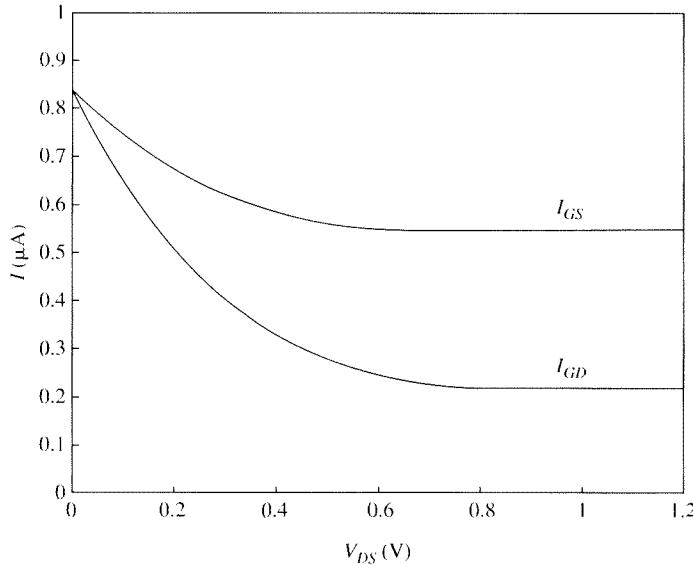
$$I_{GD} = W \int_0^L J_G(x) \frac{x}{L} dx \quad (5.12.4)$$

From (5.12.2) to (5.12.4), we can verify that

$$I_{GS} + I_{GD} = I_{GC} \quad (5.12.5)$$

Evaluation of the preceding integrals is complex; the interested reader is referred to Ref. 212 for details. These currents are in addition to the drain-source current I_{DS} and any possible impact ionization current (Sec. 5.8). I_{DS} is defined as the channel current in the absence of gate leakage.²¹¹

The partitioning of the gate-to-channel current into source and drain components is shown in Fig. 5.28 vs. V_{DS} . As expected, this partition is 50-50 when $V_{DS} = 0$; as V_{DS} is increased, the gate-to-drain voltage decreases, and thus so does the corresponding tunneling current component. Most current then flows from gate to source. With large V_{DS} , saturation occurs and the channel, according to the usual long-channel simplified picture (Sec. 4.7) becomes pinched off. The potential difference across the channel then becomes the saturation voltage, V'_{DS} , and the voltage between the gate and the tip of the pinched-off channel becomes $V_{GS} - V'_{DS}$, which is independent of further increases in V_{DS} , under the usual long-channel assumptions. Thus the value of the gate-drain current is set by that voltage and is not affected when V_{DS} is raised further. This voltage is positive for an enhancement-mode transistor. The gate-channel current is generated in a distributed manner along the channel, so I_{GD} depends not only on

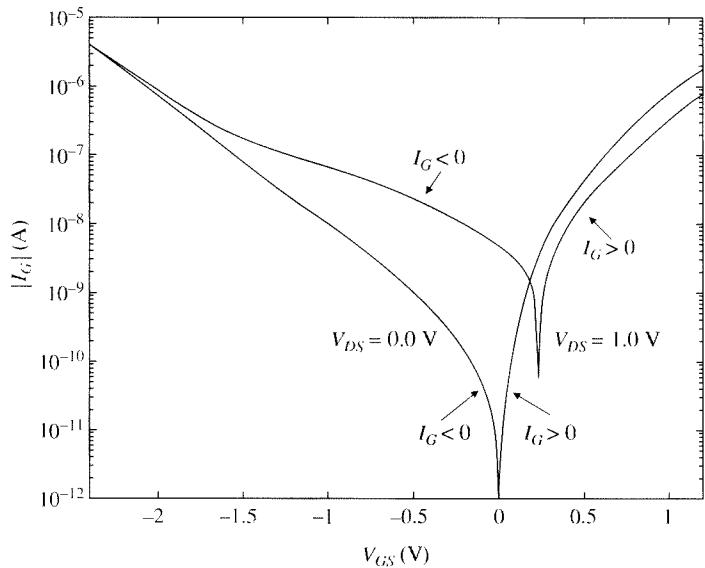
**FIGURE 5.28**

Gate-source and gate-drain components of gate leakage current (not including overlap regions) vs. V_{DS} : $t_{ox} = 1.5 \text{ nm}$, $N_A = 3 \times 10^{18} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 2 \mu\text{m}$, $V_{GS} = 1.2 \text{ V}$, $V_{SB} = 0 \text{ V}$.

$V_{GS} - V'_{DS}$ but also on the value of ψ_s along the channel; this explains why I_{GD} is positive even though $V_{GD} = 0$ for the highest V_{DS} value in Fig. 5.28.

Additional gate current occurs between the gate and the n^+ source and drain regions. Assuming for simplicity that the overlap regions are characterized by a uniform field, these currents are obtained by multiplying the corresponding current density by the overlap area, Wl_{ov} . The current density is computed from the surface potentials in the source and drain regions, using appropriate doping densities. When the device is off, because only the overlap part conducts, for minimum-sized devices, the drain-gate current that results can be of the order of 10 percent that of the gate current when the device is on; still, this is an important concern in digital operation. This current is sometimes called “edge-directed tunneling current.”

Figure 5.29 shows the magnitude of the gate current as a function of V_{GS} , for two values of V_{DS} . The values are for a large-area transistor (see caption) in order to help reveal the effects of the various current components in particular regions. To interpret these plots, let us first concentrate on the right part of the plot for $V_{DS} = 0$, with large V_{GS} . The gate-source and gate-drain voltage is then large, and we have large positive gate-source and gate-drain components, including the ones in the overlap regions. As V_{GS} is reduced, the voltage across the oxide decreases, and thus so does the gate current. With sufficiently small V_{GS} , the inversion layer disappears, and thus there is no longer a gate-channel current; the gate current is then dominated by the current in the overlap regions. With $V_{GS} = 0$, this current becomes zero as shown. If now V_{GS} is further decreased to negative values, the voltage across the overlap regions reverses polarity; the gate is now more negative than the source/drain regions, and

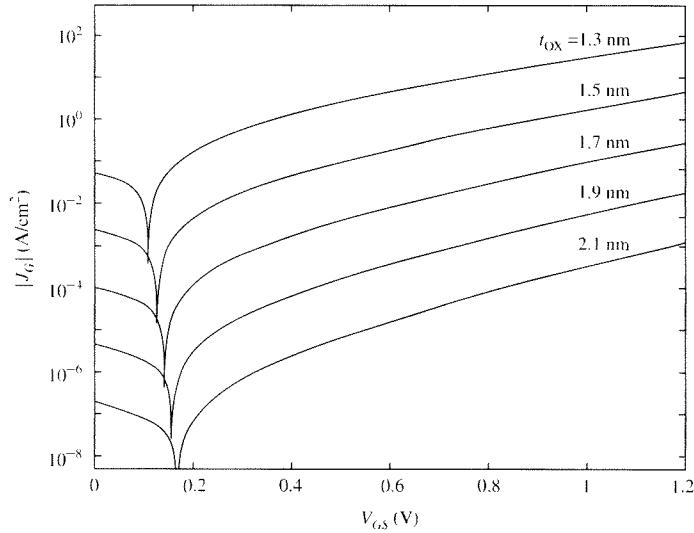
**FIGURE 5.29**

Total gate leakage channel current vs. V_{GS} for two values of V_{DS} . $t_{ox} = 1.5 \text{ nm}$, $N_A = 3 \times 10^{18} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 2 \mu\text{m}$.

thus electrons tunnel *from* the gate *to* those regions, resulting in negative gate current as indicated (note that the vertical axis necessarily shows the *absolute value* of the current, because this axis is logarithmic). Making V_{GS} more negative makes the gate current through the overlap regions more negative. Finally, at sufficiently negative V_{GS} values, the surface of the *p*-type substrate accumulates. The tunneling of electrons now occurs from the polysilicon gate to the substrate, resulting in a negative gate current, I_{GB} . The mechanism involved and the analysis required for this current are similar to the ones described previously, only somewhat simpler, as the device operates in accumulation and thus the surface potential does not vary with position. Also, no partitioning between source and drain is needed; the entire current flows between gate and body.

The second plot in Fig. 5.29 is for $V_{DS} = 1 \text{ V}$. This plot can qualitatively be interpreted in the same way as the first plot, but now the drain is positive, and thus when $V_{GS} = 0$ electrons tunnel from the gate to the positive drain in the overlap region, resulting in a negative gate current as shown. As V_{GS} is raised, this negative current decreases in magnitude, whereas the gate-source current becomes positive and starts increasing. At some point, the two currents cancel each other out and the net gate current becomes zero; this is the point where the notch is observed for that plot. A transistor would operate to the left of this notch for digital operation in the off condition, and to the right of this notch for digital operation in the on condition or for analog operation.

As already mentioned, the increase of the gate current as oxide thickness t_{ox} is reduced is drastic; a plot for various oxide thickness values is shown in Fig. 5.30.

**FIGURE 5.30**

Total gate-channel current density vs. V_{GS} for several oxide thickness values. $N_A = 1 \times 10^{18} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 2 \mu\text{m}$, $V_{DS} = 0.5 \text{ V}$, $V_{SB} = 0 \text{ V}$.

Suppressing gate leakage has become one of the leading drivers for continued CMOS technology evolution, because gate leakage has increased to the point where it is a substantial contributor to the static power dissipation of large digital circuits such as microprocessors. Controlling power dissipation has become one of the most important constraints in the design of such circuits. Because of the sensitivity of gate current to oxide thickness, it has not been possible to scale down the oxide thickness significantly in recent technology generations. This poses a problem, as drain current is proportional to the oxide capacitance per unit area, $C'_{ox} = \epsilon_{ox}/t_{ox}$, and thus if the oxide thickness is not scaled down, gate control is reduced. This has led to considerable research on other materials, with permittivities higher than ϵ_{ox} . These materials are typically hafnium-based. They allow the gate dielectric thickness to be increased, thereby reducing tunneling, with no loss in control of the inversion channel charge.

Oxide Breakdown As has already been mentioned, the gate tunneling current is a sensitive function of the transversal electric field in the oxide. When safe voltage limits across the oxide are exceeded, the magnitude of the gate current can rise abruptly to very large levels; the oxide is then said to have reached dielectric breakdown, and this can lead to permanent failure. The mechanisms of dielectric breakdown are quite involved. A summary is provided in Ref. 213. Typically, oxide fields at which breakdown occurs are about 10^7 V/cm (1 V/nm) or even higher. This means, for example, that a device with a 2-nm oxide can withstand about 2 V before it breaks down. Typically, devices are used at voltages up to half the oxide breakdown limit or less.

5.13 Junction Leakage; Band-to-Band Tunneling; GIDL

The high doping concentration used for the source/drain and body in modern devices can lead to large reverse-bias junction leakage currents. As expected from Sec. 1.5, large doping concentrations lead to narrow depletion regions around the source and drain junctions, which essentially means that the reverse-bias voltage must be dropped over a small distance, resulting in high electric fields. These can cause impact ionization in the junction depletion regions (via mechanism similar to those discussed in Sec. 5.8), and the leakage can be higher than that expected classically from the description in Sec. 1.5. The large potential gradient also means that the bands must bend considerably over a small distance, as shown in Fig. 5.31. This leads to a form of tunneling through the narrow depletion region, similar to what we have through thin oxides (Sec. 5.12). In terms of energy bands, this is schematically indicated in Fig. 5.31; an electron can tunnel from the valence band on the *p* side to the conduction band on the *n* side.²¹⁴ leaving behind a hole. This is called *band-to-band tunneling*. The reasons for this phenomenon are again quantum mechanical in nature, just as was the case with tunneling through the oxide, discussed in Sec. 5.12. The electrons and holes generated in this manner are separated by the field and contribute to the leakage current. Currents due to band-to-band tunneling become considerable at fields of 10^6 V/cm or more. For a one-sided *pn* junction, it is easy to calculate from (1.5.8) and (1.5.16) that this field value is reached with substrate dopings of about 2×10^{18} cm⁻³, when the reverse-bias voltage is 1 V. Derivations, which are involved, predict a current of the following form:²¹⁴

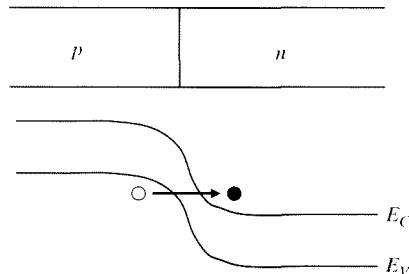
$$I_{BTB} = A_{BTB} V_R \mathcal{E}_{\max} e^{-B_{BTB} f \mathcal{E}_{\max}} \quad (5.13.1)$$

where V_R is the reverse-bias voltage, \mathcal{E}_{\max} is the maximum electric field magnitude [e.g., from (1.5.8) for an one-sided junction], and A_{BTB} and B_{BTB} are fitting parameters, with A_{BTB} proportional to the junction area. Current densities of over 1 A/cm² can occur in modern devices with maximum reverse junction bias.[†] The preceding phenomenon can be enhanced by interface traps at the silicon surface (surface states, see Sec. 2.2), resulting in what is called “trap-assisted band-to-band tunneling,” or simply “trap-assisted tunneling.” The leakage components need to be evaluated separately for the bottom wall and for the perimeter wall of the drain-body and source-body junctions, due to the different doping and edge effects in each of these walls.

For extremely short channels (less than 10 nm), direct tunneling can also occur between drain and source.²⁰⁷

Gate-Induced Drain Leakage Junction leakage can be enhanced by the electric field caused by the gate overlap with the drain region. To understand this, consider a transistor in the vicinity of the drain, as shown in Fig. 5.32. In Fig. 5.32a, we show the

[†] In an attempt to reduce the weak-inversion current when the device is off, it is sometimes attempted to apply negative bias to the body with respect to the source in order to increase the threshold voltage. Unfortunately, this increases the drain-body voltage and can make band-to-band tunneling worse, as expected from the preceding discussion.

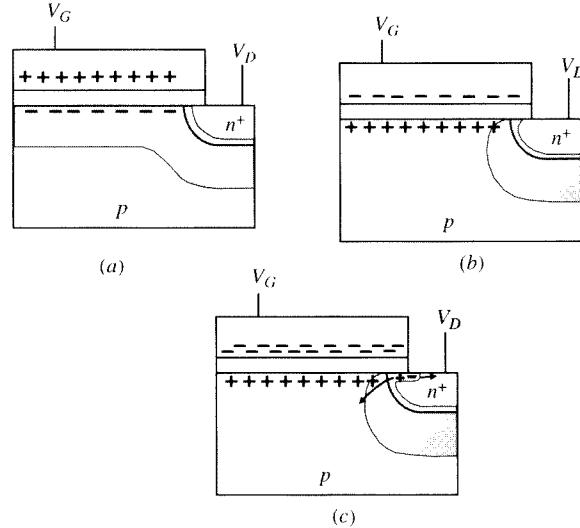
**FIGURE 5.31**

Energy band diagram for a *pn* junction, schematically indicating band-to-band tunneling.

situation when the gate voltage is large and the channel is inverted. Assuming nonsaturation, the potential of the inversion layer with respect to the bulk near the drain is similar to that from the n^+ drain region to the bulk. The depletion region width is then roughly uniform in the vicinity of the drain, as shown by the shaded area (a uniform substrate doping is assumed). Note that we show the extent of the depletion region in both the body and the n^+ drain region (the latter has not been shown explicitly in previous figures); because the doping is heavier in the drain region, the extent of the depletion region there is narrower. The lighter substrate doping results in a considerable depletion region extent into the *p*-type body. Consider now a low gate voltage, below the flatband voltage, e.g., when the device is off. This results in negative charges on the gate; these cause surface accumulation, as shown in Fig. 5.32*b*. Now the surface contains a very large number of holes, which make it act as if the *p*-type doping were locally very high. The depletion region width thus narrows as shown. Because practically the same voltage must now be dropped over a narrower depletion region, a large electric field results. It is also possible for the negative charges on the gate to repel the electrons at the surface of the n^+ region, resulting in local depletion there, as shown in Fig. 5.32*c*; large electric fields are now found in that region as well. (In fact, it is even possible for the n^+ region surface to invert).

The presence of the large electric fields in the cases shown in Fig. 5.32 causes band-to-band tunneling (discussed previously), thus generating electron-hole pairs, as shown schematically in Fig. 5.32*c*. The drain electric field causes the resulting electrons to move toward the drain, and the holes to move toward the bulk, as shown. These holes contribute to the bulk current. The preceding leakage mechanism²¹⁵ is referred to as *gate-induced drain leakage*, or GIDL (the corresponding acronym for leakage at the source, GISL, is not used much, because for many circuits, V_{SB} is zero and the preceding effect is predominantly seen in the drain current). Trap-assisted tunneling may enhance GIDL.²¹⁶ The enhanced leakage currents due to GIDL are experimentally observed to increase as gate bias decreases. The GIDL current obeys a law similar to that in (5.13.1):^{217,41,43}

$$I_{\text{GIDL}} = A_{\text{GIDL}} \mathcal{E}_{n,\max}^2 e^{-B_{\text{GIDL}}/\mathcal{E}_{n,\max}} \quad (5.13.2)$$

**FIGURE 5.32**

Drain-body junction in the cases of (a) substrate surface inversion, (b) substrate surface accumulation, and (c) substrate surface accumulation accompanied by drain region surface depletion.

where A_{GIDL} and B_{GIDL} are fitting parameters, with A_{GIDL} proportional to the channel width W , and $\mathcal{E}_{n,\max}$ is the maximum field in the GIDL region of the n^+ drain, which occurs just under the oxide in the gate-drain overlap region. As with computation of the impact ionization current in Sec. 5.8, there is no simple, first-principles analysis that leads to an analytically tractable, theoretically accurate model for GIDL. Rather, engineering approximations have to be made to arrive at a sufficiently accurate yet simple model. To this end, pseudo-two-dimensional analyses are typically used.

We now briefly discuss such an analysis,⁴³ to illustrate the level of approximations sometimes forced on the modeler in need of simple equations. Assuming first that the oxide field in the overlap region, $\mathcal{E}_{\text{ox,ov}}$, is purely transversal (perpendicular to the interface), we have (Appendix A)

$$\mathcal{E}_{\text{ox,ov}} = \frac{Q'_{\text{G,ox}}}{\epsilon_{\text{ox}}} = \frac{C'_{\text{ox}} \psi_{\text{ox,ov}}}{\epsilon_{\text{ox}}} \quad (5.13.3)$$

where the subscript “ov” refers to the gate-drain overlap region. To determine the oxide potential $\psi_{\text{ox,ov}}$, we can apply the potential balance equation, (2.3.1), for the *overlap* MOS structure consisting of the gate, the oxide, and the n^+ drain region. We have

$$V_{GD} = \psi_{\text{ox,ov}} + \psi_{s,ov} + \phi_{MS,ov} \quad (5.13.4)$$

Using (5.13.3) gives the field on the oxide side of the interface. To find the field \mathcal{E}_{ny} on the n^+ side, we note that there is a field discontinuity because there is a

permittivity change and also because of the presence of the effective interface charge, Q'_o . From Appendix A, we have

$$\epsilon_s \mathcal{E}_{ny} - \epsilon_{ox} \mathcal{E}_{ox,ov} = Q'_o \quad (5.13.5)$$

From the previous three equations, we obtain

$$\mathcal{E}_{ny} = \frac{C'_ox}{\epsilon_s} (V_{GD} - V_{FB,ov} - \psi_{s,ov}) \quad (5.13.6)$$

where $V_{FB,ov} = \phi_{MS,ov} - Q'_o/C'_ox$ is the flatband voltage of the overlap structure. This value has been obtained assuming that the field is exactly transversal, which is not the case; the situation is two-dimensional, and the voltage V_{DB} can be expected to cause a longitudinal field component. Assuming that the longitudinal component is proportional to V_{DB} , we can express it as

$$\mathcal{E}_{nx} = \frac{V_{DB}}{l_{eff}} \quad (5.13.7)$$

where $1/l_{eff}$ is a constant of proportionality to be determined empirically. We can now combine the transversal and longitudinal components of the field using the Pythagorean theorem:

$$\mathcal{E}_{n,max} = \sqrt{\mathcal{E}_{nx}^2 + \mathcal{E}_{ny}^2} \quad (5.13.8)$$

This can now be used in (5.13.2) to estimate the GIDL current. It goes without saying that a number of difficult-to-justify steps have been taken in this analysis, although most steps are at least intuitively reasonable. This type of approach is typical in modeling several phenomena that do not lend themselves to rigorous, simple analysis. It should be noted that the preceding only models the most basic mechanisms in operation. For example, it does not include trap-assisted tunneling and impact ionization, which have been reported to also contribute to GIDL.²¹⁶ Nevertheless, the values of the empirical parameters employed above can be chosen to improve accuracy in the presence of such effects. Other approaches are also used.¹⁹⁷

Figure 5.33 shows drain current without and with contributions from GIDL, based on simulations from models for a 90-nm technology. Although the additional leakage currents are of the order of nA, and occur for gate biases below strong inversion, there are situations where the GIDL currents are quite important; for example, where dynamic body biasing is used to minimize leakage (see preceding footnote).

5.14 Leakage Currents—Particular Cases

We have discussed several mechanisms that lead to leakage currents in modern MOS devices. To put everything together, we now provide a summary of the main leakage mechanisms for three particular cases, shown in Fig. 5.34, and meant as examples

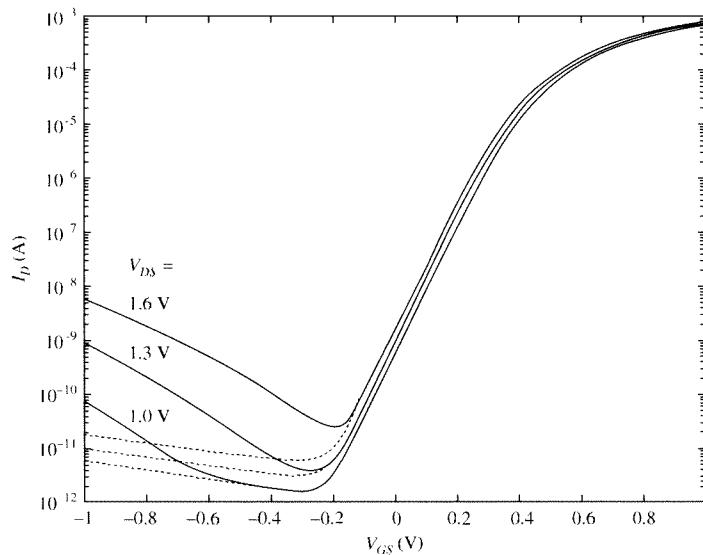
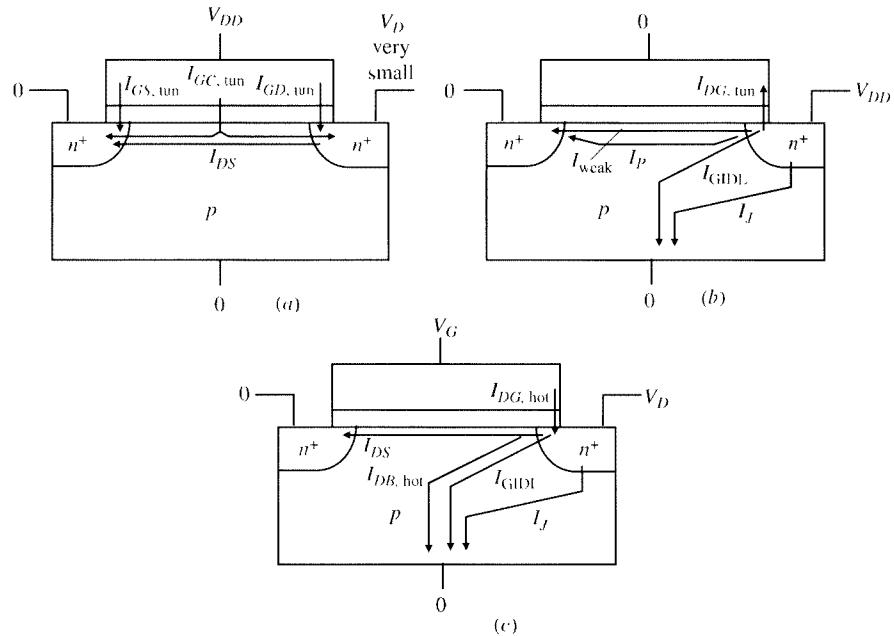


FIGURE 5.33
 $\log I_D$ vs. V_{GS} without GIDL (broken lines) and with GIDL (solid lines). $t_{ox} = 2.6$ nm, $N_A = 6 \times 10^{17}$ cm $^{-3}$, $W = 2$ μm , $L = 0.1$ μm .

of commonly encountered situations in circuits. In all cases, the source and body are assumed grounded. The mechanisms and quantitative description of the currents in the following discussion have been discussed in Secs. 5.8, 5.12, and 5.13.

Figure 5.34a shows a transistor in digital on operation; the gate voltage is large, assumed here to be equal to the power supply voltage, V_{DD} . A drain-source current I_{DS} flows. The device is assumed to be part of a CMOS inverter, in which case its drain voltage will be very low and is assumed here to be practically zero for simplicity. It can be seen that a large positive voltage exists across the entire oxide. Thus, tunneling current flows throughout, consisting of a gate-channel component (which has source and drain components), a gate-source overlap component, and a gate-drain overlap component, in the directions shown.

Figure 5.34b shows the transistor in digital off operation, with zero voltage applied at the gate; the drain is voltage is high, and is assumed to be V_{DD} for simplicity. A channel current I_{weak} flows due to weak inversion, as indicated; *this current is often considered a leakage current in digital work*, just like any current that should ideally be zero in the off condition, but is not. The weak-inversion current is affected by the small-dimension effects we have discussed earlier in this chapter, notably by DIBL (Sec. 5.5). A punchthrough current, I_P , may flow in parallel to the channel current, if punchthrough effects are present; this path can be along the surface, or it can be a subsurface current, depending on transistor implantation details. A junction current I_J flows due to the reverse-bias voltage between the drain and body. A GIDL current I_{GIDL} flows between drain and body. Finally, a tunneling current $I_{DG,tun}$ flows between the drain and the gate, in the overlap region, as shown. All currents shown are positive in the direction indicated.

**FIGURE 5.34**

Leakage currents in the cases of (a) digital ON operation, (b) digital OFF operation, and (c) analog operation.

Figure 5.34c shows a transistor in analog operation. Voltage V_G is assumed large enough to cause operation in strong inversion, and voltage V_D is large enough to allow operation in saturation. A large channel current, I_{DS} , flows. A junction current I_J and a GIDL current I_{GIDL} flow as in the previous case, although their values will be different due to the different bias condition. Finally, in the pinchoff region, hot electron effects cause a drain-body current $I_{DB,hot}$ and possibly a drain-gate current $I_{DG,hot}$.

5.15 The Quest for Ever-Smaller Devices

5.15.1 Introduction

One of the hallmarks of the semiconductor industry over many years has been the continuous scaling down of circuit dimensions with the introduction of every new technology generation.^{218–221,223–229} This has led to a continuous increase in speed and circuit complexity per unit of chip area, and perhaps even more importantly, to a continuous decrease in cost per transistor and cost per circuit function. The process of reducing certain key dimensions, from one technology generation to the next, is referred to as *scaling*.

Until about the year 2000, scaling could be done in a rather systematic, co-ordinated manner, according to well-established scenarios. Today, the scaling down

of certain dimensions has almost reached its limit, for example, scaling down oxide thickness greatly increases dc gate leakage (Sec. 5.12). Thus, the scenarios of the past can no longer be applied, and now technology generations are designed on a case-by-case basis. Nevertheless, it is interesting to take a look at the scaling scenarios that carried us through several technology generations in the past; these go by the name of *classical scaling*. A look at classical scaling is in order because it conveys a general way of thinking, whose value can be carried beyond classical scaling. Indeed, parts of the considerations involved in classical scaling are still useful today, although they must be supplemented with additional considerations, such as leakage current through the oxide. Classical scaling is described in Sec. 5.15.2. Modern scaling is described in Sec. 5.15.3.

5.15.2 Classical Scaling

In this section, we will assume that scaling can be performed without encountering the problem of oxide leakage; this assumption will be removed in Sec. 5.15.3. We will also assume that the substrate doping is uniform, unless stated otherwise.

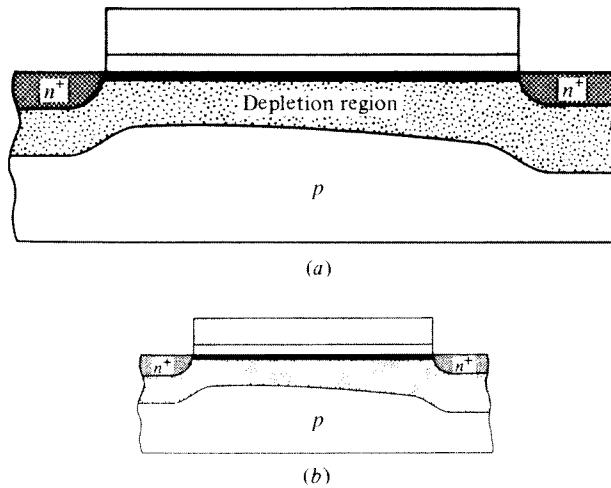
We will first describe a set of classical scaling rules aimed at reducing size in such a way that the resulting device, along with its depletion region, is simply a scaled version of a large device, as is shown in Fig. 5.35. Then no significant side effects should appear, and one can analyze the scaled device by using well-known large-device concepts, thus taking advantage of the significant experience gained with large devices over the years. A process proposed to achieve the reduction shown in Fig. 5.35b will now be briefly described.²¹⁹ It will be seen to result in a “scaled” device, in which internal electric field shape and maximum magnitude is the same as in the original device; hence the name *constant-field scaling* is used to describe this process.

Let us assume that a large device is scaled in all three dimensions by a factor $1/\kappa$, where κ is larger than 1 (say, between 1 and 3). This means that L , W , oxide thickness, and junction depth are all scaled by the same factor. Thus, areas (e.g., gate area) are scaled by $(1/\kappa)^2$. The depletion region underneath a junction has a depth d given by [see (1.5.16) with (1.5.3)]

$$d = \sqrt{\frac{2\epsilon_s}{qN_A}} \sqrt{\phi_{bi} + V} \quad (5.15.1)$$

where V is the reverse bias and ϕ_{bi} the built-in potential. The form of this equation is also valid for the depletion region underneath the inversion layer; for example, in strong inversion, the equation is valid with $\phi_{bi} = \phi_0$. We will assume ϕ_{bi} is small compared with V .[†] Then we can scale d by $1/\kappa$, by scaling $\sqrt{V/N_A}$ by $1/\kappa$ in (5.15.1). Let us scale N_A by κ and V by $1/\kappa$. To achieve a properly scaled V throughout

[†] This assumption is not valid for all bias conditions; it has historically been made for simplicity. Thus, a device scaled as described will deviate from ideal constant field scaling behavior when the reverse bias is small.

**FIGURE 5.35**

(a) A MOS transistor; (b) a MOS transistor obtained from (a) by "classical" scaling.

the channel, all operating voltages as well as the threshold voltage must be scaled by $1/\kappa$. From basic electrostatics (Appendix A), one can see that under this scaling the shape and maximum magnitude of the electric field in the structure will remain the same. Thus undesirable high field effects will not occur. As seen in (1.5.25) and (2.2.4), capacitances *per unit area* C' are inversely proportional to distances, so they are scaled by κ . However, capacitances C *per se* are given by $C'A$, where A is the area, and thus are seen to scale by $\kappa(1/\kappa^2) = 1/\kappa$.

The body effect coefficient γ is seen from (2.4.26a) to scale by $1/\sqrt{\kappa}$. In device equations, the threshold voltage V_T appears in differences with bias voltages, such as in (4.7.24). Since the latter are scaled by $1/\kappa$, V_T should also be scaled by $1/\kappa$ as already mentioned. Consider (4.7.19). From the preceding discussion, if ϕ_0 is very small compared with V_{SB} , the term $\gamma\sqrt{\phi_0 + V_{FB}}$ will scale by $1/\kappa$. To scale $V_{FB} + \phi_0$ in that equation, one should be able to control V_{FB} , which cannot be done independently for unimplanted devices. However, it will be seen in Chap. 9 that an "equivalent" V_{FB} can be controlled through ion implantation.

Charges *per unit area* Q' are not scaled [see (1.5.18) or (3.4.8a) and use the preceding results]. Therefore, charges Q are scaled by $1/\kappa^2$, since areas are scaled by this factor.

What is now the effect of the preceding scaling procedure on the drain current? Let us look at (4.7.24). If α is close to 1, it is seen that since all voltages are scaled by $1/\kappa$, the quantity in brackets is scaled by $1/\kappa^2$. At the same time C'_{ox} is scaled by κ . Thus, assuming that μ is practically independent of doping concentration (Sec. 4.11) and thus does not change under constant-field scaling, the current in (4.7.24) is scaled by $1/\kappa$.

Consider now the slope of $\ln I_{DS}$ vs. V_{GS} in weak inversion for a constant V_{DS} . As seen from (4.8.11), that slope is proportional to $1/n$, with n given by (4.8.14). Since

γ scales by $1/\sqrt{\kappa}$ and $V_{SB} + 2\phi_F$ by $1/\kappa$ (assuming V_{SB} is large), n and S in Sec. 4.8 remain the same, and the slope of $\ln I_{DS}$ vs. V_{GS} does not scale. For digital circuits, this is undesirable, since it makes it more difficult to turn a device off. For example, it takes as much reduction in gate voltage to reduce I_{DS} by one decade as it does for large devices. Since the total voltage swings possible have been scaled by $1/\kappa$, the gate-voltage reduction needed to turn the device off represents a larger fraction of the total swing, and noise margins in digital circuits are reduced. The width of the moderate-inversion region²³⁰ does not scale either. Thus this region now becomes a larger part of the power supply voltage, and more attention must be paid to it.^{†230, 231}

Since both voltages and currents are scaled by $1/\kappa$, power dissipation is scaled by $1/\kappa^2$. However, since device areas have been scaled by $1/\kappa^2$, the density of devices per unit area is scaled by κ^2 ; thus, the power per unit of chip area is not scaled.

Since all device currents are scaled by $1/\kappa$ and the various capacitances are scaled by the same factor, the rate of change of charging these capacitances, $dV/dt = I/C$, will not scale. However, these capacitances now only need to be charged to voltages scaled down by $1/\kappa$, and thus the time needed to charge them scales by $1/\kappa$; hence, the speed of digital circuits increases by κ . Since the power dissipation for a transistor has scaled by $1/\kappa^2$, the “power delay product” (a figure of merit used for digital circuits) scales by $1/\kappa^3$. Constant-field scaling is summarized in Table 5.1.

Consider now the metal and polysilicon lines used to form the gates and interconnections. Since we now have a fabrication process capable of small dimensions, let us attempt to scale the width of these lines by $1/\kappa$. The new process may also require reducing the height of these lines, since trying to make very thin, but tall, lines can run into fabrication problems; let us attempt to scale that height by $1/\kappa$ also. Then the cross-sectional area of the lines is scaled by $1/\kappa^2$. Since the current that these lines carry has been seen to scale by $1/\kappa$, the current *density* in these lines will scale by κ . This is very undesirable, since the increased current density can cause what is known as electromigration, a phenomenon in which atoms are carried by the flow of current and can result in line failure. For aluminum lines, the current density should not be larger than about $1 \text{ mA}/\mu\text{m}^2$. Another problem with scaling interconnection lines is that the resistance of the lines is proportional to length and inversely proportional to the cross-sectional area and thus scales by κ . The parasitic capacitances of these lines to the substrate scale by $1/\kappa$, and thus the corresponding time constant does not scale. If the lines are long, this can cause a problem, since it prevents us from taking advantage of the fact that the speed of transistors has scaled by κ . Also, since the resistance of these lines scales by κ and the current through them by $1/\kappa$, the voltage drop across them does not scale. Thus, a larger fraction of the total voltages available, which have been scaled by $1/\kappa$, is now wasted across interconnect lines. Because of the preceding problems, the height of interconnect lines is reduced less significantly.

Additional problems are caused by the “contact windows,” etched through the thick oxide in order to make contacts between various layers. If the area of these

[†] In moderate inversion, the slopes of the inversion layer charge (Fig. 2.17) and of I_{DS} for low V_{DS} (Fig. 4.33) vs. gate voltage are found to decrease. This has been attributed to the finite-inversion-layer capacitance in this region²³⁰ [see (2.7.23)], and for this reason, sometimes moderate-inversion effects are referred to as “finite-inversion-layer capacitance effects.”

TABLE 5.1
Constant-field scaling

Quantity	Scaling factor
Device dimensions (L , W , t_{ox} , d_j)	$1/\kappa$
Area	$1/\kappa^2$
Packing density (devices per unit of chip area)	κ^2
Doping concentration, N_A	κ
Bias voltages and V_T	$1/\kappa$
Bias currents	$1/\kappa$
Power dissipation for a given circuit	$1/\kappa^2$
Power dissipation per unit of chip area	1
Capacitances, C	$1/\kappa$
Capacitances per unit area, C'	κ
Charges, Q	$1/\kappa^2$
Charges per unit area, Q'	1
Electric field intensity	1
Body effect coefficient, γ	$1/\sqrt{\kappa}$
Transistor transit time, τ	$1/\kappa$
Transistor power-delay product	$1/\kappa^3$

windows is scaled by $1/\kappa^2$, their resistances will scale by κ^2 . For currents scaled by $1/\kappa$, this means that the voltage drop across the contacts will scale by κ , i.e., in the *opposite* direction from the bias voltages, which were scaled by $1/\kappa$. Another undesirable effect of scaling is an increase in the resistance of the source and drain n^+ regions due to the decreased junction depth. The effect of the source and drain resistances on the transistor characteristics has been considered in Sec. 4.12.

The scaling scenario just discussed presents certain problems. As already discussed, the weak-inversion region width does not scale. Hence, the voltage swings required for turning the device from off to on may be an unacceptably large fraction of the total voltage available. In addition, established chip interface requirements must often be obeyed for which voltage levels are fixed, and, hence, the voltage cannot be scaled. The rules followed in such cases, for decreasing device dimensions while keeping the voltages unchanged, have classically been referred to as *constant-voltage scaling*. In these, W , L , and N_A are scaled as before. However, if the oxide thickness is scaled by the same factor, the resulting field can be exceedingly high since voltages are not scaled; this can cause oxide leakage and mobility degradation. To alleviate this problem somewhat, oxide thickness is usually scaled less drastically.²²⁹ One column in Table 5.2 summarizes these scaling rules. The resulting effect on the various quantities (such as those entered in Table 5.1) is considered in Prob. 5.33. It is to be expected that undesirable effects associated with high electric fields can become severe under constant-voltage scaling.

To avoid the extreme cases of constant-field and constant-voltage scaling, compromise scaling rules have been proposed. For example, geometric dimensions and substrate doping are scaled as in the case of constant-field scaling, but voltages are scaled less drastically. This has been termed *quasi-constant-voltage scaling*²³² and is summarized in the corresponding column of Table 5.2. As is evident from (5.15.1),

TABLE 5.2
Scaling rules

Quantity	Scaling factor			
	Constant-field scaling	Constant voltage scaling $1 < \kappa' < \kappa$	Quasi-constant-voltage scaling $1 < \kappa' < \kappa$	Generalized scaling $1 < \kappa' < \kappa$
W, L	$1/\kappa$	$1/\kappa$	$1/\kappa$	$1/\kappa$
t_{ox}	$1/\kappa$	$1/\kappa'$	$1/\kappa$	$1/\kappa$
N_A	κ	κ	κ	κ^2/κ'
V, V_T	$1/\kappa$	1	$1/\kappa'$	$1/\kappa'$

under this type of scaling the depletion region widths do not scale by the same factor as W , L , and t_{ox} . This is avoided if the scaling factor for N_A is modified appropriately, resulting in the *generalized scaling* rules shown in the last column of Table 5.2.²²⁵

5.15.3 Modern Scaling

By Dimitri Antoniadis, MIT

As has already been mentioned, the classical scaling approaches have reached their limits.^{231a-g} Developing the next generation technology is now a very detailed, custom process, as numerous very-small-dimension effects must be considered. Notable among these is dc gate current, discussed in Sec. 5.12. In this section, we consider state-of-the-art CMOS technologies at the time of writing of this book, and also speculate on the future of scaling. Our emphasis is on logic circuits, which drive technology development.

There are three types of logic families:²³³ (1) the so-called high-performance (HP) CMOS, which encompasses technologies used for the fastest electronic products such as servers, workstations, and some desktop computers; (2) the low operating power (LOP) CMOS family, which strives to maintain a balance between performance and power dissipation, with notebook computers being one of the main target applications; and (3) low standby power (LSTP) CMOS, which finds application in portable handheld devices such as cell phones and personal digital assistants. There are several differences between these families, with HP MOSFETs having the most aggressively scaled dimensions and, in digital operation, also the highest off-state current, I_{off} , i.e., current at $V_{GS} = 0$ V and $V_{DS} = V_{DD}$, where V_{DD} is the system power supply voltage.

Although the considerations involved in classical scaling are still important, other technological aspects have been driving modern device scaling. The balance between the needs for low I_{off} , determined by the threshold voltage, and high I_{on} (on-state drive current in digital operation, determined by the supply voltage for a given device size) have kept threshold voltage and supply voltage roughly constant. Fine adjustment of the threshold is a key to balancing the speed-power trade-offs between the different types of technology families. Other areas that are being pursued in modern scaling are doping profile engineering (e.g., the introduction of halo implants, discussed in Secs. 5.4.2 and 9.4); the use of new materials, such as hafnium-based gate dielectrics rather than silicon dioxide, and metal rather than polysilicon gates; the use

of strain engineering to increase mobility (Sec. 9.6) and thus current drive capability; and the investigation of novel device structures. We will highlight all those issues below by reviewing briefly the scaling history of the last few HP CMOS technology generations and by discussing prospective device structures for further future scaling.

The basic tenet of the famous Moore's law is that circuit area density (number of devices per unit area) scales by a factor of 2 each 18 to 24 months. For this area density scaling, linear dimensions must scale by a factor of roughly $\sqrt{2}$. The key feature dimension that must scale in CMOS is the so-called *minimum contacted gate pitch*, defined as the gate center-to-center distance in an array of parallel minimum-length gates, with a source/drain region in between them, containing a minimum-sized metal contact. In modern CMOS technology, the minimum contacted gate pitch is typically about 3.6 times larger than the numeral name of a given technology (which, as mentioned in Sec. 1.7, has historically been derived from certain minimum features in corresponding DRAM technology); for example, it is of the order of 160 nm for the 45-nm generation depending on the manufacturer. In the years 2008–2009, the so-called 65-nm CMOS generation was in full production, the 45-nm generation production was ramping up, and the first 32-nm generations were disclosed in technical conferences.

Close examination of the last three generations in production, i.e., 90-, 65-, and 45-nm, shows that the gate oxide thickness has been almost constant, consisting of a heavily nitrided silicon dioxide (SiON) material with about 1.2-nm equivalent-SiO₂ thickness. This is the minimum thickness that can be tolerated in terms of gate current leakage (Sec. 5.12). Consequently, the minimum gate length has scaled slowly, from about 45 to 50 nm in the 90-nm generation to 35 to 40 nm in the 45-nm generation, and this has been primarily achieved with channel doping engineering based on strong halo doping around the source and drain junctions. Similarly, the power supply voltage has not scaled much either (from about 1.2 V to 1.1 V correspondingly) in order to enable some circuit performance increase with each succeeding generation while maintaining near-constant V_T and hence near-constant I_{off} .

Because the traditional scaling pathway (reducing oxide thickness and channel length) to improve device performance could no longer be followed, with the 90-nm technology generation, channel strain was introduced. Because of the piezoresistive effect in silicon, tensile strain in *n*FETs and compressive strain in *p*FETs (along the channel direction) has increased carrier mobility significantly. Then with the 45-nm technology, at least one manufacturer introduced a composite gate dielectric consisting of hafnium silicate and SiO₂, with a dielectric constant significantly higher than that of SiON ($k \approx 15\text{--}20$ as compared with 4.2, respectively.) Increasing the dielectric constant allows for increased capacitance with a thicker dielectric and thus makes possible low gate leakage. The gate channel capacitance is further increased via the use of metal instead of polysilicon as the gate material, hence eliminating the polysilicon-gate depletion effect (Sec. 5.10). It is expected that so-called high- k /metal-gate stacks will find wider acceptance with the 32-nm generation and beyond.

The basic structure of the MOSFET in all generations to date has remained unchanged. The channel is in the plane of the silicon wafer, and whether the wafer is all silicon (so-called bulk silicon) or consists of a silicon film on insulator, which in turn is on a bulk-silicon wafer (so-called SOI), the MOSFET can be analyzed using the classical approach presented in this book, save for certain peculiarities associated with a

floating body. The reason why current SOI devices in production are still considered “classical structures” is that the SOI channel film thickness, t_s , is larger than the channel depletion depth, d_B . These so-called partially depleted SOI MOSFETs are scaled like bulk-silicon devices, using the approaches that have already been described.

Because of the difficulty in scaling down the effective oxide thickness, and despite the advent of high- k dielectrics, it is expected that at some point in the continuing evolution of CMOS technologies, so-called nonclassical MOSFET device structures, with better scalability than the classical device structures, will have to be adopted. The exact generation where this is likely to happen is not certain at present, but it is unlikely that such structures will be adopted before the 22-nm generation. These nonclassical device structures include the (planar) fully depleted SOI (FDSOI) structure, where $t_s < d_B$, and a variety of nonplanar structures, of which the so-called FinFET structure^{234,235} is a typical representative (Fig. 5.36a). Although still in experimental and possibly early development phases, the FinFET is a practical implementation of the double-gate MOSFET, in which the channel body is very thin, and therefore fully depleted, and there are two gates at the same voltage symmetrically controlling the inversion charge. In some cases, when the channel thickness is small, the two inversion layers may merge into one indistinguishable layer centered in the channel body in what is often referred to as “volume-inversion.”²³⁶ Whereas in

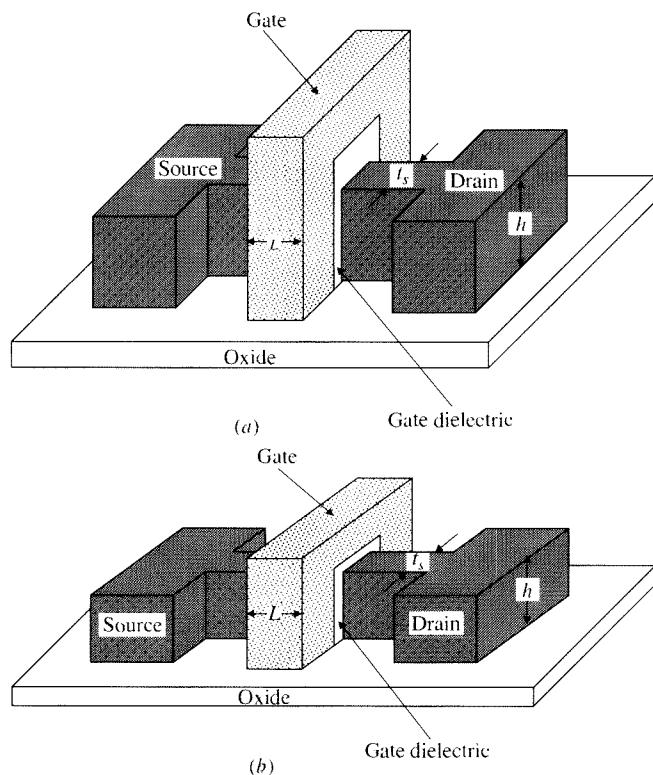


FIGURE 5.36

Simplified illustrations of (a) the FinFET (symmetrical double-gate) MOSFET structure and (b) the tri-gate structure.

principle, the two gates could be independently biased, the most common FinFET structure has them tied together, as shown in the figure. It can be appreciated from this discussion that the gates have tighter electrostatic coupling to the channel than in a planar single-gate structure. The height-to-thickness ratio, h/t_s , of the typical FinFET would be on the order of 2 to 4, and only the two vertical surfaces can be inverted because the insulator on the top surface is kept thick. Note that the inversion charge sheet in this structure lies on the two vertical planes under the gate. Therefore, the effective width for the FinFET is $W = 2h$.

Another type of nonplanar structure is the “trigate” structure²³⁷ in which the height-to-thickness ratio is about 1, and all surfaces facing the gate have thin oxide and can be inverted (Fig. 5.36b). Because the gate wraps the channel on three sides, the electrostatic coupling is even tighter than that of the FinFET. The ultimate tight coupling is achieved in yet another structure called “gate-all-around (GAA)” or “nanowire channel” where the fin in Figure 5.36b is replaced by a cylindrical semiconductor channel that is fully wrapped by the gate.²³⁸

The characteristic of all these nonclassical device architectures is that the channel structure is used to improve scalability by promoting enhanced electrostatic coupling between gate and channel. This is a result purely of the laws of electrostatics and can be understood by using the concepts introduced in Sec. 5.5 and, more specifically, the results of the pseudo-two-dimensional analysis represented in (5.5.1), where the short-channel-induced V_T reduction is modeled in terms of the characteristic length or “gauge,” λ , given by (5.5.2). The meaning of the gauge²¹³ is that a condition on the channel length, $L > m\lambda$, should be met to guarantee a minimum level of electrostatic integrity, defined as a maximum acceptable limit for threshold voltage reduction due to DIBL and charge sharing. A minimum value of m is about 4.²³⁹ Equation (5.5.2) has been adapted for FDSOI, FinFET, and GAA structures.^{168,169,200,201,206,209–217,233–241} An example of minimum channel lengths, calculated based on such equations for an equivalent oxide thickness of 1.0 nm and assuming $m = 4$, is shown in Figure 5.37. This plot is just a rough approximation; for more accuracy in

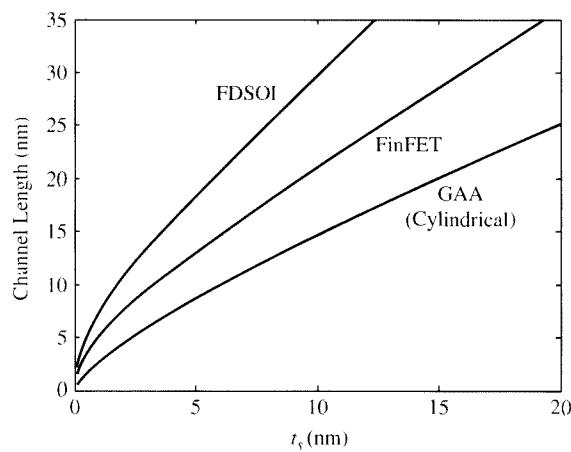


FIGURE 5.37
Minimum channel length vs. channel thickness, t_s , for gate oxide thickness $t_{ox} = 1.0$ nm, assuming $m = 4$, for fully depleted SOI (FDSOI), FinFet, and GAA cylindrical MOSFET structures.

understanding the effects of device geometry on scalability, proper two- and three-dimensional numerical analysis must be performed, e.g., as in Ref. 242.

Given these electrostatic scaling considerations, with foreseeable technological developments over the next several years, it could be possible to realize MOSFETs with channel lengths down to about 5 nm. However, such channel lengths would require a nanowire channel with diameter around 2–3 nm! At such nanoscale dimensions, many other physical considerations arise from the finite extent of the semiconductor lattice; e.g., such a structure involves only 10 monolayers across the diameter of the cylinder and would be subject to quantum mechanical carrier confinement. Although qualitatively the device behavior should be similar to that of a classical MOSFET, it is difficult at this point to predict its quantitative behavior. This area of nanoscale electronic device fabrication and operation is under intense investigation at the time of writing of this book.

References

1. D. Vandorpe, J. Borel, G. Merckel, and P. Saintot, "An accurate two-dimensional numerical analysis of the MOS transistor," *Solid-State Electronics*, vol. 15, p. 547, 1972.
2. D. P. Kennedy and P. C. Murley, *IBM Journal of Research and Development*, vol. 17, p. 2, 1973.
3. M. S. Mock, "A two-dimensional mathematical model of the insulated gate field-effect transistor," *Solid-State Electronics*, vol. 16, pp. 601–609, 1973.
4. R. Kasai and T. Kimura, "Two-dimensional structure analysis of short-channel C-MOS and transient analysis of the C-MOS circuits," Technical Group, Institute of Electronics and Communications Engineers of Japan, ED76–5, 1976 (in Japanese).
5. J. A. Greenfield and R. W. Dutton, "Nonplanar VLSI device analysis using the solution of Poisson's equation," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 1520–1532, 1980.
6. S. Selberherr, A. Schütz, and H. W. Pötzl, "MINIMOS—A two-dimensional MOS transistor analyzer," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 1540–1549, 1980.
7. B. T. Brown and J. J. H. Miller (editors), *Numerical Analysis of Semiconductor Devices*, Boole Press, Dublin, 1970.
8. E. M. Buturla, P. E. Cottrel, B. M. Grossman, K. A. Salzburg, M. B. Lawlor, and C. T. McMullen, "Three-dimensional finite element simulation of semiconductor devices," *Proceedings of the IEEE International Solid-State Circuits Conference*, San Francisco, 1980, pp. 76–77.
9. A. Yoshii, H. Kitazawa, M. Tomizawa, S. Horiguchi, and T. Sudo, "A three-dimensional analysis of semiconductor devices," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 184–189, 1982.
10. A. Husain and S. G. Chamberlain, "Three-dimensional simulation of VLSI MOSFET's: The three-dimensional simulation program WATMOS," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 631–638, 1982.
11. W. L. Engl, H. K. Dirks, and B. Meinerzhagen, "Device modeling," *Proceedings of the IEEE*, vol. 71, pp. 10–33, 1983.
12. M. R. Pinto, C. S. Rafferty, and R. W. Dutton, "PISCES-II: Poisson and continuity equation solver," Stanford Electronics Laboratory Technical Report, September 1984.
13. C. L. Wilson, P. Roitman, and L. Blue, "High accuracy physical modeling of submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 1246–1258, 1985.
14. C. L. Wilson and J. L. Blue, "Accurate current calculations in two-dimensional MOSFET models," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2060–2068, 1985.
15. B. Meinerzhagen, H. K. Dirks, and W. L. Engl, "Quasi-simultaneous solution method: A new highly efficient strategy for numerical MOST simulations," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2131–2138, 1985.
16. J. G. Ruch, "Electron dynamics in short-channel field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 652–654, 1972.

17. T. Kobayashi and K. Saito, "Two-dimensional analysis of velocity overshoot effects in ultrashort-channel Si MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 788–792, 1985.
18. G. G. Shahidi, D. A. Antoniadis, and H. I. Smith, "Electron velocity overshoot at room and liquid nitrogen temperatures in silicon inversion layers," *IEEE Electronic Device Letters*, vol. EDL-9, pp. 94–96, 1988.
19. K. K. Thornber, "Relation of drift velocity to low-field mobility and high-field saturation velocity," *Journal of Applied Physics*, vol. 51, pp. 2127–2133, 1980.
20. S. R. Hofstein and G. Warfield, "Carrier mobility and current saturation in the MOS transistor," *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 129–138, 1965.
21. G. Baum and H. Beneking, "Drift velocity saturation in MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-17, pp. 481–482, 1970.
22. G. Baum, "Driftgeschwindigkeitssättigung bei MOS-Feldeffekttransistoren," *Solid-State Electronics*, vol. 13, pp. 789–798, 1970.
23. G. Merckel, J. Borel, and N. Z. Cupcea, "An accurate large-signal MOS transistor model for use in computer-aided design," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 681–690, 1972.
24. H. C. Poon, " V_{th} and beyond," presented at the Workshop on Device Modelling for VLSI, Burlingame, Calif., March 29, 1979; also L. Cong, Bell Laboratories, private communication.
25. M. El Nokali and H. Miranda, "A simple model for the MOS transistor in saturation," *Solid-State Electronics*, vol. 29, pp. 591–596, 1986.
26. F. M. Klaassen and W. C. J. de Groot, "Modeling of scaled-down MOS transistors," *Solid-State Electronics*, vol. 23, pp. 237–242, 1980.
27. T. Poorer and J. H. Satter, "A DC model for an MOS transistor in the saturation region," *Solid-State Electronics*, vol. 23, pp. 765–772, 1980.
28. J. R. Brews, "Physics of the MOS transistor," chap. I, *Silicon Integrated Circuits, Part A*, D. Kahng (editor), Applied Solid-State Science Series, Academic Press, New York, 1981.
29. S. Liu and L. W. Nagel, "Small-signal MOSFET models for analog circuit design," *IEEE Journal of Solid-State Circuits*, vol. SC-17, pp. 983–998, 1982.
30. F. N. Trofimenkoff, "Field-dependent mobility analysis to the field-effect transistor," *Proceedings of the IEEE*, vol. 53, pp. 1765–1766, 1965.
31. R. H. Crawford, *MOSFET in Circuit Design*, McGraw-Hill, New York, 1967.
32. B. Hoeneisen and C. A. Mead, "Current-voltage characteristics of small size MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 382–383, 1972.
33. P. Smith, M. Inoue, and J. Frey, "Electron velocity in Si and GaAs at very high electric fields," *Applied Physics Letters*, vol. 37, pp. 797–798, 1980.
34. M. H. White, F. Van de Wiele, and J. P. Lambot, "High-accuracy MOS models for computer-aided design," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 899–906, 1980.
35. C. G. Sodini, P.-K. Ko, and J. L. Moll, "The effect of high fields on MOS device and circuit performance," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1386–1393, 1984.
36. R. Coen and R. S. Muller, "Velocity of surface carriers in inversion layers of silicon," *Solid-State Electronics*, vol. 23, pp. 35–40, 1980.
37. J. R. Hauser, "A new and improved physics-based model for MOS transistors," *IEEE Transactions on Electron Devices*, vol. 52, no. 12, pp. 2640–2647, 2005.
38. D. M. Caughey and R. E. Thomas, "Carrier mobilities in silicon empirically related to doping and field," *Proceedings of the IEEE*, vol. 55, pp. 2192–1766, 1967.
39. D. L. Sharfetter and H. K. Gummel, "Large-signal analysis of a silicon Read diode oscillator," *IEEE Transactions on Electron Devices*, vol. ED-16, no. 1, pp. 64–77, 1969.
40. K. Joardar, K. K. Gullapalli, C. C. McAndrew, M. E. Burnham, and A. Wild, "An improved MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 45, no. 1, pp. 134–148, 1998.
41. R. van Langevelde, A. J. Scholten, and D. B. M. Klaassen, "Physical background of MOS model 11," Philips Nat. Lab. Unclassified Report NL-UR 2003/00239, 2003 [Online]. Available at http://www.nxp.com/models/mos_models/model11/index.html. Accessed Sept. 29, 2009.
42. G. Gildeblat, X. Li, H. Wang, W. Wu, A. Jha, R. van Langevelde, A. J. Scholten, G. D. J. Smit, and D. B. M. Klaassen, "Theory and modeling techniques used in the PSP model," *Technical Proceedings of the Nanotechnology Conference*, pp. 604–609, 2006.

43. G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, 2006.
44. R. van Langevelde and F. M. Klaassen, "Accurate drain conductance modeling for distortion analysis in MOSFETs," *Proceedings of the International Electron Devices Meeting*, pp. 313–316, 1997.
45. T. Sakurai and A. R. Newton, "A simple MOSFET model for circuit analysis," *IEEE Transactions on Electron Devices*, vol. 38, p. 887, 1991.
- 45a. P. Ratnam and C. A. T. Salama, "A new approach to the modeling of nonuniformly doped short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1289–1298, 1984.
- 45b. F. J. Lai and J. Y. Sun, "Analytical one-dimensional model for lightly doped drain (LDD) MOSFET devices," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2803–2811, 1985.
- 45c. G. S. Huang and C. Y. Wu, "An analytic I-V model for lightly doped drain (LDD) MOSFET devices," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 1311–1321, 1987.
- 45d. K. Y. Toh, P. K. Ko, and R. G. Meyer, "An engineering model for short-channel MOS devices," *IEEE Journal of Solid-State Circuits*, vol. SC-23, pp. 950–958, 1988.
- 45e. Y. Hu, R. V. H. Booth, and M. H. White, "An analytical model for the lateral channel electric field in LDD structures," *IEEE Transactions on Electron Devices*, vol. ED-37, pp. 2254–2263, 1990.
46. S. R. Hofstein and F. P. Heinman, "The silicon insulated-gate field-effect transistor," *Proceedings of the IEEE*, vol. 51, pp. 1190–1202, 1963.
47. C. Goldberg, "Pinch off in insulated-gate field-effect transistors," *Proceedings of the IEEE*, vol. 52, pp. 414–415, 1964.
48. V. K. G. Reddi and C. T. Sah, "Source to drain resistance beyond pinch-off in metal-oxide semiconductor transistors (MOS)," *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 139–141, 1965.
49. J. E. Schroeder and R. S. Muller, "IGFET analysis through numerical solution of Poisson's equation," *IEEE Transactions on Electron Devices*, vol. ED-15, pp. 954–961, 1968.
50. H. Shichman and D. A. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-3, pp. 285–289, 1968.
51. D. Frohman-Bentchkowsky and A. S. Grove, "Conductance of MOS transistors in saturation," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 108–113, 1969.
52. R. S. C. Cobbold, *Theory and Applications of Field-Effect Transistors*, Wiley-Interscience, New York, 1970.
53. G. A. Armstrong and J. A. Magowan, "The distribution of mobile carriers in the pinch-off region of an insulated-gate field-effect transistor and its influence on device breakdown," *Solid-State Electronics*, vol. 14, pp. 723–733, 1971.
54. D. Vandorpe, J. Borel, G. Merckel, and P. Saintot, "An accurate two-dimensional numerical analysis of the MOS transistor," *Solid-State Electronics*, vol. 15, pp. 547–557, 1972.
55. A. Popa, "An injection level dependent theory of the MOS transistor in saturation," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 774–781, 1972.
56. P. Rossel, H. Martinot, and G. Vassilieff, "Accurate two-sections model for MOS transistor in saturation," *Solid-State Electronics*, vol. 19, pp. 51–56, 1976.
57. Y. A. El-Mansy and A. R. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Transactions on Electron Devices*, vol. ED-24, pp. 254–262, 1977.
58. G. Merckel, "CAD models of MOSFETs," *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
59. F. M. Klaassen, "Review of physical needs for MOS transistors," *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
60. B. Hofflinger, H. Sibbert, and G. Zimmer, "Model and performance of hot-electron MOS transistors for VLSI," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 513–520, 1979.
61. M. E. Banna and M. El-Nokali, "A pseudo-two-dimensional analysis of short-channel MOSFETs," *Solid-State Electronics*, vol. 31, pp. 269–274, 1988.
62. P. K. Ko, "Approaches to scaling," pp. 1–37, in *Advanced MOS Device Physics*, N. G. Einspruch and G. Gildenblat (editors), VLSI Electronics, vol. 18, Academic Press, New York, 1989.

63. N. D. Arora, *MOSFET Models for VLSI Circuit Simulation—Theory and Practice*, Computational Microelectronics Series, S. Selberherr (editor), Springer-Verlag, Vienna and New York, 1993.
64. H. C. de Graaff and F. M. Klaassen, "Compact transistor modeling for circuit design," Springer-Verlag, Vienna and New York, 1990.
65. F. M. Klaassen and R. M. D. Velghe, "Compact modeling of the MOSFET drain conductance," *Proceedings of the European Device Research Conference*, pp. 418–422, 1989.
66. R. M. D. A. Velghe, D. B. M. Klaassen, and F. M. Klaassen, "Compact MOS modeling for analog circuit simulation," *Proceedings of the International Electron Devices Meeting*, pp. 485–488, 1993.
67. G. T. Cheney and R. A. Kotch, "A simple theory for threshold voltage modulation in IGFETs," *Proceedings of the IEEE*, vol. 56, pp. 837–888, 1968.
68. H. S. Lee, "An analysis of the threshold voltage for short-channel IGFETs," *Solid-State Electronics*, vol. 16, pp. 1407–1414, 1973.
69. R. C. Varschney, "Simple theory for threshold voltage modulation in short-channel MOS transistor," *Electronics Letters*, vol. 9, pp. 600–602, 1973.
70. H. C. Poon, L. D. Yau, R. L. Johnston, and D. Beecham, "D.C. model for short-channel IGFETs," *Proceedings of the International Electron Devices Meeting*, Washington, D.C., 1973, pp. 156–159.
71. L. D. Yau, "A simple theory to predict the threshold voltage of short-channel IGFETs," *Solid-State Electronics*, vol. 17, pp. 1059–1063, 1974.
72. K. O. Jeppson, "Influence of the channel width on the threshold voltage modulation in MOSFETs," *Electronics Letters*, vol. 11, pp. 297–299, 1975.
73. K. E. Kroell and G. K. Ackermann, "Threshold voltage of narrow channel field effect transistors," *Solid-State Electronics*, vol. 19, pp. 77–81, 1976.
74. W. P. Noble and P. E. Cottrell, "Narrow width effects in insulated gate field effect transistors," *Proceedings of the International Electron Devices Meeting*, pp. 582–586, 1976.
75. W. R. Bandy and D. P. Kokalis, "A simple approach for accurately modeling the threshold voltage of short-channel MOSTs," *Solid-State Electronics*, vol. 20, pp. 675–680, 1977.
76. D. J. Coe, H. E. Brakman, and K. H. Nicholas, "A simple approach for accurately modeling the threshold voltage of short-channel MOSTs," *Solid-State Electronics*, vol. 20, p. 993, 1977.
77. G. Merckel, "Short channels—Scaled down MOSFETs," *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
78. R. R. Troutman and A. G. Fortino, "Simple model for threshold voltage in short-channel IGFETs," *IEEE Transactions on Electron Devices*, vol. ED-24, pp. 1266–1268, 1977.
79. G. W. Taylor, "Subthreshold conduction in MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 337–350, 1978.
80. P. P. Wang, "Device characteristics of short-channel and narrow width MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 779–786, 1978.
81. E. Sun, "Short-channel MOS modeling for CAD," *Proceedings of the Twelfth Annual Asilomar Conference on Circuits, Systems, and Computers*, Pacific Grove, Calif., 1978, pp. 493–499.
82. W. Fichter and H. W. Potzl, "MOS modeling by analytical approximations. I. Subthreshold current and threshold voltage," *International Journal of Electronics*, vol. 46, pp. 33–55, 1979.
83. L. M. Dang, "A simple current model for short-channel IGFET and its application to circuit simulation," *IEEE Journal of Solid-State Circuits*, vol. SC-14, pp. 358–367, 1979.
84. R. R. Troutman, "VLSI limitations from drain-induced barrier lowering," *IEEE Journal of Solid-State Circuits*, vol. SC-14, pp. 383–391, 1979.
85. G. W. Taylor, "The effects of two-dimensional charge sharing on the above-threshold characteristics of short-channel IGFETs," *Solid-State Electronics*, vol. 22, pp. 701–717, 1979.
86. T. Toyabe and S. Asai, "Analytical models of threshold voltage and breakdown voltage of short-channel MOSFET's derived from two-dimensional analysis," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 453–461, 1979.
87. H. Masuda, M. Makai, and M. Kubo, "Characteristics and limitation of scaled-down MOSFETs due to two-dimensional field effect," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 980–986, 1979.
88. G. Merckel, "A simple model of the threshold voltage of short and narrow channel IGFETs," *Solid-State Electronics*, vol. 23, pp. 1207–1213, 1980.

89. A. Vladimirescu and S. Liu, "The simulation of MOS integrated circuits using SPICE 2," Memorandum No. UCB/ERL M80/7, Electronics Research Laboratory, University of California, Berkeley, February 1980.
90. P. K. Chatterjee and J. E. Leiss, "An analytic charge-sharing predictor model for sub-micron MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1980, pp. 28–33.
91. K. N. Ratnakumar, J. D. Meindl, and D. L. Scharfetter, "New IGFET short channel threshold voltage model," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1981, pp. 204–206.
92. P. P. Guebels and F. Van de Wiele, "A charge sheet model for small geometry MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1981, pp. 211–214.
93. T. N. Nguyen and J. D. Plummer, "Physical mechanisms responsible for short channel effects in MOS devices," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1981, pp. 596–599.
94. A. A. Naem and A. R. Boothroyd, "Compensation tendency of short-channel and narrow-channel effects in small-geometry IGFETs," *Electronics Letters*, vol. 18, pp. 135–136, 1982.
95. R. Kasai, K. Yokoyama, A. Yoshii, and T. Sudo, "Threshold-voltage analysis of short- and narrow-channel MOSFETs by three-dimensional computer simulation," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 870–876, 1982.
96. L. A. Akers and C. S. Chao, "A closed-form threshold voltage expression for a small-geometry MOSFET," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 776–778, 1982.
97. L. A. Akers and J. J. Sanchez, "Threshold voltage models of short, narrow, and small geometry MOSFETs: A review," *Solid-State Electronics*, vol. 25, pp. 621–641, 1982.
98. P. P. Guebels and F. Van de Wiele, "A small geometry MOSFET model for CAD applications," *Solid-State Electronics*, vol. 26, pp. 267–273, 1983.
99. C. R. Ji and C. T. Sah, "Analysis of the narrow gate effect in submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 1672–1677, 1983.
100. C. R. Viswanathan, B. C. Burkey, G. Lubberts, and T. J. Tredwell, "Threshold voltage in short-channel MOS devices," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 932–940, 1985.
101. A. M. Asenov, E. N. Stefanov, B. Z. Antov, and P. K. Vitanov, "Numerical analysis of MOS transistor effective channel width," *Electronics Letters*, vol. 21, pp. 595–597, 1985.
102. T. W. Tang, Q. L. Zhang, and D. H. Navon, "Analytical model for predicting threshold voltage in submicrometer-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 1890–1895, 1985.
103. C. T. Wang, "A threshold voltage expression for small-size MOSFET's based on an approximate three-dimensional analysis," *IEEE Transactions on Electron Devices*, vol. ED-23, pp. 160–164, 1986.
104. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
105. F. M. Klaassen, "MOS device modelling," in *Design of MOS VLSI Circuits for Telecommunications*, Y. Tsividis and P. Antognetti (editors), Prentice Hall, Englewood Cliffs, N.J., 1985.
106. *MEDICI User's Manual*, Technology Modeling Associates, Santa Clara, Calif., 1996.
107. G. W. Taylor, "Subthreshold conduction in MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 337–350, 1978.
108. M. El Banna and M. El-Nokali, "A pseudo-two-dimensional analysis of short-channel MOSFETs," *Solid-State Electronics*, vol. 31, pp. 269–274, 1988.
109. J. D. Kendall and A. R. Boothroyd, "A two-dimensional analytical solution of the Poisson and current continuity equations for the short-channel MOSFET," *Solid-State Electronics*, vol. 33, pp. 537–551, 1990.
110. M. Conti and C. Turchetti, "On the short-channel theory for MOS transistor," *IEEE Transactions on Electron Devices*, vol. ED-38, pp. 2657–2663, 1991.
111. T. A. Fjeldy and M. Shur, "Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, pp. 137–145, 1993.
112. Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, "Threshold voltage model for deep-submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, pp. 86–95, 1993.

113. B. Iniguez, "Comments on 'Threshold voltage model for deep-submicrometer MOSFETs,'" *IEEE Transactions on Electron Devices*, vol. 42, p. 1712, 1995.
114. A. Kloss and A. Kostka, "A new physics-based, predictive compact model for small geometry MOSFETs including two-dimensional calculations with a close link to process and layout data," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 147–150, 1996.
115. N. D. Arora, R. Rios, C-L Huang, and K. Raol, "PCIM: A physically based continuous short-channel IGFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 988–997, 1994.
116. N. D. Arora and M. Sharma, "Modeling the anomalous threshold voltage behavior of submicron MOSFETs," *Electron Device Letters*, vol. EDL-13, pp. 92–94, 1992.
117. C. Rafferty, H-H. Vuong, S. Eshraghi, M. Giles, M. Pinto, and S. Hilienius, "Explanation of reverse short-channel effect by defect gradients," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 311–314, 1993.
118. A. G. Lewis and J. Y. Chen, "Current trends in MOS process integration," pp. 40–117, *Advanced MOS Device Physics*, N. G. Einspruch and G. Gildenblat (editors), VLSI Electronics, vol. 18, Academic Press, New York, 1989.
119. A. Kurosawa, T. Shibata, and H. Iozuka, "A new bird's beak free isolation technology for VLSI devices," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 384–387, 1981.
120. L. A. Akers, M. Sugino, and J. M. Ford, "Characterization of the inverse-narrow-width effect," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 2476–2484, 1987.
121. K. K. Hsueh, J. J. Sanchez, T. A. Demassa, and L. A. Akers, "Inverse-narrow-width effects and small-geometry MOSFET threshold voltage model," *IEEE Transactions on Electron Devices*, vol. ED-36, pp. 325–338, 1988.
122. K. Ohe, S. Odanaka, K. Moriyama, T. Hora, and G. Fuse, "Narrow-width effects of shallow trench-isolated CMOS with n⁺-polysilicon gate," *IEEE Transactions on Electron Devices*, vol. ED-36, pp. 1110–1115, 1989.
123. R. C. Vankemmel and K. M. De Meyer, "A study of the corner effects in trench-like isolated structures," *IEEE Transactions on Electron Devices*, vol. ED-37, pp. 168–175, 1990.
124. R. M. Swanson and J. D. Meindl, "Fundamental performance limits of MOS integrated circuits," *Proceedings of the International Solid-State Circuits Conference*, Philadelphia, February 1975, pp. 110–111.
125. M. J. Deen and Z. X. Yan, "Substrate bias effects on drain-induced barrier lowering in short-channel PMOS devices," *IEEE Transaction on Electron Devices*, vol. 37, no. 7, pp. 1707–1713, 1990.
126. PSP website. Available at: <http://pspmodel.asu.edu>. Accessed Sept. 24, 2009.
127. C. Turehetti and G. Masetti, "A charge-sheet analysis of short-channel enhancement-mode MOSFETs," *IEEE Journal of Solid-State Circuits*, vol. SC-21, pp. 267–275, 1986.
128. P. Richman, *MOSFETs and Integrated Circuits*, John Wiley, New York, 1973.
129. R. A. Stuart and W. Eccleston, "Punchthrough currents in short-channel M.O.S.T. devices," *Electronics Letters*, vol. 9, pp. 586–588, 1973.
130. J. R. Brews, "Geometrical factors in avalanche punchthrough erase," *IEEE Transactions on Electron Devices*, vol. ED-24, pp. 1108–1116, 1977.
131. J. J. Barnes, K. Shimohigashi, and R. W. Dutton, "Short-channel MOSFETs in the punchthrough current mode," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 446–453, 1979.
132. F. S. Hsu, R. S. Muller, C. Hu, and P-K. Ko, "A simple punchthrough model for short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 1354–1359, 1983.
133. W. W. Lattin and J. L. Rutledge, "Impact ionization current in MOS devices," *Solid-State Electronics*, vol. 16, p. 1043, 1973.
134. T. Kamata, K. Tanabashi, and K. Kobayashi, "Substrate current due to impact ionization in MOSFETs," *Japanese Journal of Applied Physics*, vol. 15, p. 1127, 1976.
135. E. Sun, J. Moll, J. Berger, and B. Alders, "Breakdown mechanism in short-channel MOS transistors," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1978, pp. 478–482.
136. J. Matsunaga, M. Konaka, S. Kohyama, and H. Iizuku, "Design limitations due to substrate currents and secondary impact ionization electrons in NMOS LSIs," *Proceedings of the Eleventh International Conference on Solid-State Devices*, Tokyo, 1979, p. 45.

137. P. K. Chatterjee, "VLSI dynamic NMOS design constraints due to drain induced primary and secondary impact ionization," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1979, pp. 14–17.
138. B. Eitan and D. Frohman-Bentchkowsky, "Holding time degradation in dynamic MOS RAM by injection-induced electron currents," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 1515–1519, 1981.
139. Y. El-Mansy, "MOS device and technology constraints in VLSI," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 567–573, 1982.
140. S. Tam, P. Ko, F. C. Hsu, C. Hu, and P. S. Muller, "Hot electron-induced excess currents in n-channel MOSFETs," Device Research Conference, Colorado State University, Fort Collins, 1982.
141. F.-C. Hsu, P.-K. Ko, S. Tam, C. Hu, and R. S. Muller, "An analytical breakdown model for short-channel MOSFET's," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 1735–1740, 1982.
142. W. Muller, L. Risch, and A. Schutz, "Short-channel MOS transistors in the avalanche-multiplication regime," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 1778–1784, 1982.
143. R. Kuhnert, C. Werner, and A. Schutz, "A novel impact-ionization model for 1-μm MOSFET simulation," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 6, pp. 1057–1063, 1985.
144. H. Hana, Y. Okamoto, and H. Ohnuma, *Japanese Journal of Applied Physics*, vol. 9, p. 1103, 1970.
145. T. H. Ning, C. M. Osburn, and H. N. Yu, "Effect of electron tapping on IGFET characteristics," *Journal of Electronics Materials*, vol. 6, pp. 65–76, 1977.
146. T. H. Ning, P. W. Cook, R. H. Dennard, C. M. Osburn, S. E. Shuster, and H. N. Yu, "1-μm MOSFET VLSI technology: Part IV—Hot-electron design constraints," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 346–353, 1979.
147. P. E. Cottrell, R. R. Troutman, and T. H. Ning, "Hot electron emission in n-channel IGFETs," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 520–533, 1979.
148. R. B. Fair and R. C. Sun, "Threshold voltage instability in MOSFETs due to channel hot-hole emission," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 83–94, 1981.
149. S. Tam, F.-C. Hsu, C. Hu, R. S. Muller, and P. K. Ko, "Hot-electron currents in very short channel MOSFET's," *IEEE Electron Device Letters*, vol. EDL-4, pp. 249–251, 1983.
150. C. Hu, S. C. Tam, F. C. Hsu, P. K. Ko, T. Y. Chan, and K. W. Terrill, "Hot-electron induced MOSFET degradation—Model, monitor, and improvement," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 375–385, 1985.
151. E. Takeda, "Hot-carrier effects in submicrometer MOS VLSI's," *Proceedings of the IEEE*, vol. 131, part I, pp. 153–162, 1984.
152. T. Tsuchiya and J. Frey, "Relationship between hot-electrons/holes and degradation of p- and n-channel MOSFETs," *IEEE Electron Device Letters*, vol. EDL-6, pp. 8–11, 1985.
153. K. R. Hofmann, C. Werner, W. Weber, and G. Dorda, "Hot-electron and hole emission effects in short n-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 691–699, 1985.
154. Y. El-Mansy, "MOS device and technology constraints in VLSI," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 567–573, 1982.
155. K. Brennan and K. Hess, "A theory of enhanced impact ionization due to the gate field and mobility degradation in the inversion layer of MOSFETs," *IEEE Electron Device Letters*, vol. EDL-7, pp. 86–88, 1986.
156. W. Müller, L. Risch, and A. Schütz, "Short-channel MOS transistors in the avalanche-multiplication regime," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 1778–1784, 1982.
157. J. Mar, S. S. Li, and S. Y. Yu, "Substrate current modelling for circuit simulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits*, vol. CAD-1, pp. 183–186, 1982.
158. M. Fukuma and W. Lui, "MOSFET substrate current model including energy transport," *IEEE Electron Devices Letters*, vol. EDL-8, pp. 214–216, 1987.
159. J. W. Slotboom, G. Streutker, G. J. T. Devids, and P. B. Hartog, "Surface impact ionization in silicon devices," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 494–497, 1987.
160. C. T. Wang, "An improved hot-electron-emission model for simulating the gate-current characteristic of MOSFETs," *Solid-State Electronics*, vol. 31, pp. 229–231, 1988.

161. J. Higman, I. C. Kizilyalli, and K. Hess, "Nonlocality of the electron ionization coefficient in n-MOSFETs: An analytical approach," *IEEE Electron Devices Letters*, vol. EDL-9, pp. 399–401, 1988.
162. N. Goldsman and J. Frey, "Electron energy distribution for calculation of gate leakage current in MOSFETs," *Solid-State Electronics*, vol. 31, pp. 1089–1092, 1988.
163. B. Meinerzhagen, "Consistent gate and substrate current modeling based on energy transport and the lucky electron concept," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 504–507, 1988.
164. C. G. Hwang and R. W. Dutton, "Substrate current model for submicrometer MOSFETs based on mean free path analysis," *IEEE Transactions on Electron Devices*, vol. ED-36, pp. 1348–1354, 1989.
165. G-S Huang and C-Y Wu, "An analytic saturation model for drain and substrate current of conventional and LDD MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-37, pp. 1667–1677, 1990.
166. Y. A. El-Mansy and D. M. Caughey, "Modeling weak avalanche multiplication currents in IGFETs and SOS transistors for CAD," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 31–34, 1975.
167. C. Hu, "Hot-carrier effects," pp. 119–160. *Advanced MOS Device Physics*, N. G. Einspruch and G. Gildenblat (editors), VLSI Electronics, vol. 18, Academic Press, New York, 1989.
168. N. D. Arora and M. S. Sharma, "MOSFET substrate current model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 38, no. 6, pp. 1392–1398, 1991.
169. A. G. Chynoweth, "Ionization rates for electrons and holes in silicon," *Physical Review*, vol. 109, pp. 1537–1540, 1958.
170. S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, and J. F. Shepard, "Design and characteristics of the lightly doped drain-source (LDD) insulated-gate field-effect transistor," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 1359–1366, 1980.
171. E. Takeda, H. Kume, Y. Nakagome, T. Makino, A. Shimizu, and S. Asai, "An As-P ($n^+ - n^-$) double diffused rain MOSFET for VLSIs," *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 652–657, 1983.
172. H. Mikoshiba, T. Horiuchi, and K. Hamano, "Comparison of drain structures in n -channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 140–144, 1986.
173. J. J. Sanchez, K. K. Hsueh, and T. A. DeMassa, "Drain-engineered hot-electron-resistant device structures—A review," *IEEE Transactions on Electron Devices*, vol. ED-36, pp. 1125–1131, 1989.
174. L. C. Parillo, "VLSI process integration," *VLSI Technology*, S. M. Sze (editor), McGraw-Hill, New York, 1983.
175. S. Wolf, "Silicon processing in the VLSI era," *The Submicron MOSFET*, vol. 3, Lattice Press, Sunset Beach, Calif., 1995.
176. Y. Zhou, D. Connerney, R. Carroll, and T. Luk, "Modeling MOS snapback for circuit-level ESD simulation using BSIM3 and VBIC models," *Proceedings of the Sixth International Symposium on Quality of Electronic Design*, pp. 476–481, 2005.
177. S. Y. Chou, D. A. Antoniadis, and H. I. Smith, "Observation of electron velocity overshoot in sub-100-nm-channel MOSFETs in Si," *IEEE Electron Device Letters*, vol. EDL-6, pp. 665–667, 1985.
178. T. Kobayashi and K. Saito, "Two-dimensional analysis of velocity overshoot effects in ultrashort-channel Si MOSFETs," *IEEE Transaction on Electron Devices*, vol. ED-32, pp. 782–792, 1985.
179. G. G. Shahidi, D. A. Antoniadis, and H. I. Smith, "Electron velocity overshoot at room and liquid nitrogen temperatures in silicon inversion layers," *IEEE Electron Device Letters* EDL-9, vol. 94–96, 1988.
180. G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. A. Rishton, and E. Ganin, "High transconductance and velocity overshoot in NMOS devices at the 0.1 μm gate-length level," *IEEE Electron Device Letters*, vol. EDL-9, pp. 464–466, 1988.
181. S. E. Laux and M. V. Fischetti, "Monte Carlo simulation of submicrometer Si n-MOSFETs at 77 and 300 K," *IEEE Electron Device Letters*, vol. 9, pp. 467–469, 1988.
182. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go?" *Proceedings of the IEEE International Electron Devices Meeting*, pp. 553–556, 1992.

324 OPERATION AND MODELING OF THE MOS TRANSISTOR

183. K. Natori, "Ballistic metal-oxide-semiconductor field effect transistors," *Journal of Applied Physics*, vol. 76, p. 4879, 1994.
184. J.-H. Song, Y.-Y. Park, and H. S. Min, "Drain current enhancement due to velocity overshoot effects and its analytical modeling," *IEEE Transactions on Electron Devices*, vol. 43, pp. 1870–1875, 1996.
185. M. Lundstrom, "Elementary scattering theory of the Si MOSFET," *IEEE Electron Device Letters*, vol. EDS-18, pp. 361–363, 1997.
186. M. R. Pinto, E. Sangiorgi, and J. Bude, "Silicon MOS transconductance scaling into the velocity overshoot regime," *IEEE Electron Device Letters*, vol. 14, pp. 375–378, 1993.
187. A. Lochtefeld and D. A. Antoniadis, "On experimental determination of carrier velocity in deeply scaled NMOS: How close to the thermal limit?" *IEEE Electron Device Letters*, vol. 22, no. 2, pp. 95–97, 2001.
188. A. Khakifirooz, O. Nayfeh, and D. A. Antoniadis, "A simple semiempirical short-channel MOSFET current-voltage model continuous across all regions of operation and employing only physical parameters," *IEEE Transactions on Electron Devices*, vol. 56, pp. 1674–1680, 2009.
189. C.-Y. Lu, M. Sung, H. C. Kirsch, S. T. Hillenius, T. E. Smith, and L. Manchanda, "Anomalous C-V characteristics of implanted poly MOS structure in n⁺/p⁺ dual-gate CMOS technology," *IEEE Electron Device Letters*, vol. 10, pp. 192–194, 1989.
190. N. D. Arora, R. Rios, and C.-L. Huang, "Modeling the polysilicon depletion effect and its impact on submicrometer CMOS circuit performance," *IEEE Transactions on Electron Devices*, vol. 42, pp. 935–943, 1995.
191. B. Riccò, R. Versari, and D. Esseni, "Characterization of polysilicon-gate depletion in MOS structures," *IEEE Electron Device Letters*, vol. 17, pp. 103–105, 1996.
192. G. Gildenblat, T.-L. Chen, and P. Bendix, "Analytical approximation for perturbation of MOSFET surface potential by polysilicon depletion layer," *Electronics Letters*, vol. 35, no. 22, pp. 1974–1976, 1999.
193. N. D. Arora, R. Rios, and C.-L. Huang, "Modeling the polysilicon depletion effect and its impact on submicrometer CMOS circuit performance," *IEEE Transactions on Electron Devices*, vol. 24, no. 5, pp. 935–943, 1995.
194. J.-M. Sallese, M. Bucher, and C. Lallement, "Improved analytical modeling of polysilicon depletion in MOSFETs for circuit simulation," *Solid-State Electronics*, vol. 44, no. 6, pp. 905–912, 2000.
195. C. Galup-Montoro and M. C. Schneider, *MOSFET Modeling for Circuit Analysis and Design*, World Scientific, Singapore, 2007.
196. M. Miura-Mattausch, H. J. Mattausch, and T. Ezaki, *The Physics and Modeling of MOSFETs*, World Scientific, Singapore, 2008.
197. BSIM website. Available at <http://www-device.eecs.berkeley.edu/~bsim3/>. Accessed Sept. 29, 2009.
198. F. Stern and W. E. Howard, "Properties of semiconductor surface inversion layers in the electric quantum limit," *Physical Review B*, vol. 163, pp. 816–835, 1967.
199. F. Stern, "Quantum properties of surface space-charge layers," *CRC Critical Revnes in Solid-State Sciences*, pp. 499–514, 1974.
200. M. J. van Dort, P. H. Woerlee, A. J. Walker, C. A. H. Juffermans, and H. Lifka, "Quantum-mechanical threshold voltage shifts of MOSFETs caused by high levels of channel doping," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 495–498, 1991.
201. M. J. van Dort, P. H. Woerlee, and A. J. Walker, "A simple model for quantization effects in heavily-doped silicon MOSFETs at inversion conditions," *Solid-State Electronics*, vol. 37, no. 3, pp. 411–414, 1994.
202. C. Y. Wong, J. Y.-C. Sun, Y. Taur, C. S. Oh, R. Angelucci, and B. Davari, "Doping of n⁺ and p⁺ polysilicon in dual-gate CMOS process," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 238–241, 1988.
203. R. Rios and N. D. Arora, "Determination of ultra-thin gate oxide thickness for CMOS structures using quantum effects," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 613–616, 1994.
204. M. J. van Dort, P. H. Woerlee, A. J. Walker, C. A. H. Juffermans, and H. Lifka, "Influence of high substrate doping levels on the threshold voltage and the mobility of deep submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-39, pp. 932–938, 1994.

205. J. W. Slotboom and H. C. de Graaf, "Measurements of bandgap narrowing in Si bipolar transistors," *Solid-State Electronics*, vol. 19, pp. 857–862, 1976.
206. K. Morikawa, H. Ueno, D. Kitamatu, M. Tanaka, T. Okagaki, M. Miura-Mattausch, and H. J. Mattausch, "Quantum effect in sub-0.1 μm MOSFET with pocket technologies and its relevance for the on-current condition," *Japanese Journal of Applied Physics*, vol. 41, no. 4B, pp. 2359–2362, 2002.
207. M. S. Lundstrom and J. Guo, *Nanoscale Transistors: Device Physics, Modeling and Simulation*, Springer, New York, 2006.
208. B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelectronics—I. MOS technology," *Solid-State Electronics*, vol. 15, pp. 819–829, 1972.
209. R. Tsu and L. Esaki, "Tunneling in a finite superlattice," *Applied Physics Letters*, vol. 22, pp. 562–564, 1973.
210. R. van Langevelde, A. J. Scholten, R. Duffy, F. N. Cubaynes, M. J. Knitel, and D. B. M. Klaassen, "Gate current: Model, DL extraction and impact on RF performance," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 289–292, 2001.
211. W.-K. Shih, R. Rios, P. Paekan, K. Mistry, and T. Abbott, "A general partition scheme for gate leakage current suitable for MOSFET compact models," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 293–296, 2001.
212. X. Gu, T.-L. Chen, G. Gildenblat, G. O. Workman, S. Veeraraghavan, S. Shapiro, and K. Stiles, "A surface potential-based compact model of n-MOSFET gate-tunneling current," *IEEE Transactions on Electron Devices*, vol. 51, no. 1, pp. 127–135, 2004.
213. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge, 1998.
214. R. B. Fair and H. W. Wivell, "Zener and avalanche breakdown in As-implanted low voltage Si p-n junctions," *IEEE Transactions on Electron Devices*, vol. ED-23, pp. 512–518, 1976.
215. J. Chen, T. Y. Chan, C. Chen, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Letters*, vol. EDL-8, no. 11, pp. 515–517, 1987.
216. R. Inagaki, M. Miura-Mattausch, and Y. Inoue, "A GIDL-current model for advanced MOSFET technologies implemented into HiSiM2," *Proceedings of the International Conference on Communications, Circuits, and Systems*, pp. 1057–1061, 2007.
217. E. Kane, "Zener tunneling in semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 12, pp. 181–188, 1959.
218. "The international technology roadmap for semiconductors," Semiconductor Industry Association, <http://www.itrs.net>.
219. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. SC-9, pp. 256–268, 1974.
220. H.-N. Yu, A. Reisman, C. M. Osborn, and D. L. Critchlow, "1 μm MOSFET VLSI technology, Part I: An Overview," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 318–324, 1979.
221. R. H. Dennard, F. H. Gaensslen, E. J. Walker, and P. W. Cook, "1 μm MOSFET VLSI technology: Part II—Device designs and characteristics for high-performance logic applications," *IEEE Transactions on Electron Devices*, vol. ED-26, pp. 325–333, 1979.
222. E. O. Johnson, "The insulated-gate field-effect transistor—A bipolar transistor in disguise," *RCA Review*, vol. 34, pp. 80–94, 1973.
223. Y. El-Mansy, "MOS device and technology constraints in VLSI," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 567–573, 1982.
224. VLSI Laboratory, Texas Instruments, "Technology and design challenges of MOS VLSI," *IEEE Journal of Solid-State Circuits*, vol. SC-17, pp. 442–448, 1982.
225. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized scaling theory and its application to a 1/4 micron MOSFET design," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 452–462, 1984.
226. A. Reisman, "Device, circuit, and technology scaling to micron and submicron dimensions," *Proceedings of the IEEE*, vol. 71, pp. 550–565, 1983.
227. J. H. King, "A novel approach to silicon gate CMOS device scaling," *Solid-State Electronics*, vol. 26, pp. 879–891, 1983.

228. E. Sangiorgi, E. A. Hofstatter, R. K. Smith, P. F. Bechtold, and W. Fichtner, "Scaling issues related to high field phenomena in submicrometer MOSFETs," *IEEE Electron Device Letters*, vol. EDL-7, pp. 115–118, 1986.
229. P. K. Chatterjee, W. R. Hunter, T. C. Holloway, and Y. T. Lin, "The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI," *Electron Device Letters*, vol. EDL-1, pp. 220–223, 1980.
230. Y. Tsividis, "Moderate inversion in MOS devices," *Solid-State Electronics*, vol. 25, pp. 1099–1104, 1982; see also Erratum, *ibid.*, vol. 26, p. 823, 1983.
231. G. Baccarani and M. R. Wordeman, "Transconductance degradation in thin-oxide MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 1295–1304, 1983.
- 231a. R. W. Keys, "Physical limits in digital electronics," *Proceedings of the IEEE*, vol. 63, pp. 740–767, May 1975.
- 231b. O. G. Folberth and J. H. Bleher, "The fundamental limitations of digital semiconductor technology," *Microelectronics Journal*, vol. 9, pp. 33–41, 1979.
- 231c. K. N. Ratnakumar and J. D. Meindl, "Performance limits of E/D NMOS VLSI," *Proceedings of the International Solid-State Circuits Conference*, San Francisco, February 1980, pp. 72–73.
- 231d. C. Svensson, "VLSI physics," *Integration*, vol. 1, pp. 3–19, 1983.
- 231e. H. Shichijo, "A re-examination of practical performance limits of scaled n-channel and p-channel MOS devices for VLSI," *Solid-State Electronics*, vol. 26, pp. 969–986, 1983.
- 231f. J. Pfeister, J. D. Shott, and J. D. Meindl, "Performance limits of NMOS and CMOS," *Proceedings of the International Solid-State Circuits Conference*, San Francisco, 1984, pp. 158–159.
- 231g. J. R. Pfeister, J. D. Shott, and J. D. Meindl, "Performance limits of CMOS ULSI," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 333–343, February 1985.
232. J. Meindl, "Circuit scaling limits for ultra large scale integration," *Proceedings of the International Solid-State Circuits Conference*, New York, February 1981, pp. 36–37.
233. [Online]. Available at <http://www.itrs.net/Links/2008ITRS/Home2008.htm>. Accessed Sept. 29, 2009.
234. D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, "A fully depleted lean-channel transistor (DELTA)—A novel vertical ultra-thin SOI MOSFET," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 833–836, 1989.
235. X. Huang, W. C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y. K. Choi, K. Asano, V. Subramanian, T. J. King, J. Bokor, and C. Hu, "Sub-50-nm FinFET: PMOS," *Proceedings of the International Electron Devices Meeting*, pp. 67–70, 1999.
236. F. Balestra, S. Cristoloveanu, M. Benachir, J. Brini, and T. Elewa, "Double-gate silicon-on-insulator transistors with volume inversion: A new device with greatly enhanced performance," *IEEE Electron Device Letters*, vol. EDL-8, no. 9, pp. 410–412, 1987.
237. B. S. Doyle, S. Datta, M. Dockzy, S. N. Harelund, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High performance fully-depleted tri-gate CMOS transistors," *IEEE Electron Device Letters*, vol. 24, no. 3, pp. 263–265, 2003.
238. J. P. Colinge, X. Baie, V. Bayot, and E. Grivei, "A silicon-on-insulator quantum wire," *Solid-State Electronics*, vol. 39, no. 1, pp. 49–51, 1996.
239. K. Suzuki, T. Tanaka, Y. Tosaka, H. Hori, and Y. Arimoto, "Scaling theory for double-gate SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, no. 12, pp. 2326–2329, 1993.
240. R. H. Yan, A. Ourmazd, and K. F. Lee, "Scaling the Si MOSFET: From bulk to SOI to bulk," *IEEE Transactions on Electron Devices*, vol. ED-39, pp. 1704–1710, 1992.
241. C. Auth and J. Plummer, "Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFETs," *IEEE Electron Device Letters*, vol. 18, pp. 74–76, 1997.
242. J.-T. Park, and J. P. Colinge, "Multiple-gate SOI MOSFETs: Device design guidelines," *IEEE Transactions on Electron Devices*, vol. 49, no. 12, pp. 222–229, 2002.

Problems

- 5.1.** Derive expressions corresponding to (5.2.9) and (5.2.10) assuming, instead of (5.2.3), a two-section piecewise-linear model for $v_d(\mathcal{E}_v)$ (i.e., assume that $|v_d| = \mu|\mathcal{E}_v|$ for $|\mathcal{E}_v| \leq \mathcal{E}_c$, and $|v_d| = |v_d|_{\max}$ for $|\mathcal{E}_v| > \mathcal{E}_c$).
- 5.2.** Prove (5.2.12) and show that, in the absence of velocity saturation effects ($L\mathcal{E}_c$ approaching infinity), V'_{DS} as given there reduces to $(V_{GS} - V_T)/\alpha$.
- 5.3.** Assume that the longitudinal electric field magnitude $|\mathcal{E}_v|$, which drives velocity saturation, varies linearly along the channel, from a value of zero at the source to a value of $2V_{DS}/L$ at the drain.⁴¹ Use this in the velocity saturation model (5.2.13), with $n=2$, to derive a closed form expression, similar to (5.2.10), for the overall effect of velocity saturation [the expression will be more complex than (5.2.10)]. The assumed field variation is clearly unrealistic for small V_{DS} . Why is this not an issue?
- 5.4.** Give the complete equations for nonsaturation and saturation in the presence of velocity saturation and channel length modulation effects, using a model based on (5.2.11), (5.3.1), and (5.3.3). The value of V'_{DS} should be slightly modified from (5.2.12) to ensure continuity of both I_{DS} and dI_{DS}/dV_{DS} at $V_{DS} = V'_{DS}$.
- 5.5.** (a) Consider a device with $\mu C'_{ox} = 500 \mu\text{A/V}^2$, $\alpha = 1.1$, $\mathcal{E}_c = 2.5 \times 10^4 \text{ V/cm}$, and $V_T = 0.4 \text{ V}$. By using the model of Example 5.1, plot I_{DS} vs. V_{DS} up to 1.2 V, with V_{GS} as a parameter ($V_{GS} = 0.6, 0.9, 1.2 \text{ V}$), and for $W=L=0.8, 0.4, \text{ and } 0.2 \mu\text{m}$.
(b) Repeat for the preceding values of L , but keep W constant at 0.8 μm . Discuss the results obtained. Neglect the effect of channel length and width on the threshold voltage.
- 5.6.** Derive (5.3.3) using (1.2.29) (see Appendix A).
- 5.7.** Prove (5.3.9) and state the condition(s) for its validity.
- 5.8.** Consider a model consisting of (4.7.18) in nonsaturation, and (5.3.1) and (5.3.3) in saturation. Assume the two regions are adjacent at the point $V_{DS} = \hat{V}_{DS}$. Find the value of \hat{V}_{DS} that will guarantee continuity of $I_{DS}(V_{DS})$ and of its slope.
- 5.9.** Consider a transistor with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2 \text{ nm}$, $V_{FB} = -0.9 \text{ V}$, and junction depth of 100 nm, biased at $V_{GS} = 0.6 \text{ V}$, $V_{SB} = 0$. Consider channel length modulation as described by (5.3.5). Plot I_p vs. V_{DS} for values of V_{DS} between $V'_{DS} + 0.2 \text{ V}$ and $V'_{DS} + 1 \text{ V}$. How long should the device channel be to ensure a I_p/L of 1 percent or less for the voltage range considered?
- 5.10.** Examine carefully the arguments related to charge sharing up to (5.4.3). Identify the points where arbitrary assumptions were used.
- 5.11.** Prove (5.4.7) and simplify it for the case of small d_B/d_j .
- 5.12.** Equation (5.4.7) was derived for L large enough so that a trapezoidal region could be defined as shown in Fig. 5.7. Investigate the case where L is so small that the trapezoid becomes a triangle. Derive expressions for \hat{Q}_B/Q_B and \hat{V}_T , in this case. Assume punch-through does not occur.
- 5.13.** Consider the device of Prob. 5.9; assume $V_{DS} = 0$. Plot \hat{V}_T vs. L for L between 0.2 and 5 μm , and for $V_{SB} = 0, 0.75, \text{ and } 1 \text{ V}$.
- 5.14.** Proceeding along the lines of the derivation of (5.4.7), derive a similar expression for the case of a relatively small nonzero V_{DS} , assuming that the edge of the depletion region in the center part of the device is horizontal at a depth equal to the average value of the source and drain depletion regions. From this find \hat{V}_T .
- 5.15.** Prove (5.4.15).

- 5.16.** Equation (5.4.15) was derived for $V_{DS} = 0$. Attempt to extend this result to the case of a small nonzero V_{DS} .
- 5.17.** For the device of Prob. 5.9, and assuming STI isolation and that the channel is long, plot \hat{V}_T vs. W for W between 0.5 and 5 μm and for $V_{SB} = 0, 0.75$, and 1.5 V ($V_{DS} = 0$). *Assume $C_F = 10^{-1} \text{ F}/\mu\text{m}$*
- 5.18.** Consider a device with LOCOS isolation and with a long narrow channel, in which the effects of effective threshold increase and effective channel narrowing are both evident (Sec. 5.4.3). Suggest a measuring technique that can be used to determine the effective threshold and the effective channel width.
- 5.19.** Define an effective deletion region charge \hat{Q}_B (Sec. 5.4) in the case of channels that are *both* short *and* narrow. Give the resulting expression for \hat{V}_T in terms of \hat{Q}_B/Q_B , and attempt to justify a common approach for combining short- and narrow-channel effects when each effect acting by itself is small, which assumes that $V_T(V_{DS}) = V_T + \Delta V_{TL}(L, V_{DS}) + \Delta V_{TW}(W, V_{DS})$, where ΔV_{TL} and ΔV_{TW} are modeled as in Secs. 5.4.2 and 5.4.3. Assume STI isolation.
- 5.20.** For a device with $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2.5 \text{ nm}$, and $V_{T0} = 0.4 \text{ V}$, plot the effective threshold vs. V_{DS} , for V_{DS} between 0 and 1.2 V, for $L = 0.2, 0.15$, and 0.1 μm , assuming $\phi_0 = 2\phi_F + 5\phi_i$, and $V_{SB} = 0$, including DIBL effects. Repeat for $V_{SB} = 0.8 \text{ V}$.
- 5.21.** For the device of Prob. 5.20, plot the factor by which the off current (current with $V_{GS} = 0$ in digital operation) changes due to DIBL as a function of L , assuming $V_{DS} = 1.2 \text{ V}$, and $V_{SB} = 0$.
- 5.22.** For the device of Prob. 5.20, assume $\mu C'_ox = 500 \mu\text{A/V}^2$, $\alpha = 1.1$, $W = 2 \mu\text{m}$, and $L = 0.1 \mu\text{m}$. Plot I_{DS} vs. V_{DS} , for V_{DS} between 0 and 1.2 V, and for $V_{GS} = 0.6, 0.8$, and 1.0 V, including DIBL effects.
- 5.23.** Develop a formula for roughly estimating the value of V_{DS} at which punchthrough begins.
- 5.24.** For the device of Prob. 5.9, how much does V_{DS} need to be above V'_{DS} in order for the substrate current due to impact ionization to rise to 10 percent of the drain-source current? Assume $K_i = 2 \text{ V}^{-1}$ and $V_i = 20 \text{ V}$.
- 5.25.** For the case of polysilicon depletion, prove (5.10.5)–(5.10.7) and (5.10.9).
- 5.26.** Assume that the device of Prob. 5.9 is made with a polysilicon gate with *n*-type doping concentration of $5 \times 10^{19} \text{ cm}^{-3}$. Estimate the resulting change in its effective oxide thickness and threshold voltage.
- 5.27.** Assume that for a transistor exhibiting impact ionization as in (5.8.4), $K_i = 1 \text{ V}^{-1}$ and $V_i = 20 \text{ V}$. How much above V'_{DS} must V_{DS} be in order for the impact ionization body current to be 1 percent of the drain current? 10 percent?
- 5.28.** Provide the detailed proofs of (5.10.11) and (5.10.14).
- 5.29.** For the device of Prob. 5.9, what is the lowest polysilicon gate doping concentration needed to ensure that the increase in threshold voltage due to poly depletion is no more than 50 mV? How much will the weak-inversion slope factor deteriorate for such doping?
- 5.30.** Prove (5.11.2).
- 5.31.** Propose an approximate, empirical relation for the gate leakage current vs. oxide thickness for the data given in Fig. 5.30, for a gate-source voltage of 1.2 V.
- 5.32.** It is claimed that for any *n*MOS short-channel device, we can decrease the drain current at will compared with its off value (i.e., the value of the current when the device is off in digital operation, with $V_{GS} = 0$) by making V_{GS} sufficiently negative. Is this true?
- 5.33.** Extend Table 5.2, giving values for all entries in Table 5.1, for the cases of constant voltage, quasi-constant voltage, and generalized scaling.

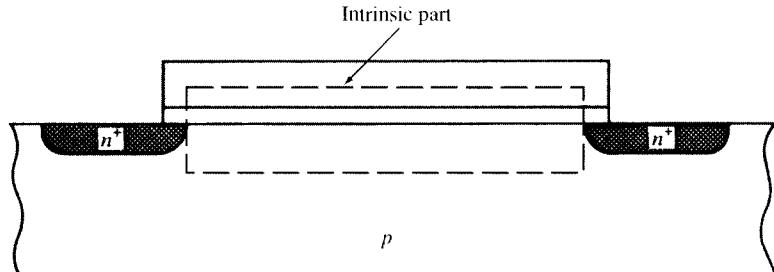
The MOS Transistor in Dynamic Operation—Large-Signal Modeling

6.1 Introduction

The MOS transistor has been treated in previous chapters with the assumption that all terminal voltages are constant. However, the device is usually employed in circuits with *varying* terminal voltages. Such “dynamic” operation causes the transistor charges to vary, and the charge changes must be supplied from the outside world by extra currents flowing through the device terminals; here “extra” refers to currents not predicted by dc theory.

The subject of this chapter is the evaluation of charges and terminal currents under dynamic operation, without placing restrictions on the magnitude of the variations; i.e., we will deal with the *large-signal* dynamic operation of the MOS transistor.^{1-71†} We will concentrate first on the device part between source and drain, containing the inversion layer, the depletion region, the oxide, and the lower part of the gate (recall that the gate is actually very tall, as seen in Fig 1.38a or 1.40). In most figures, we do not show it as such in order to save space). This part is shown enclosed in a broken line in Fig. 6.1. It is called the *intrinsic* part, it is the part mainly responsible for transistor action, and it is the subject of Secs. 6.1–6.7. The rest of the device constitutes the *extrinsic* part and is responsible for parasitic effects, which can limit overall performance. We will postpone discussion of the extrinsic device part until Sec. 6.8.

† The references include some papers mainly on small-signal operation, which, nevertheless, contain material pertinent to our discussion in this chapter. Chapters 7 and 8 contain an extensive discussion of small-signal operation and many more related references.

**FIGURE 6.1**

The intrinsic part of a transistor is enclosed within the broken line. The part outside is the extrinsic part.

Unless stated otherwise in this chapter, we will assume long and wide channels and a uniform substrate, and we will ignore the effects of the resistances of the substrate and the gate. Also, we will assume that no hot electron effects and no gate leakage effects are present; the gate and substrate currents in this chapter are not due to such effects, but rather to changes in the gate and substrate charges, respectively.

We assume that the speeds involved are low enough so that we do not have to consider the magnetic fields resulting from the varying currents; we will thus continue to use basic electrostatic relations, only with constant charges and voltages replaced by time-varying quantities. Finally, we assume that the transistor dimensions are much smaller than the wavelength corresponding to the maximum frequency of interest, so that the transistor obeys Kirchhoff's current law and thus remains overall electrically neutral.

6.2 Quasi-Static Operation

We consider a fictitious device as shown in Fig. 6.2, with source and drain omitted to emphasize that only the intrinsic part is under consideration. The device is driven by four dc voltages, V_D , V_G , V_B , and V_S , defined with respect to some arbitrary reference point denoted by the ground symbol. Let four dc currents I_D , I_G , I_B , and I_S be defined as *entering* the device, as shown in the figure. We have seen that current flow is caused by the *transport* of electrons in the inversion layer (an *n*-channel device is assumed). Defining the *transport current* (or "conduction" current) as flowing from the drain through the channel to the source and denoting it by I_T , we have

$$I_D = I_T \quad (6.2.1a)$$

$$I_G = 0 \quad (6.2.1b)$$

$$I_B = 0 \quad (6.2.1c)$$

$$I_S = -I_T \quad (6.2.1d)$$

where we have neglected the extremely small leakage current through the insulator and the small leakage current through the depletion region. We have already derived

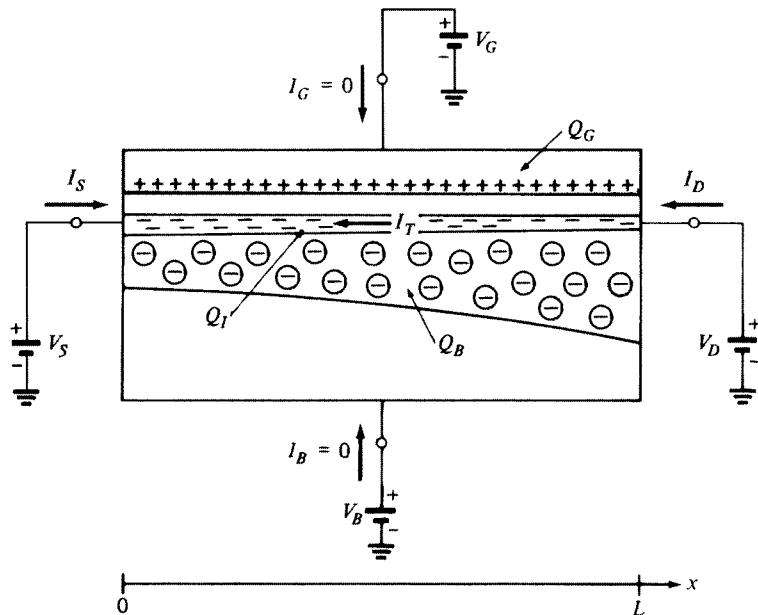


FIGURE 6.2
Definition of currents and intrinsic charges under dc excitation.

several expressions for I_T (called I_{DS} in Chap. 4), which can be written in the general form[†]:

$$I_T = h_T(V_D, V_G, V_B, V_S) \quad (6.2.2)$$

where $h_T(\)$ is an appropriate function dependent on the model used to describe the device's dc behavior. Such models have been discussed in Chap. 4.

In previous chapters, we have considered the inversion layer, gate, and depletion region charges per unit area (Q'_I , Q'_G , and Q'_B). These quantities depend, in general, on the position x along the channel. We will need to determine the corresponding *total* charges Q_I , Q_G , and Q_B . Consider Q_I as an example. A chunk of the inversion layer, with width W and length Δx at position x along the channel, will contain a charge $Q'_I(W\Delta x)$, where Q'_I depends on x . Thus, the total inversion layer charge will be

$$Q_I = \int_0^L Q'_I W dx$$

[†] In most cases in previous chapters we have given the drain current as a function of voltages of the form V_{KL} , where K and L can denote any two transistor terminals. We can write $V_{KL} = V_K - V_L$, in which case an expression in the form of (6.2.2) is obtained. In general cases where the drain current expression includes surface potential values of the source and drain ends of the channel (Sec. 4.3), we can, in principle, use for those potentials an approximate *explicit* expression in terms of $V_{SB} = V_S - V_B$, $V_{DS} = V_D - V_B$, and $V_{GB} = V_G - V_B$; again, the result will be an explicit expression for the drain current in the form of (6.2.2).

or

$$Q_I = W \int_0^L Q'_I dx \quad (6.2.3a)$$

Similarly,

$$Q_G = W \int_0^L Q'_G dx \quad (6.2.3b)$$

$$Q_B = W \int_0^L Q'_B dx \quad (6.2.3c)$$

We will undertake the evaluation of these integrals in Sec. 6.4. For now, we only note that, as might be expected, the final results will depend on the terminal voltages

$$Q_I = f_I(V_D, V_G, V_B, V_S) \quad (6.2.4a)$$

$$Q_G = f_G(V_D, V_G, V_B, V_S) \quad (6.2.4b)$$

$$Q_B = f_B(V_D, V_G, V_B, V_S) \quad (6.2.4c)$$

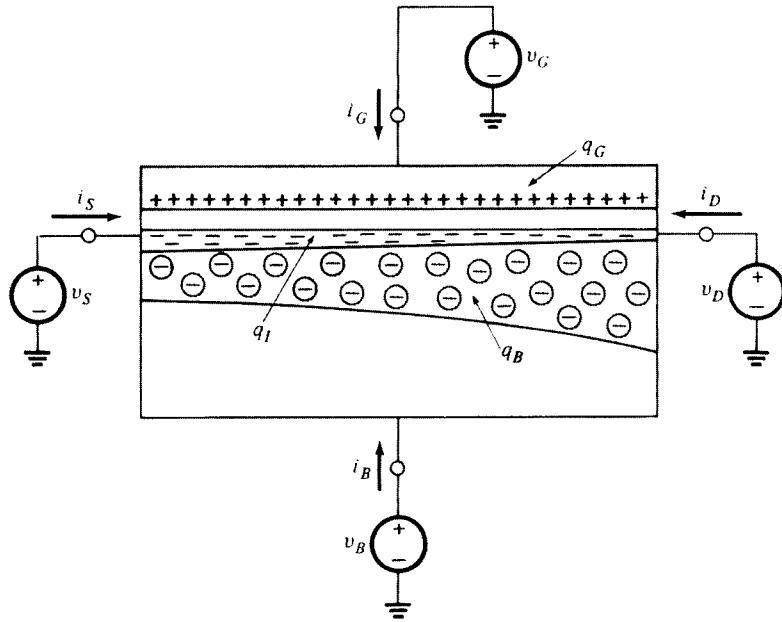
Q_G and Q_B can be interpreted as charges “stored” in the device; however, *the interpretation of Q_I requires more care*. Q_I is due to electrons in the inversion layer. These electrons are not really stored in the device. They enter through the source and eventually leave through the drain, being continuously replaced by new electrons entering through the source. Q_I is simply the total charge of the electrons that happen to be in the inversion layer at any given instant. This quantity is constant in Fig. 6.2 despite the fact that the “individual electrons” giving rise to it may be different at different instants.

We will now allow the terminal voltages to vary with time. Total time-varying quantities will be denoted by lowercase symbols with uppercase subscripts, as shown in Fig. 6.3. We will assume that *the variation of the terminal voltages is sufficiently slow, so that the device operates quasi-statically*,^{5,8,10,15,16,18–21,25,26,28–39,41–53,55–58,60,61} as defined below.

Quasi-static operation assumption: Let $v_D(t)$, $v_G(t)$, $v_B(t)$, and $v_S(t)$ be the varying terminal voltages; then at any position, the charges per unit area at any time t' are assumed identical to those that would be found if dc voltages were used instead: $V_D = v_D(t')$, $V_G = v_G(t')$, $V_B = v_B(t')$, and $V_S = v_S(t')$.

Under this assumption, the total charges q_I , q_G , and q_B can still be found from (6.2.3), and will be given by

$$q_I(t) = f_I(v_D(t), v_G(t), v_B(t), v_S(t)), \quad \text{quasi-static operation} \quad (6.2.5a)$$

**FIGURE 6.3**

Definition of currents and intrinsic charges in the presence of varying terminal voltages. Lowercase symbols with capital subscripts denote total time-varying quantities.

$$q_G(t) = f_G(v_D(t), v_G(t), v_B(t), v_S(t)), \quad \text{quasi-static operation} \quad (6.2.5b)$$

$$q_B(t) = f_B(v_D(t), v_G(t), v_B(t), v_S(t)), \quad \text{quasi-static operation} \quad (6.2.5c)$$

where f_I , f_G , and f_B represent the *same* functions as in (6.2.4). However, in contrast to the calculation of the charges, the currents *cannot* be evaluated by using the current formulas for static operation. For example, since q_G is varying, there will be a nonzero gate current; this current cannot be predicted using the relation (6.2.1b); that equation is only valid for static operation. The proper way to calculate the currents in quasi-static operation will be discussed shortly.

It may be intuitively clear that the assumption of quasi-static operation will fail if the terminal voltages vary too fast. For example, if a step waveform is used for one of them, the charges will exhibit some inertia and cannot be expected to readjust themselves instantaneously. Such cases are not considered in this and the next several sections. Instead, we are assuming that the terminal voltages vary sufficiently slowly for the quasi-static approximation to be valid. The limits of validity of the quasi-static approximation, and the reasons for its failing in extreme cases, will be discussed in Sec. 6.6. Non-quasi-static analysis will be discussed in Sec. 6.7 and, for small signals, in Sec. 8.4.

One can obtain intuition about the quasi-static approximation by using the fluid dynamical analog introduced in Sec. 1.6. Consider the case shown in Fig. 6.4.

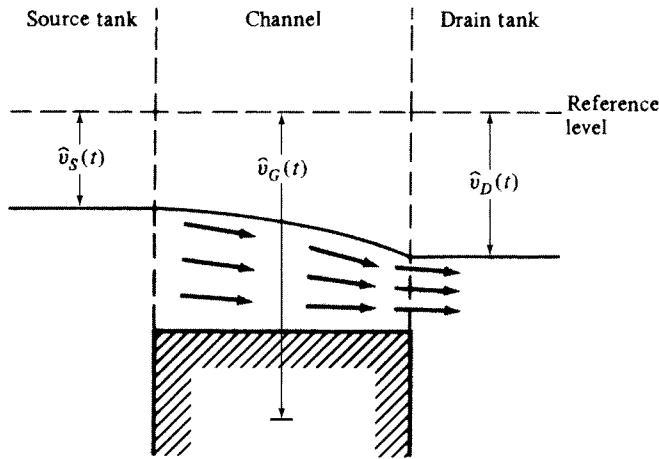


FIGURE 6.4
A fluid dynamical analog for illustrating charge motion in a transistor.

Assume the piston is moving, and let $\hat{v}_G(t)$ be the depth of its handle below the reference level. If the piston is moving sufficiently slowly, the fluid distribution at any instant t' will be practically as if $\hat{v}_G(t)$ has been frozen permanently at the value $\hat{v}_G(t')$. Then the quasi-static approximation holds for the fluid dynamical analog. Obviously, if $\hat{v}_G(t)$ varies fast, this will no longer be the case, since the fluid is not given enough time to accommodate itself. Similar comments hold if \hat{v}_S or \hat{v}_D is varied.

6.3 Terminal Currents in Quasi-Static Operation

We will now evaluate the terminal currents of the *idealized* device in Fig. 6.3, assuming *quasi-static operation*. (The total terminal currents of a real device will include current components owing to extrinsic parasitic capacitances. Such capacitances are discussed in Sec. 6.8.1.)

Assuming there is no gate leakage and, therefore, no gate transport current, all gate current in Fig. 6.3 is associated with a changing gate charge:

$$\boxed{i_G(t) = \frac{dq_G}{dt}} \quad (6.3.1)$$

Similarly, assuming no leakage in the bulk, the transport current in the depletion region is zero. Then all bulk current is associated with free electrons changing the charge in that region, by depleting or “covering” acceptor atoms there. More specifically, assume that the bulk charge increases by an amount Δq_B . This means that free electrons of total charge $-\Delta q_B$ must have left the acceptors they were previously covering and must have exited through the substrate terminal (we assume

no electron-hole recombination). This is equivalent to an opposite charge, $+\Delta q_B$, *entering* the device through the substrate terminal. Thus, the movement of mobile charges through the substrate terminal can be related to the change in the charge of the depletion region, and we can write

$$i_B(t) = \frac{dq_B}{dt} \quad (6.3.2)$$

Finally, the sum of the drain and source currents represents the total current entering the channel, which will change the inversion layer charge:^{5,8,10,15,16,18–21}

$$i_D(t) + i_S(t) = \frac{dq_I}{dt} \quad (6.3.3)$$

We need separate expressions for $i_D(t)$ and $i_S(t)$. *The reader is cautioned that, under the assumption of quasi-static operation, conflicting ways to evaluate these quantities are suggested in the literature.* This is because of conflicting interpretations of the “charges associated with the drain and with the source,” quantities to be discussed later. We emphasize that just looking at equations can be very misleading in the development of relations for $i_D(t)$ and $i_S(t)$. *A correct interpretation of the charges appearing in such equations is essential.* This requires making several fine points as explained in the following.

We begin by noting that at dc, $i_D(t) = -i_S(t)$ [see (6.2.1)]. Hence, from (6.3.3), $dq_I/dt = 0$; that is, q_I is constant. In the general case of time-varying voltages, though, q_I will be varying too. Then $dq_I/dt \neq 0$, and $i_D(t)$ cannot be equal to $-i_S(t)$ in (6.3.3). To illustrate this point, assume that in Fig. 6.3, voltages v_D , v_S , and v_B are kept constant but v_G is varying as in Fig. 6.5a. We will assume that the variations of v_G are slow enough so that quasi-static operation is maintained. Then $q_I(t)$ will be given by an equation of the form (6.2.5a); its magnitude is plotted in Fig. 6.5b. The corresponding currents $i_D(t)$ and $-i_S(t)$ are shown in Fig. 6.5c. To interpret this figure, note that $i_D(t)$ is a measure of the number of electrons leaving the device through the drain per unit time and $-i_S(t)$ is a measure of the number of electrons entering the device through the source per unit time.[†] If “dc-like” behavior were obeyed at all times, we would have $-i_S(t) = i_D(t) = i_T(t)$. The quantity $i_T(t)$ would be given by (6.2.2) after replacing V_G in it by $v_G(t)$. This is shown by the dashed line in Fig. 6.5c. However, what we actually have is shown by the solid lines and can be explained as follows. When $v_G(t)$ is increasing, $|q_I(t)|$ must increase, as shown in Fig. 6.5b. Thus the number of electrons in the channel must increase. For this to happen, the rate of supply of electrons from the source, $-i_S(t)$, must temporarily become larger than the rate of removal of electrons from the drain, $i_D(t)$. As can be deduced from the quantitative results presented later, this is accomplished by $-i_S(t)$ becoming temporarily larger than $i_T(t)$, and by $i_D(t)$ becoming temporarily smaller than $i_T(t)$, as in Fig. 6.5c. These comments hold for both up-going transitions in the figure.

[†] Recall that negative charges moving from left to right correspond to positive current flowing from right to left.

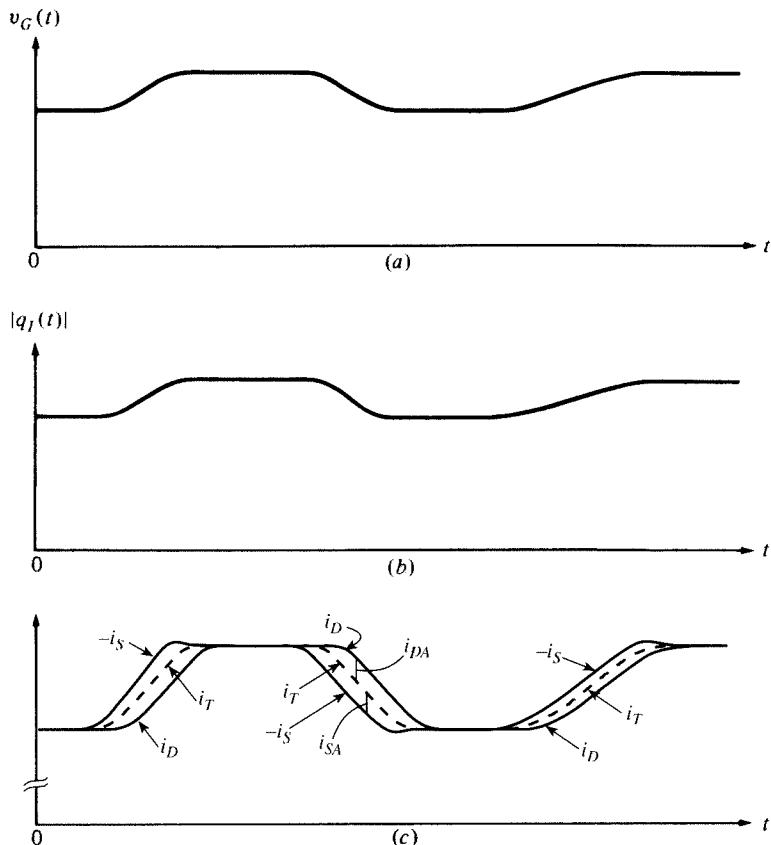


FIGURE 6.5

(a) Gate voltage waveform for the device of Fig. 6.3; all other terminal voltages are assumed fixed; (b) corresponding inversion layer charge magnitude waveform assuming quasi-static operation; (c) total drain current and the negative of the total source current; the transport and charging current components are also shown.

If $v_G(t)$ is *decreasing* instead, the opposite will be true. Here $|q_I(t)|$, and thus the number of electrons in the channel, must decrease. For this to happen, the rate of removal from the drain must temporarily exceed the rate of supply from the source. This is accomplished by $i_D(t)$ temporarily becoming larger than $i_T(t)$, and by $-i_S(t)$ temporarily becoming smaller than $i_T(t)$, as shown in the falling part of the plots in Fig. 6.5c.

The difference between the curve for i_D and the curve for $-i_S$ in Fig. 6.5c must be equal to the total rate of change of q_I , as follows from (6.3.3). By integrating (6.3.3) it can be deduced that the area between the solid lines at each transition is equal to the total change in $|q_I|$. If $|q_I(t)|$ in Fig. 6.5b goes up, down, and up again by equal amounts, the three areas enclosed by solid lines at each transition in Fig. 6.5c will all be equal. However, because the first up-going transition is faster than the second up-going transition, in the former i_D and $-i_S$ need to deviate more from i_T to accomplish the same change in q_I .

From these arguments it follows that, in general, $i_D(t)$ and $i_S(t)$ will differ from their transport values $i_T(t)$ and $-i_T(t)$, respectively. Denoting the differences by $i_{DA}(t)$ and $i_{SA}(t)$, we can write

$$i_D(t) = i_T(t) + i_{DA}(t) \quad (6.3.4a)$$

$$i_S(t) = i_T(t) + i_{SA}(t) \quad (6.3.4b)$$

From these equations and (6.3.3) we have

$$i_{DA}(t) + i_{SA}(t) = \frac{dq_I}{dt} \quad (6.3.5)$$

Thus, we can conveniently view $i_T(t)$ as wholly responsible for the transport effect, and $i_{DA}(t)$, $i_{SA}(t)$ as wholly responsible for changing q_I . (For this reason these two currents are sometimes called “charging” currents.) It should be clear that we cannot identify particular electrons as causing transport current and others as charging the inversion layer. All electrons entering the source can be envisioned to eventually leave through the drain [see paragraph following (6.2.4c)]. However, since $i_D(t) \neq -i_S(t)$, the behavior an external observer sees is indistinguishable from what one *would* have if some of the charge went to contribute to a transport current while the rest went to change q_I , the latter often viewed as “stored” in the inversion layer at any given instant. To make this imaginary picture complete, let us associate $i_{DA}(t)$ and $i_{SA}(t)$ with two *fictitious* charges. If $i_{DA}(t)$ causes the inversion layer charge to change by an amount Δq_D in time Δt , we can write

$$i_{DA}(t) = \frac{dq_D}{dt} \quad (6.3.6a)$$

and, if $i_{SA}(t)$ causes the inversion layer charge to change by an amount Δq_S in time Δt , we can write[†]

$$i_{SA}(t) = \frac{dq_S}{dt} \quad (6.3.6b)$$

Expressions for q_D and q_S will be given shortly. The last three equations give

$$\frac{dq_D}{dt} + \frac{dq_S}{dt} = \frac{dq_I}{dt} \quad (6.3.7)$$

Intuition about the preceding ideas can be increased with the help of the fluid dynamical analog in Fig. 6.4. Assume \hat{v}_G has been fixed for a long time, so that the fluid

[†] The definitions of q_D and q_S here are purposely imprecise. Until further notice, q_D and q_S can be any functions of time, the derivatives of which predict the correct values for $i_{DA}(t)$ and $i_{SA}(t)$, respectively. The reader is urged to avoid assigning physical significance to q_D and q_S at this point. A discussion of this will follow.

has attained a steady-state distribution and the total rate of flow into the channel from the source tank is equal to the rate of flow out of the channel and into the drain tank. Now assume that \hat{v}_G is slowly increased (the piston is slowly moved downward). The total amount of fluid in the channel must also increase, which cannot happen if the two rates of flow continue to be equal. The rate of flow from the source tank will thus temporarily become larger than what would be predicted from “dc considerations,”[†] while the rate of exit from the channel and into the drain tank will temporarily become smaller than what would be predicted from “dc considerations.” (Note by the way that, if the movement of the piston downward is sufficiently slow, then at all times the flow of water is from left to right. Although the rate of flow is smaller at the drain end, no water is seen to flow backward from the drain tank into the channel.) If the piston is slowly raised (\hat{v}_G is decreasing), the total amount of fluid in the channel must decrease. The rate of flow from the source tank will become smaller while the rate of exit into the drain tank will become larger than the rate that would be predicted by “dc considerations.”

Note that dq_S/dt and dq_D/dt in (6.3.6) and (6.3.7) correspond only to the *differences* between the actual and the dc-like values of flow at the source and the drain; they do *not* represent total flow. For example, $dq_D/dt < 0$ should not be interpreted to imply that the drain current is negative. It only implies that the actual drain current $i_D(t)$ is less than the corresponding transport value $i_T(t)$, as can be deduced by using (6.3.6a) in (6.3.4a).

To determine $i_D(t)$ and $i_S(t)$ from (6.3.4) and (6.3.6) we need dq_D/dt and dq_S/dt . Any functions $q_D(t)$ and $q_S(t)$ giving correct time derivatives will do for this purpose. An obvious simple choice is to define these functions so that¹⁶

$$q_D(t) + q_S(t) = q_I(t) \quad (6.3.8)$$

Because of this choice, the following interpretation is sometimes tempting, albeit *not* accurate: “The total charge stored in the inversion layer consists of two components, one that has come through the drain, and one that has come through the source.” The reasons that this interpretation is not accurate are as follows. First, the view of q_I as stored charge leaves much to be desired, as explained in the paragraph following (6.2.4c). Second, as can be deduced from the comment preceding (6.3.8), q_D and q_S are not unique, so it is not justified to assign to them a unique physical significance. Third, in the general case it is not correct to identify q_D and q_S as necessarily “coming” from the drain and source, respectively. For example, in the case illustrated in Fig. 6.5 in conjunction with Fig. 6.3, and provided the changes in $v_G(t)$ are sufficiently slow, electrons travel from the source through the channel to the drain at all times (just like the water in Fig. 6.4). In this picture, no electrons travel backward from the drain into the channel. Thus, all charges “come from the source,” and no charge “comes from the drain.” That does not say anything about the value of q_D , which can still be nonzero and is such that its derivative, from (6.3.6a) and (6.3.4a), is equal to $i_D(t) - i_T(t)$, this difference being indicated by i_{DA} in Fig. 6.5c. It is thus better to think of q_D as the integral of this difference and not to assign a further

[†] The “rate that would be predicted from dc considerations” at time $t = t'$ is what the rate would have been if $\hat{v}_G(t)$ had been frozen for a long time at the value $\hat{v}_G(t')$.

physical significance to it. Similar comments apply to q_S . In fact, a complete development is possible dealing with i_{DA} and i_{SA} directly, without ever defining q_D and q_S . However, in conformance with the literature, we will continue using these quantities, and we will assume that they are defined in such a way that (6.3.8) holds.[†]

Various approaches have been used in the literature for evaluating q_D and q_S , or i_{DA} and i_{SA} .^{15,16,18–21,24–30,34–37,44,70} Here we will adopt an approach that can be rigorously shown to be correct²¹ and that has been demonstrated to agree with experiment. (More on the correctness of the approach will follow later.) We begin by assuming dc operation and define two charges Q_D and Q_S as follows:

$$Q_D = W \int_0^L \frac{x}{L} Q'_I dx \quad (6.3.9a)$$

$$Q_S = W \int_0^L \left(1 - \frac{x}{L}\right) Q'_I dx \quad (6.3.9b)$$

where Q'_I is the inversion layer charge per unit area. Note that in each of these definitions, Q'_I is weighted progressively more heavily for points closer to the corresponding terminal. The sum of Q_D and Q_S is seen to be equal to the total inversion layer charge Q_I [which is given by (6.2.3a)]. The evaluation of the preceding integrals will be considered in Sec. 6.4. For now we only mention that the results will be explicit functions of the terminal voltages:

$$Q_D = f_D(V_D, V_G, V_B, V_S) \quad (6.3.10a)$$

$$Q_S = f_S(V_D, V_G, V_B, V_S) \quad (6.3.10b)$$

If the terminal voltages are allowed to vary, one can evaluate the right-hand sides in (6.3.9) by using $Q'_I(t)$ in lieu of Q'_I . If the variation of the terminal voltages is slow enough so that *quasi-static* operation is maintained, the results of this evaluation will be two quantities, denoted by $q_D(t)$ and $q_S(t)$, given by

$$q_D(t) = f_D(v_D(t), v_G(t), v_B(t), v_S(t)), \quad \text{quasi-static operation} \quad (6.3.11a)$$

$$q_S(t) = f_S(v_D(t), v_G(t), v_B(t), v_S(t)), \quad \text{quasi-static operation} \quad (6.3.11b)$$

where f_D and f_S represent the *same* functions as in (6.3.10). This is consistent with the definition of quasi-static operation given in Sec. 6.2. Thus, $q_D(t)$ and $q_S(t)$ as defined here satisfy (6.3.8). It can now be shown²¹ (Appendix H) that, for the long-channel devices we are considering, the instantaneous currents $i_D(t)$ and $i_S(t)$ in quasi-static operation will be given by (6.3.4) and (6.3.6), with $q_D(t)$ and $q_S(t)$ as given by (6.3.11), and

$$i_T(t) = h_T(v_D(t), v_G(t), v_B(t), v_S(t)) \quad (6.3.12)$$

[†] A possible definition of q_D and q_S is discussed in Appendix H.

where h_T represents the *same* function as in (6.2.2). The proof relies on the “continuity equation,” which is introduced in Sec. 6.7, and carefully considers the current and charge at each point in the channel. The evidence that the preceding approach is correct is overwhelming:

1. It is physically and mathematically sound. This fact can be appreciated if the quantities i_{DA} , i_{SA} , q_D , and q_S are carefully interpreted as explained earlier, and the detailed development²¹ is followed (Appendix H).
2. It agrees with experiment,²¹ as long as the speed of variation is maintained sufficiently low for the quasi-static operation assumption to be valid.
3. Other approaches using perturbation techniques²⁵ give equivalent results.³⁰
4. Non-quasi-static large-scale numerical models, applied in the special case of quasi-static operation, give equivalent results.²⁰
5. Small-signal models derived by using the preceding approach agree with non-quasi-static small-signal models in the special case of quasi-static operation (Chap. 8).

It should be noted that models based on the assumption of quasi-static operation are sometimes abused: they are used to predict currents in very high speed operation, in which the device does not operate quasi-statically. In such cases, the preceding theory is found not to provide satisfactory results (it should not be expected to either).

As follows from the previous discussion, the transport component can be found by using the models of Chap. 4 and will not be considered further here. To simplify our discussion, we will consider only the “charging” current components. We thus concentrate on the following equations:

$$i_{DA}(t) = \frac{dq_D}{dt} \quad (6.3.13a)$$

$$i_G(t) = \frac{dq_G}{dt} \quad (6.3.13b)$$

$$i_B(t) = \frac{dq_B}{dt} \quad (6.3.13c)$$

$$i_{SA}(t) = \frac{dq_S}{dt} \quad (6.3.13d)$$

We now consider the sum of the four terminal currents. Under dc operation, the charges in the device add up to zero, i.e., they satisfy charge neutrality; in fact, this assumption was key in deriving the models of Chap. 4. From the definition of quasi-static operation, shown framed in Sec. 6.2, it follows that in such operation the device remains charge-neutral at each instant in time. Thus, the net rate of charge flowing into it is zero at all times. This means that the total terminal currents satisfy Kirchhoff’s current law:

$$i_D(t) + i_G(t) + i_B(t) + i_S(t) = 0 \quad (6.3.14)$$

In addition, this law will also hold for the charging currents, as is obvious by using (6.3.4) in (6.3.14):

$$i_{DA}(t) + i_G(t) + i_B(t) + i_{SA}(t) = 0 \quad (6.3.15)$$

Using (6.2.5b), (6.2.5c), and (6.3.11) in (6.3.13), and applying the chain rule of differentiation, we obtain

$$i_{DA}(t) = \frac{\partial q_D}{\partial v_D} \frac{dv_D}{dt} + \frac{\partial q_D}{\partial v_G} \frac{dv_G}{dt} + \frac{\partial q_D}{\partial v_B} \frac{dv_B}{dt} + \frac{\partial q_D}{\partial v_S} \frac{dv_S}{dt} \quad (6.3.16a)$$

$$i_G(t) = \frac{\partial q_G}{\partial v_D} \frac{dv_D}{dt} + \frac{\partial q_G}{\partial v_G} \frac{dv_G}{dt} + \frac{\partial q_G}{\partial v_B} \frac{dv_B}{dt} + \frac{\partial q_G}{\partial v_S} \frac{dv_S}{dt} \quad (6.3.16b)$$

$$i_B(t) = \frac{\partial q_B}{\partial v_D} \frac{dv_D}{dt} + \frac{\partial q_B}{\partial v_G} \frac{dv_G}{dt} + \frac{\partial q_B}{\partial v_B} \frac{dv_B}{dt} + \frac{\partial q_B}{\partial v_S} \frac{dv_S}{dt} \quad (6.3.16c)$$

$$i_{SA}(t) = \frac{\partial q_S}{\partial v_D} \frac{dv_D}{dt} + \frac{\partial q_S}{\partial v_G} \frac{dv_G}{dt} + \frac{\partial q_S}{\partial v_B} \frac{dv_B}{dt} + \frac{\partial q_S}{\partial v_S} \frac{dv_S}{dt} \quad (6.3.16d)$$

To evaluate the preceding currents, we need expressions for the charges as functions of the terminal voltages.[†] Such expressions are developed in the following section.

6.4 Evaluation of Intrinsic Charges in Quasi-Static Operation

6.4.1 Introduction

The charge expressions needed to complete the evaluation of the charging currents in (6.3.16) can be developed in a straightforward manner under the assumption of *quasi-static operation* (Sec. 6.2). Thus, we observe that, under this assumption, the expressions for $q_G(t)$, $q_B(t)$, $q_I(t)$, $q_D(t)$, and $q_S(t)$ in terms of $v_G(t)$, $v_B(t)$, $v_S(t)$, and $v_D(t)$ are identical to the expressions for the charges under dc conditions Q_G , Q_B , Q_I , Q_D , and Q_S in terms of the dc voltages V_G , V_B , V_S , and V_D . The latter expressions are given in the form of integrals in (6.2.3) and (6.3.9). If each region of operation is considered separately, these integrals lead to simple functions of the terminal voltages, with the exception of the moderate-inversion region. This is shown in the next

[†] The implementation of (6.3.16) in a computer simulator requires care to avoid results inconsistent with charge conservation.^{27,34–37,44,70}

several subsections. A more general evaluation is also possible, corresponding to the all-region models of Sec. 4.3. This is discussed in Sec. 6.4.5.

The emphasis in this chapter is on illustrating the principles involved; in order to keep the expressions simple, we assume no short-channel effects are present. Using the same principles, such effects can be incorporated at the expense of complexity.^{49,52,53,56}

6.4.2 Strong Inversion

General Expressions for Nonsaturation The integrations indicated in (6.2.3) and (6.3.9) cannot be carried out directly since we do not have the corresponding charges per unit area as functions of x . Accordingly, we first perform a change of variables from the position x to V_{CB} , the strong-inversion “effective reverse bias” of the inversion layer at point x , with respect to the substrate. From (4.7.5), we have

$$dx = -\frac{\mu W}{I_{DSN}} Q'_I dV_{CB} \quad (6.4.1)$$

where I_{DSN} is the drain current in the nonsaturation region. Using the preceding relation, the variable of integration is changed in (6.2.3b) to give

$$\begin{aligned} Q_G &= W \int_{V_{SB}}^{V_{DB}} Q'_G \left(-\frac{\mu W}{I_{DSN}} Q'_I \right) dV_{CB} \\ &= -\frac{\mu W^2}{I_{DSN}} \int_{V_{SB}}^{V_{DB}} Q'_G Q'_I dV_{CB} \end{aligned} \quad (6.4.2a)$$

Similarly, (6.2.3c) becomes

$$Q_B = -\frac{\mu W^2}{I_{DSN}} \int_{V_{SB}}^{V_{DB}} Q'_B Q'_I dV_{CB} \quad (6.4.2b)$$

and (6.2.3a) becomes

$$Q_I = -\frac{\mu W^2}{I_{DSN}} \int_{V_{SB}}^{V_{DB}} Q'^2_I dV_{CB} \quad (6.4.2c)$$

Q_D and Q_S from (6.3.9) become

$$Q_D = -\frac{\mu W^2}{I_{DSN}} \int_{V_{SB}}^{V_{DB}} \frac{x}{L} Q'^2_I dV_{CB} \quad (6.4.3a)$$

$$Q_S = -\frac{\mu W^2}{I_{DSN}} \int_{V_{SB}}^{V_{DB}} \left(1 - \frac{x}{L} \right) Q'^2_I dV_{CB} \quad (6.4.3b)$$

Finally, to express x in (6.4.3) in terms of V_{CB} , we integrate (6.4.1) from $x = 0$ to an arbitrary point in the channel:

$$x = -\frac{\mu W}{I_{DSN}} \int_{V_{SB}}^{V_{CB}} Q'_I dU_{CB} \quad (6.4.4)$$

where U_{CB} is a dummy variable of integration. An equivalent result has been obtained in Sec. 4.7.3. Thus from (4.7.32), we have

$$x = L \frac{H(V_{GB}, V_{SB}, V_{CB})}{H(V_{GB}, V_{SB}, V_{DB})} \quad (6.4.5)$$

where H represents the function multiplying W/L in the nonsaturation drain current expression:

$$I_{DSN} = \frac{W}{L} H(V_{GB}, V_{SB}, V_{DB}) \quad (6.4.6)$$

It is easy to show that (6.4.5) is equivalent to (6.4.4) (Prob. 6.1).

General Expressions Including Saturation Let the value of V_{DB} at which the transistor enters saturation be denoted by V_p . Since the channel is assumed long, if V_{DB} is raised above V_p , the conditions in the channel will remain practically unaffected (Sec. 4.7.1). Thus, for example, let $g_I(V_{GB}, V_{SB}, V_{DB})$ represent the expression giving Q_I in nonsaturation, as it results from performing the integration in (6.4.2c). We will have

$$Q_I = \begin{cases} g_I(V_{GB}, V_{SB}, V_{DB}), & V_{DB} \leq V_p \\ g_I(V_{GB}, V_{SB}, V_p), & V_{DB} > V_p \end{cases} \quad (6.4.7a)$$

$$(6.4.7b)$$

Corresponding relations can be written for Q_G , Q_B , Q_D , and Q_S .

Using the preceding results, expressions of varying complexity can be found for Q_G , Q_B , Q_D , and Q_S depending on the complexity of the model used for Q'_G , Q'_B , Q'_D , Q'_S , and I_{DSN} .

Source-Referenced Simplified Model The computationally efficient simplified model derived in Sec. 4.7.3 resulted in expression (4.7.26) for the drain-to-source current:

$$I_{DS} = I'_{DS}(1 - \eta^2) \quad (6.4.8)$$

where

$$I'_{DS} = \frac{W}{L} \mu C'_{ox} \frac{(V_{GS} - V_T)^2}{2\alpha} \quad (6.4.9)$$

and η was a parameter defined in a way that makes (6.4.8) valid in *both* nonsaturation and saturation:

$$\eta = \begin{cases} 1 - \frac{V_{DS}}{V'_{DS}}, & V_{DS} \leq V'_{DS} \\ 0, & V_{DS} > V'_{DS} \end{cases} \quad (6.4.10a)$$

$$(6.4.10b)$$

where

$$V'_{DS} = \frac{V_{GS} - V_T}{\alpha} \quad (6.4.11)$$

The parameter η has been plotted in Fig. 4.25.

The bulk charge density associated with this model was given by (4.7.17), repeated here for convenience:

$$Q'_B = -C'_{ox} \left[\gamma \sqrt{\phi_0 + V_{SB}} + (\alpha - 1)(V_{CB} - V_{SB}) \right] \quad (6.4.12)$$

Using this equation in (4.7.16), and the definition of the threshold voltage in (4.7.19), we can write the inversion layer charge density as follows:

$$Q'_I = -C'_{ox} [V_{GB} - V_{SB} - V_T - \alpha(V_{CB} - V_{SB})] \quad (6.4.13)$$

We can thus find the corresponding nonsaturation *total* charges Q_B and Q_I by using these in (6.4.2b) and (6.4.2c).²⁹ We can put the results in a form valid in *both* nonsaturation and saturation by using the convenient parameter η defined in (6.4.10). After some algebra, we obtain

$$Q_I = -WLC'_{ox} (V_{GS} - V_T) \frac{2}{3} \frac{1 + \eta + \eta^2}{1 + \eta} \quad (6.4.14)$$

$$Q_B = -WLC'_{ox} \left[\gamma \sqrt{\phi_0 + V_{SB}} + \frac{\alpha - 1}{\alpha} (V_{GS} - V_T) \left(1 - \frac{2}{3} \frac{1 + \eta + \eta^2}{1 + \eta} \right) \right] \quad (6.4.15)$$

The total gate charge can be found from (6.4.2a). However, since Q_I and Q_B have already been found, it is simpler to find Q_G by using the preceding results in the charge neutrality equation:

$$Q_G + Q_o + Q_I + Q_B = 0 \quad (6.4.16)$$

where Q_o is the total equivalent interface charge. The result is

$$Q_G = WLC'_{ox} \left[\frac{V_{GS} - V_T}{\alpha} \left(\alpha - 1 + \frac{2}{3} \frac{1 + \eta + \eta^2}{1 + \eta} \right) + \gamma \sqrt{\phi_0 + V_{SB}} \right] - Q_o \quad (6.4.17)$$

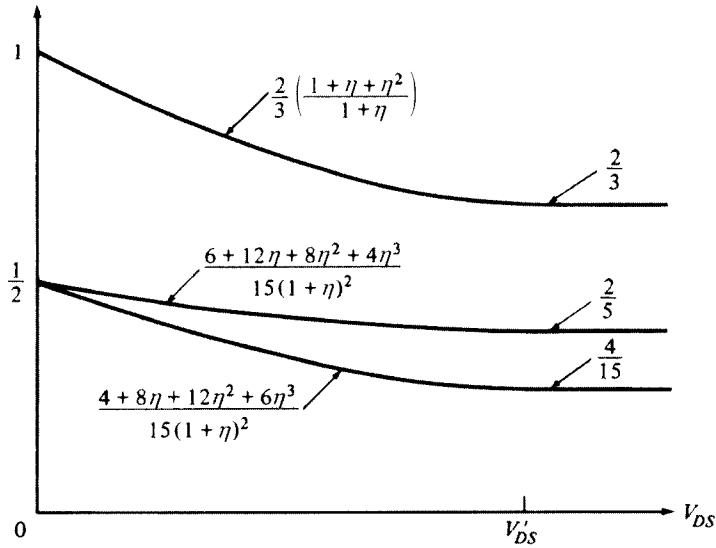


FIGURE 6.6
Quantities used in total intrinsic charge expressions vs. drain-source voltage for fixed values of V_{GS} and V_{SB} .

To find Q_D and Q_S from (6.4.3) we need to relate x to V_{CB} . This can be done by using (6.4.4) or (6.4.5):

$$x = L \frac{(V_{GS} - V_T)(V_{CB} - V_{SB}) - \frac{1}{2}\alpha(V_{CB} - V_{SB})^2}{(V_{GS} - V_T)(V_{DB} - V_{SB}) - \frac{1}{2}\alpha(V_{DB} - V_{SB})^2} \quad (6.4.18)$$

Using this and (6.4.13) in (6.4.3a), and extending the result to cover the saturation region, we obtain

$$Q_D = -WLC'_{\text{ox}}(V_{GS} - V_T) \frac{4 + 8\eta + 12\eta^2 + 6\eta^3}{15(1 + \eta)^2} \quad (6.4.19)$$

Similarly, Q_S can be found by using (6.4.3b). However, it is simpler here to obtain it from $Q_S + Q_D = Q_I$, using (6.4.14) and (6.4.19). The result is

$$Q_S = -WLC'_{\text{ox}}(V_{GS} - V_T) \frac{6 + 12\eta + 8\eta^2 + 4\eta^3}{15(1 + \eta)^2} \quad (6.4.20)$$

Plots of certain quantities appearing in (6.4.14), (6.4.15), (6.4.19), and (6.4.20) are shown in Fig. 6.6. As seen, despite the rather complicated look of the expressions, the form of the plots is rather simple. The reader may want to develop simpler functions of η , which would approximate these plots with good accuracy.

Plots of the total charges as a function of V_{DS} , with V_{GS} as a parameter, are shown in Fig. 6.7.

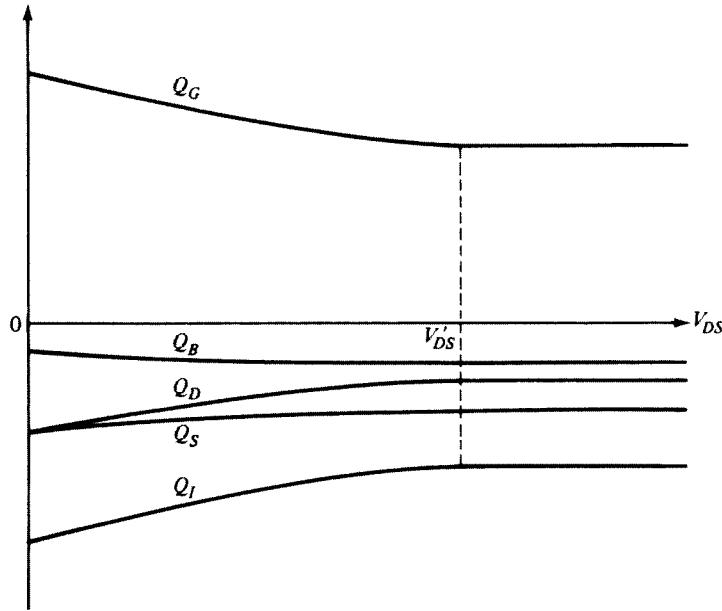


FIGURE 6.7
Total intrinsic charges vs. drain-source voltage for fixed values of V_{GS} and V_{SB} .

Charges at $V_{DS} = 0$ As a check of the previous calculations, let us determine the charges at $V_{DS} = 0$ ($\eta = 1$):

$$Q_B \Big|_{V_{DS}=0} = -WLC'_{\text{ox}} \gamma \sqrt{\phi_0 + V_{SB}} \quad (6.4.21)$$

$$Q_I \Big|_{V_{DS}=0} = -WLC'_{\text{ox}} (V_{GS} - V_T) \quad (6.4.22)$$

$$Q_D \Big|_{V_{DS}=0} = -\frac{WLC'_{\text{ox}} (V_{GS} - V_T)}{2} \quad (6.4.23)$$

$$Q_S \Big|_{V_{DS}=0} = -\frac{WLC'_{\text{ox}} (V_{GS} - V_T)}{2} \quad (6.4.24)$$

$$Q_G \Big|_{V_{DS}=0} = WLC'_{\text{ox}} [(V_{GS} - V_T) + \gamma \sqrt{\phi_0 + V_{SB}}] - Q_o \quad (6.4.25)$$

These equations make sense. Since $V_{DS} = 0$, the depletion region charge per unit area is uniform and is given by (6.4.12) with $V_{CB} = V_{SB}$. Multiplying this by the channel area WL gives (6.4.21). Similarly, (6.4.22) is simply the channel area times the uniform inversion layer charge per unit area as given from (6.4.13), with $V_{CB} = V_{SB}$.

Q_D and Q_S are seen to be half of Q_I each, which makes sense because of symmetry. Finally, (6.4.25) can result from (6.4.16), (6.4.21), and (6.4.22).

Charges in Saturation In the saturation region ($\eta = 0$), we obtain

$$Q_{B,\text{sat}} = -WLC'_{\text{ox}} \left[\gamma \sqrt{\phi_0 + V_{SB}} + \frac{\alpha-1}{3\alpha} (V_{GS} - V_T) \right] \quad (6.4.26)$$

$$Q_{I,\text{sat}} = -\frac{2}{3} WLC'_{\text{ox}} (V_{GS} - V_T) \quad (6.4.27)$$

$$Q_{D,\text{sat}} = -\frac{4}{15} WLC'_{\text{ox}} (V_{GS} - V_T) \quad (6.4.28)$$

$$Q_{S,\text{sat}} = -\frac{2}{5} WLC'_{\text{ox}} (V_{GS} - V_T) \quad (6.4.29)$$

$$Q_{G,\text{sat}} = WLC'_{\text{ox}} \left[\frac{V_{GS} - V_T}{\alpha} \left(\alpha - \frac{1}{3} \right) + \gamma \sqrt{\phi_0 + V_{SB}} \right] - Q_o \quad (6.4.30)$$

As seen, none of the preceding charges depends on the drain voltage. This is a manifestation of the fact that in saturation the drain can no longer have any influence on the intrinsic part of the device owing to pinchoff (excluding short-channel effects).

The influence of one terminal on the charge associated with another is in general nonreciprocal. For example, assume that V_S , V_B , and V_G are fixed and that V_D is varying in the saturation region. Since (6.4.30) is independent of V_D , the gate charge will remain fixed and no transient gate current will be observed. Now, assume that instead V_S , V_B , and V_D are fixed and that V_G is varying, again in the saturation region. From (6.4.28), it is apparent that Q_D will vary. Thus, a nonzero “charging current” will flow through the drain terminal in addition to the conduction current. These facts are apparent from (6.3.16a), (6.3.16b), (6.4.28), and (6.4.30). Such nonreciprocal influence of one terminal on another, although most strongly pronounced in saturation, is also apparent in nonsaturation; it vanishes only at $V_{DS} = 0$. These effects will be considered in more detail in Chap. 8.

We note here that sometimes Q_D is assumed to be zero in the saturation region. This is justified by saying that in this region the channel is isolated from the drain. However, this isolation is only responsible for maintaining i_G independent of $v_D(t)$, as already argued. It is a fact that when $v_G(t)$ changes, $i_D(t)$ will also change, and there is no reason to assume that this change will not include a charging component $i_{DA}(t)$ (Sec. 6.3). In fact, measurements and numerical simulations show that, indeed, $i_{DA}(t)$ can be nonzero in saturation. From (6.3.6a) or (6.3.16a), then, q_D must be such as to give the correct value of $i_{DA}(t)$, which cannot happen if it is identically set to zero. So Q_D is not zero in saturation, although it is independent of v_D . There is no contradiction between “isolation” due to pinchoff and a nonzero value for Q_D , if Q_D is interpreted as explained in Sec. 6.3.

The expressions we have given so far for the charges are adequate for most quasi-static transient response calculations in strong inversion.

Symmetric Model The general equations (6.4.2) and (6.4.3) can be used to develop charge expressions corresponding to the complete symmetric model of Sec. 4.7.1. One must work in this case with the charges per unit area used in developing that model, which are given by (4.7.7) and (4.7.8). The resulting charge expressions have the attractive feature that they are completely symmetric with respect to V_{SB} and V_{DB} , just as was the case for the corresponding model for the current in Sec. 4.7.1. However, these expressions are rather involved.²¹

6.4.3 Moderate Inversion

As was the case for drain current modeling, no simple closed-form expressions have been derived for the charges in the moderate-inversion region in terms of the terminal voltages. Some charge models neglect this region and assume that the weak-inversion and strong-inversion expressions hold in adjacent regions of V_{GS} . The limit point between these regions is often taken to be $V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{SB}}$, which is what we have denoted by V_M (the bottom of the moderate-inversion region). The resulting error in the moderate-inversion charges is not large. However, if it is attempted to differentiate these charges in order to obtain capacitance expressions (Chap. 7), large errors will result as we will see.

To model better the charges in moderate inversion, sometimes empirical expressions are used, resulting in curves that connect smoothly the weak-inversion curves with those for strong inversion. When this is done, it is important to maintain continuity not only of the charges at the transition points but also of the *derivatives* of those charges with respect to the terminal voltages. In this way, the charge expressions will result in continuous capacitance expressions (Chap. 7). As is true with all empirical expressions, it is also important that the parameters used in them be related to process parameters in a correct manner, so that predictions can be made in cases where experimental data are not available.

As will be seen in Sec. 6.4.5, it is possible to derive a model valid in all regions, including moderate inversion as a special case. However, as was the case with general drain current modeling, the price to be paid is computational complexity.

6.4.4 Weak Inversion

In weak inversion, the calculation of the charges is easy. First, we note that the depletion region charge per unit area is given by (4.3.12), repeated here:

$$Q'_B = -\gamma C'_{ox} \sqrt{\psi_s} \quad (6.4.31)$$

As seen in Sec. 4.8, in the weak-inversion region (as well as in depletion), the surface potential ψ_s is practically independent of position, and is given by

$$\psi_s \approx \psi_{sd} = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2 \quad (6.4.32)$$

Thus, Q'_B is also independent of position. The total depletion charge is thus simply WLQ'_B , which gives

$$Q_B = WL C'_{ox} \gamma \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right) \quad (6.4.33)$$

Consider now the charge neutrality equation $Q_G + Q_o + Q_I + Q_B = 0$. For the purposes of calculating Q_G , we can use the fact that in weak inversion $Q_I \ll Q_B$ (Sec. 4.8). Thus

$$Q_G \approx -Q_B - Q_o \quad (6.4.34)$$

To find Q_I accurately, we can use the observation in Sec. 4.8 that Q'_I varies as a straight line with position between its value at the source and its value at the drain, as shown in Fig. 4.29. Thus

$$Q'_I(x) = Q'_{I0} + \frac{x}{L} (Q'_{IL} - Q'_{I0}) \quad (6.4.35)$$

where expressions for Q'_{I0} and Q'_{IL} have been given in Sec. 4.8. Since we have Q'_I as an explicit function of x , we can use it in (6.2.3a) directly to find Q_I . The result is (Prob. 6.7)

$$Q_I = WL \frac{Q'_{I0} + Q'_{IL}}{2} \quad (6.4.36)$$

Similarly, from (6.3.9) and (6.4.35), we obtain (Prob. 6.7)

$$Q_D = WL \left(\frac{Q'_{I0}}{6} + \frac{Q'_{IL}}{3} \right) \quad (6.4.37)$$

$$Q_S = WL \left(\frac{Q'_{I0}}{3} + \frac{Q'_{IL}}{6} \right) \quad (6.4.38)$$

In practice, the previous three charges are often neglected completely in the computation of transients for the following reason. The source and drain junction depletion regions (in the *extrinsic* part of the device, outside the channel area) contain charges that can be evaluated as in Sec. 1.5; see also Sec. 6.8. Of those charges, the ones in the n^+ material must be changed (when V_{SB} and V_{DB} are varied) by electrons supplied through the source and drain terminals. These charges and the corresponding charging currents are much larger than the ones associated with the inversion layer for typical channel lengths. Thus, in weak inversion, the source and drain

charging currents are dominated by the extrinsic part of the device, and the following simplification is often used:

$$Q_I \approx Q_D \approx Q_S \approx 0 \quad (6.4.39)$$

This should not be interpreted to mean that the weak-inversion layer is irrelevant in dynamic operation. In fact, it turns out that in weak inversion, speed is limited by non-quasi-static effects in the inversion layer, rather than by extrinsic parasitics.

6.4.5 All-Region Model

We have seen in Secs. 6.4.2 and 6.4.4 that, if strong inversion or weak inversion is considered separately, simple expressions become possible for the charges in each of the two regions. However, it is also possible to develop a general expression for each charge that will be valid in *all* regions of inversion, just as was done for the drain current in Sec. 4.3. These expressions will even be valid in moderate inversion, a region for which no simple expression has been derived. As usual, considerable complexity must be accepted to make such generality possible.

The principle on which the derivation of general charge expressions is based is simple. The general expression for the inversion layer charge per unit area, as used in Sec. 4.3, is

$$Q'_I = -C'_{ox} \left(V_{GB} - V_{FB} - \psi_s - \gamma \sqrt{\psi_s} \right) \quad (6.4.40)$$

and the corresponding general expression for the depletion region charge per unit area is

$$Q'_B = -\gamma C'_{ox} \sqrt{\psi_s} \quad (6.4.41)$$

Equation (4.3.4) gives the drain-to-source current, assuming both drift and diffusion are present. From this equation we have

$$dx = -\frac{\mu W}{I_{DS}} Q'_I d\psi_s + \frac{\mu W}{I_{DS}} \phi_t dQ'_I \quad (6.4.42)$$

Appropriate use of the preceding three equations in (6.2.3) and (6.3.9) results in general expressions for the total charges Q_I , Q_B , Q_G , Q_D , and Q_S .[†] As an example, by using (6.4.42) in (6.2.3a), we obtain

$$Q_I = -\frac{W^2 \mu}{I_{DS}} \int_{\psi_{t0}}^{\psi_{L}} Q_I'^2 d\psi_s + \frac{W^2 \mu}{I_{DS}} \phi_t \int_{Q_{t0}'}^{Q_L'} dQ_I' \quad (6.4.43a)$$

or

$$Q_I = -\frac{W^2 \mu}{I_{DS}} \int_{\psi_{t0}}^{\psi_{L}} Q_I'^2 d\psi_s + \frac{W^2 \mu}{I_{DS}} \phi_t \frac{1}{2} (Q_L'^2 - Q_{t0}'^2) \quad (6.4.43b)$$

[†] A different, more complicated procedure involves the quasi-Fermi potential in the channel.²⁸

where ψ_{s0} and ψ_{sL} are the surface potentials at the source and the drain ends of the channel, respectively, and I_{DS} is found as in Sec. 4.3. The first integral in (6.4.43b) can be evaluated after expressing Q'_I in terms of ψ_s , using (6.4.40). The result will be in terms of ψ_{s0} and ψ_{sL} , which can be determined from (4.3.15). The quantity in parentheses in (6.4.43b) can similarly be found by expressing Q'_{IL} in terms of ψ_{sL} and Q'_{I0} in terms of ψ_{s0} . Note that ψ_{sL} and ψ_{s0} must be very accurately known, as discussed in Sec. 4.3.

As a check of (6.4.43b) consider the special case of strong inversion in which only drift current is assumed. Then, only the first term will be present in the right-hand side of (6.4.42) (Sec. 4.3). Thus, only the term containing the integral will be present in (6.4.43b). By using in it (4.7.1) and (4.7.4), this term is seen to reduce to (6.4.2c), which was developed specifically for strong inversion. If weak-inversion operation is considered instead, the current is practically all due to diffusion and only the second term will be present in (6.4.42). Thus, only the second term will appear in (6.4.43b). By using in it I_{DS} from (4.8.3), this term is seen to reduce to (6.4.36), which was developed specifically for weak inversion. In moderate inversion, both terms in (6.4.43b) will be significant.

The rest of the charges can be similarly found, starting from (6.2.3b), (6.2.3c), (6.3.9a), and (6.3.9b), and can be expressed in terms of ψ_{s0} and ψ_{sL} . The procedure is rather long but mathematically straightforward and involves changes of variables and integration (Prob. 6.8).[†]

Simplified All-Region Models Alternatively, the expression in (6.4.43b) can be evaluated in a manner consistent with the simplified all-region models of Sec. 4.4 (Prob. 6.9). Consider, for example, the model in (4.4.22b), repeated here:^{57,61}

$$I_{DS} = \frac{W}{L} \mu \left[\frac{-1}{2\alpha C_{ox}'} (Q''_{IL} - Q''_{I0}) + \phi_t (Q'_{IL} - Q'_{I0}) \right] \quad (6.4.44)$$

The development of this model in Sec. 4.4.4 can be traced to the observation⁴⁷ that, for a given V_{GB} , Q_I varies almost linearly with ψ_s . This resulted in (4.4.15), repeated here:

$$\frac{\partial Q'_I}{\partial \psi_s} = \alpha C_{ox}' \quad (6.4.45)$$

This relation can be used to change the variable of integration in (6.4.43b) from ψ_s to Q'_I . This results in the following expressions for the total charge Q_I :

$$Q_I = \frac{W^2 \mu}{I_{DS}} \left[-\frac{Q''_{IL} - Q''_{I0}}{3\alpha C_{ox}'} + \frac{\phi_t}{2} (Q'_{IL} - Q'_{I0}) \right] \quad (6.4.46)$$

[†] We note that, for implementation in a computer program, these expressions may have to be modified since they are not numerically robust. For example, as ψ_{sL} approaches ψ_{s0} , both the numerator and the denominator in some of the expressions may approach zero.

Using (6.4.44) in this expression and removing the common factor $(Q'_{IL} - Q'_{I0})$ from both numerator and denominator, we obtain^{57,61}

$$Q_I = WL \frac{\frac{2}{3}(Q'^2_{IL} + Q'_{IL}Q'_{I0} + Q'^2_{I0}) - \alpha\phi_t C'_{ox} (Q'_{IL} + Q'_{I0})}{Q'_{IL} + Q'_{I0} - 2\alpha\phi_t C'_{ox}} \quad (6.4.47)$$

which can be evaluated for known bias voltages, by relating Q'_{IL} to ψ_{sL} and Q'_{I0} to ψ_{s0} , as before. This expression gives Q_I in all regions of inversion. As a check, consider strong inversion and operation in saturation. Then the last terms in the numerator and denominator of (6.4.47) (the development of which terms can be traced to the assumption that diffusion currents are present) will be negligible, and Q'_{IL} will be zero; thus (6.4.47) gives $Q_I = \frac{2}{3}(WL)Q'_{I0}$. With the help of (4.7.7) and (4.7.8), this is easily seen to reduce to (6.4.27).

Similarly, one can develop expressions for the other charges, starting from (6.2.3b), (6.2.3c), (6.3.9a) and (6.3.9b). Such development, for the simplified general models we have discussed in Sec. 4.4, is discussed in the literature.^{57,61,72-76} Results for the body-referenced simplified general model of Sec. 4.4.2, which employed the concept of symmetric linearization, can be found on the book's website, www.oup.com/us/tsividis_mcandrew

Expressions have also been proposed using interpolation semiempirical models.⁵⁸

6.4.6 Depletion and Accumulation

In digital circuits, transistors are switched between conduction and, ideally, cutoff.[†] Therefore, the charges in the latter state are of importance in calculating the transient response of such circuits. The cutoff region consists of two regions—depletion and accumulation. In both of these, the inversion layer charge is totally negligible (Sec. 2.4):

$$Q_I = 0 \quad (6.4.48)$$

The bulk charge density, which therefore equals the total silicon charge density Q'_C , is related to the surface potential needed to support this charge by (2.5.1). The surface potential is positive in depletion, resulting in negative bulk charge; the surface potential is instead negative in accumulation, resulting in the accumulation of positively charged holes. From the potential balance equation, (2.3.1), we have, for the oxide potential, $\psi_{ox} = V_{GB} - \psi_s - \phi_{MS}$. The gate charge per unit area, $Q'_G = C'_{ox}\psi_{ox}$, multiplied by the gate area, gives

$$\begin{aligned} Q_G &= WLC'_{ox} (V_{GB} - \psi_s - \phi_{MS}) \\ &= WLC'_{ox} (V_{GB} - V_{FB} - \psi_s) - Q_o \end{aligned} \quad (6.4.49)$$

[†] In some processes, use of a low supply voltage has led to the use of low threshold voltage, in order to provide sufficient current in the on condition. Thus, when the gate-source voltage is made equal to 0, the device may be in weak inversion rather than cutoff.

where we have used the definition of the flatband voltage in (2.2.6). The bulk charge Q_C can now be found from the charge balance equation (2.3.3):

$$Q_C = -Q_G - Q_o \quad (6.4.50)$$

In some of the literature, the presence of Q_o in the above equations is overlooked. The charges can be found by numerically solving the above equations together with (2.5.1). Simple results become possible deep in depletion and deep in accumulation (i.e., as long as V_{GB} is not too close to V_{FB}), as discussed below.

Deep in Depletion Deep in depletion (i.e., for ψ_s above zero by several ϕ_i), the situation is similar to that in weak inversion in Sec. 6.4.4, where the negligible inversion layer charge played no role in calculating Q_B and Q_G ; we thus get the same equations, repeated here for convenience:

$$Q_B = -WLC'_{ox} \gamma \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right) \quad (6.4.51)$$

$$Q_G = -Q_B - Q_o \quad (6.4.52)$$

Deep in Accumulation Deep in accumulation (i.e., for ψ_s below zero by several ϕ_i), ψ_s varies little with V_{GB} (see Fig. 2.12); it can then be seen from (6.4.49) that the plot of Q_G vs. V_{GB} becomes approximately a straight line. From (6.4.50), the same will be true for the plot of Q_C vs. V_{GB} .

6.4.7 Plots of Charges vs. V_{GS}

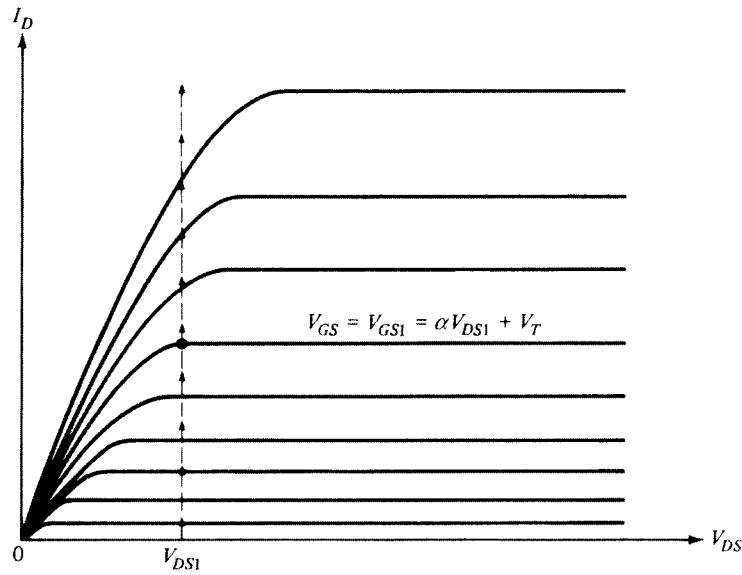
In Fig. 6.7, we have seen the charges plotted as a function of V_{DS} , for a fixed V_{GS} in strong inversion. To show the behavior of the charges in all regions of operation, we can plot the various charges vs. V_{GS} for a given fixed V_{DS} . This corresponds to moving up a vertical line in the I_{DS} - V_{DS} characteristics, as shown in Fig. 6.8. The regions encountered as V_{GS} increases are accumulation, depletion, weak inversion, moderate inversion, strong-inversion saturation, and strong-inversion nonsaturation, in that order. Recall that, for the approximate strong-inversion model, nonsaturation is defined by

$$\text{Nonsaturation: } V_{DS} \leq \frac{V_{GS} - V_T}{\alpha} \quad (6.4.53)$$

Thus, for a given $V_{DS} = V_{DS1}$, the range of V_{GS} for which nonsaturation is obtained is

$$\text{Nonsaturation: } V_{GS} \geq \alpha V_{DS} + V_T \quad (6.4.54)$$

The critical value of the right-hand side of this equation is marked with a filled circle in Fig. 6.8. The charges vs. V_{GS} for the fixed value of $V_{DS} = V_{DS1}$ are shown in Fig. 6.9.

**FIGURE 6.8**

Drain current vs. drain-source voltage, with gate-source voltage as a parameter. V_{DS1} is a specific drain-source voltage value for which the plots of Fig. 6.9 are obtained.

6.4.8 Use of Intrinsic Charges in Evaluating the Terminal Currents

The expressions we have developed for the charges in individual regions of operation can be written as functions of V_D , V_G , V_B , and V_S by substituting in them $V_{GS} = V_G - V_S$, $V_{DS} = V_D - V_S$, and $V_{SB} = V_S - V_B$. As discussed in Sec. 6.4.1, if the terminal voltages vary slowly enough so that quasi-static operation is maintained, the same expressions can be used to determine the time-varying charges. Thus the partial derivatives in (6.3.16) can be determined. In other words, we have

$$\frac{\partial q_K}{\partial v_L} = \left. \frac{\partial Q_K}{\partial V_L} \right|_{V_L=v_L}, \quad \text{quasi-static operation} \quad (6.4.55)$$

where each of K and L can stand for D , G , B , or S . Thus, if the variation of terminal voltages with time is known, the terminal currents can be found from (6.3.4), (6.3.12), (6.3.16), and (6.4.55).

A similar approach can be taken if the all-region model is used (Sec. 6.4.5), since the surface potentials ψ_{s0} and ψ_{sL} can be related to V_{SB} and V_{DB} , respectively, through (4.3.15). The algebra, however, is considerably more complicated.

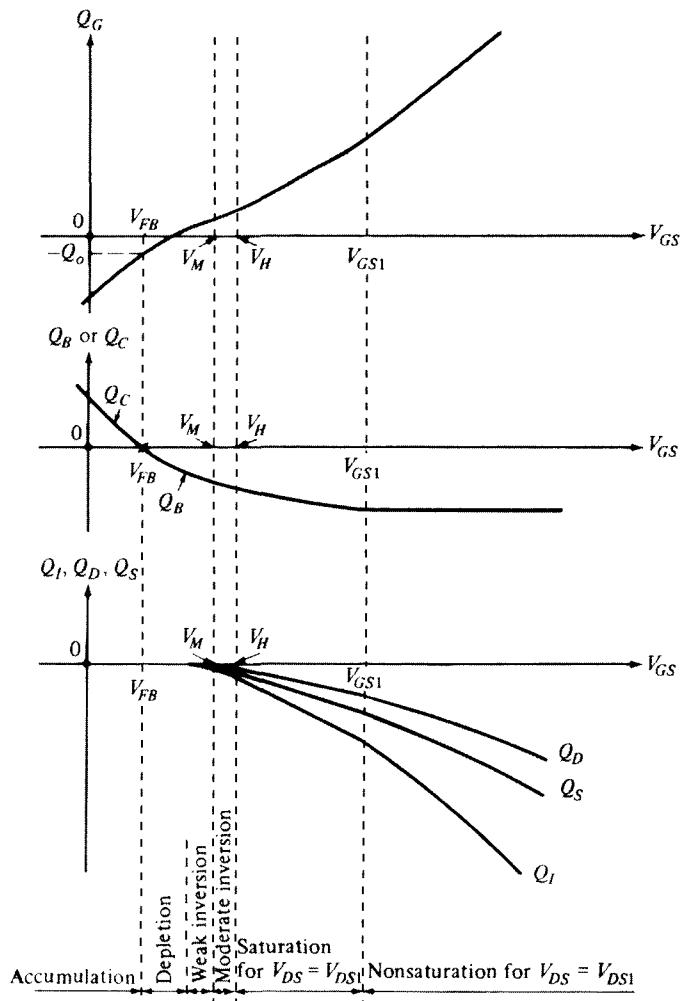


FIGURE 6.9
Total intrinsic charges vs. gate-source voltage for a specific value of $V_{DS} = V_{DS1}$, as indicated in Fig. 6.8.

6.5 Transit Time under DC Conditions

The transit time (Sec. 1.3.1) associated with dc operation in a transistor is the average time it takes for an electron to travel the length of the channel:

$$\tau = \frac{|Q_I|}{I_{DS}} \quad (6.5.1)$$

Having evaluated the inversion layer charge in the previous section, this is a convenient time to calculate τ for four cases of interest. We will make use of τ in the following section.

1. *Strong-inversion nonsaturation with very small V_{DS} .* From (6.4.22), we have $|Q_I| \approx C'_{ox}WL(V_{GS} - V_T)$, and from (4.7.24a), with very small V_{DS} , $I_{DS} \approx \mu C'_{ox}(W/L)(V_{GS} - V_T)V_{DS}$. Hence, from (6.5.1),

$$\boxed{\tau \approx \frac{L^2}{\mu V_{DS}}} \quad (6.5.2)$$

Note that, since with negligible V_{DS} the channel is approximately uniform and the drift velocity approximately fixed, we could have used (1.3.11) to obtain this result.

2. *Strong-inversion saturation.* Assume that no velocity saturation takes place. From (6.4.27), $|Q_I| = \frac{2}{3}WLC'_{ox}(V_{GS} - V_T)$, and from (4.7.24b), $I_{DS} = \frac{1}{2}\mu C'_{ox}(W/L)(V_{GS} - V_T)^2/\alpha$. We thus obtain, from (6.5.1),

$$\boxed{\tau = \frac{4}{3}\tau_o} \quad (6.5.3)$$

where

$$\boxed{\tau_o = \frac{\alpha L^2}{\mu(V_{GS} - V_T)}} \quad (6.5.4)$$

For a device with $\mu = 400 \text{ cm}^2/(\text{V}\cdot\text{s})$ and $L = 0.2 \text{ }\mu\text{m}$, operated with $(V_{GS} - V_T) = 1 \text{ V}$ and with $\alpha = 1.1$, the value of τ is 1.1 ps.

3. *Weak inversion with $V_{DS} > 5\phi_t$.* Here, from Sec. 4.8, we easily see that $Q'_{II} \approx 0$, and from (6.4.36), $|Q_I| = \frac{1}{2}|Q'_{I0}|WL$. From (4.8.10), $I_{DS} = \mu(W/L)\phi_t|Q'_{I0}|$. Hence

$$\boxed{\tau = \frac{L^2}{\mu(2\phi_t)}} \quad (6.5.5)$$

With $\mu = 400 \text{ cm}^2/(\text{V}\cdot\text{s})$ and $L = 0.2 \text{ }\mu\text{m}$, τ is about 19 ps. Note that in all three cases, the transit time is proportional to the *square* of L . This is so because Q_I is proportional to L , and I_{DS} is inversely proportional to L . To see this effect another way, consider the case of strong-inversion nonsaturation with very small V_{DS} . The channel is nearly uniform and the field is everywhere approximately equal to V_{DS}/L . The drift velocity of the electrons is proportional to this field. Increasing L , say, m times decreases the drift velocity m times and these electrons now have to travel m times the distance. So the time it takes them to travel the length of the channel will increase m^2 times. Applying the simple

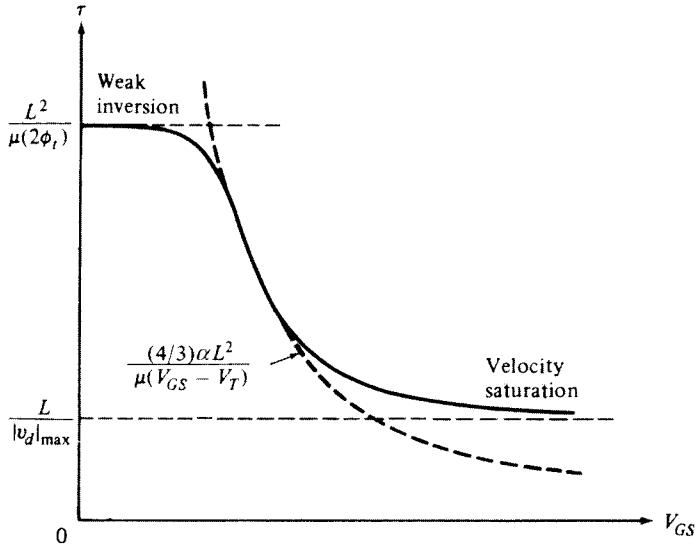


FIGURE 6.10
Transit time vs. gate-source voltage for operation in the “flat” part of the I_{DS} - V_{DS} characteristics.

model of Sec. 1.3.3 to weak inversion, a similar argument holds, only there the “driving force” is not the field but rather the gradient of the charge in the channel, $|Q'_{10}|/L$.

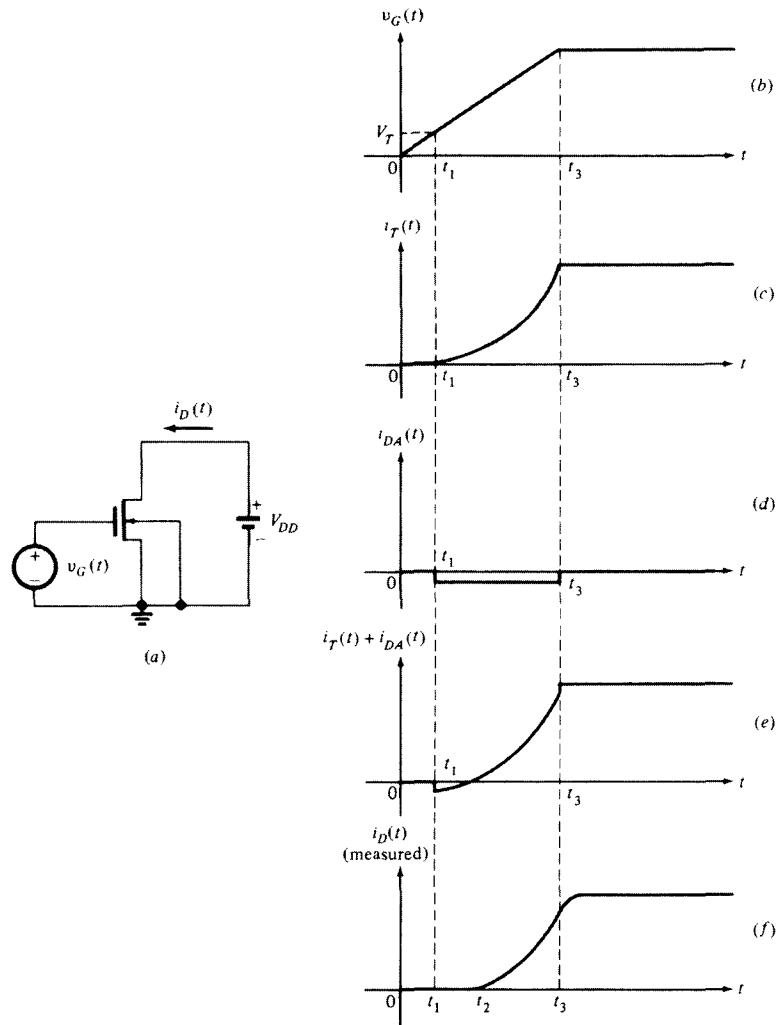
4. *Velocity saturation.* If velocity saturation is present over part of the channel, the previous arguments break down.⁴⁰ Although the value of the transit time in this case can be calculated by using material from Sec. 5.2, we will limit our discussion to a simple estimate. We note that τ will be larger than what one would have if the electrons were moving at maximum speed over *all* of the channel. Thus

$$\tau > \frac{L}{|v_d|_{\max}} \quad (6.5.6)$$

Figure 6.10 shows the transit time of a device operating in the “flat” part of the I_{DS} - V_{DS} characteristics as a function of V_{GS} (solid line). As V_{GS} is increased, V'_{DS} increases (Sec. 4.7.3) and V_{DS} must be raised if the device is to be kept in the saturation region. The increase in V'_{DS} is accompanied by an increased electric field and, thus, if L is small, velocity saturation sets in over much of the channel. Thus, τ cannot be decreased at will by increasing V_{GS} , as one could have concluded by carelessly applying (6.5.3) and (6.5.4).

6.6 Limitations of the Quasi-Static Model

From our discussion so far, we expect that the quasi-static model will be valid if the terminal voltages vary sufficiently slowly so that the charge distribution in the

**FIGURE 6.11**

(a) A transistor with varying excitation; (b) gate-source voltage; (c) transport current calculated from dc model; (d) drain charging current calculated by assuming quasi-static operation; (e) sum of (c) and (d); (f) form of actual $i_D(t)$ observed in practice. A sufficiently large V_{DD} is assumed, so that even the maximum value of v_G corresponds to operation in the saturation region. Extrinsic effects are assumed negligible.

channel can follow with negligible inertia. A quantitative definition of the term "sufficiently slowly" is difficult to come by. Whether results obtained from using the quasi-static model are trustworthy or not depends on the type of voltage waveforms applied to the terminals, on the regions of operation involved, on the type of result desired (current waveform shape, delay, rise time, etc.), on the accuracy sought, etc. In practice, some rules of thumb have been developed semiempirically, by

using a simple basic case shown in Fig. 6.11a. Here only v_G is varying, as shown in Fig. 6.11b. We will consider only the intrinsic device effects. The drain current contains a transport part $i_T(t)$ and a charging part $i_{DA}(t)$, as in (6.3.4a):

$$i_D(t) = i_T(t) + i_{DA}(t) \quad (6.6.1)$$

For approximate calculations in digital circuit applications, it is often assumed that a transistor is in the off state if $v_{GS} < V_T$ and in strong inversion if $v_{GS} > V_T$. Using this simplification here implies that the device goes abruptly from off to strong-inversion saturation operation at $t = t_1$. We will also assume that V_{DD} is large enough so that the device never goes into nonsaturation. The conductive part of the current can be found from any dc model, such as (4.7.24b), and is shown in Fig. 6.11c. The charging current $i_{DA}(t)$ can be found from (6.3.16a). Since v_D , v_S , and v_B are constant, this equation gives

$$i_{DA}(t) = \frac{\partial q_D}{\partial v_G} \frac{dv_G}{dt} \quad (6.6.2)$$

In strong inversion saturation, the value of q_D in quasi-static operation $Q_{D,\text{sat}}$ is given by (6.4.28). Thus, $\partial q_D / \partial v_G$ is $-\frac{4}{15} WLC'_{ox}$, a negative constant independent of v_G . (In a more general case, though, one can expect the partial derivatives $\partial q_k / \partial v_l$ to be functions of the terminal voltages.) Since dv_G/dt is constant for the rising part of v_G , $i_{DA}(t)$ is of the form shown in Fig. 6.11d. Adding the two drain current components then produces $i_D(t)$, as shown in Fig. 6.11e.

Measurements of $i_D(t)$ reveal a waveform of the type shown in Fig. 6.11f assuming negligible extrinsic effects.^{19–21} This is seen to differ from our “quasi-static” result of Fig. 6.11e, notably in two respects:

1. For some time after t_1 , the quasi-static model predicts a negative drain current, which is not observed in practice *once extrinsic device effects are subtracted out* (Prob. 6.11). Instead, the drain current remains at zero until some time t_2 , as shown in Fig. 6.11f. This can be explained as follows. For $t < t_1$, the channel is empty. At $t = t_1$, the conditions in the channel become favorable for electrons; the latter enter the channel through the source and move toward the drain. However, no drain current will be observed until the electrons *reach* the drain. This happens at time t_2 , as shown in Fig. 6.11f. In contrast to this picture, the quasi-static model assumes that at *any* time t' after t_1 the channel contains electrons *throughout* its length, as would be the case if v_G had been frozen for a long time at the value $v_G(t')$, i.e., the nonzero time it takes for the electrons to reach the drain is not considered.
2. At $t = t_3$, $i_D(t)$ in Fig. 6.11e is predicted to jump to its dc steady-state value instantaneously. This is a consequence of the assumption that the charge adjusts itself with no inertia. In reality this is not the case, as illustrated in Fig. 6.11f.

The quasi-static model is thus seen to fail when it comes to predicting the fine details of the drain current waveform. However, for many applications this is of no

great consequence, since the fine waveform details are often of no great interest. This is, for example, the case in much of digital circuit design, where quantities of interest are initial values, final values, output rise or fall times, delays between output and input, etc. Comparisons to measurements and to non-quasi-static numerical results have been used to suggest that, for much of digital circuit work, the quasi-static model can be used with acceptable results if the rise time t_R of the waveforms[†] involved satisfies the condition^{20,21}

$$t_R > 20\tau_o \quad (6.6.3)$$

where τ_o is given by (6.5.4) by using the maximum value of V_{GS} from Fig. 6.11b, assuming no velocity saturation is present. It is emphasized that this is only a *rough* rule of thumb.[‡] For example, depending on the application, the factor of 20 could instead be 15 or 25. As a numerical example of the order of magnitude involved here, consider a device with $\mu = 400 \text{ cm}^2/(\text{V}\cdot\text{s})$, $V_T = 0.3 \text{ V}$, $L = 0.2 \mu\text{m}$, $\alpha = 1.1$, and $V_{GS,\max} = 6 \text{ V}$. Then (6.6.3) requires $t_R > 22 \text{ ps}$.

In common fabrication technologies, in which the transistor exhibits significant extrinsic parasitic capacitances, the preceding limit is not restrictive. The speed of operation is slowed down in these technologies because of such “stray” capacitances, and internal waveforms on a chip will often have rise times larger than this limit. On top of this, even if occasionally a gate voltage waveform has a rise time somewhat below this limit, the total transient currents and delays due to the stray capacitance of the device being driven can be significant and can mask the errors due to intrinsic effects predicted by quasi-static models. This is not true with technologies that achieve very low stray capacitance values, such as SOI technologies (Sec. 1.7). In such cases, as well as in cases where pushing the speed limit is attempted by using bulk technologies, results obtained using quasi-static models should be looked at with suspicion.

The Issue of Drain/Source Charge Partition We have seen that q_D and q_S can be evaluated from (6.3.9a) and (6.3.9b). These relations satisfy $q_D + q_S = q_I$, but the ratio q_D/q_S depends on the bias voltages. Consider, for example, operation in strong-inversion saturation. Then the previous relations lead to (6.4.28) and (6.4.29). Comparing these with (6.4.27), we see that q_D is 40 percent, and q_S 60 percent, of the total charge q_I , which is often referred to as a “40/60 partition.” Other partitions have also been proposed. One of these is 50/50, which is clearly not appropriate for the nonsymmetric situation that exists in the saturation region. Another partition used is 0/100, which means $q_S = q_I$ and $q_D = 0$. However, the claim that $q_D = 0$ has been disputed in the two paragraphs following (6.4.30). This claim is sometimes defended by pointing to the fact that, if q_D is taken identically zero, $i_{DA}(t)$ from (6.6.2) will be zero, too; thus no negative current region will occur in $i_D(t)$ plots, which agrees with

[†] Defined for this purpose as $t_3 - t_1$ in Fig. 6.11b.

[‡] The origin of this result will be understood after non-quasi-static analysis is described in Sec. 6.7.

measurements, unlike the case in Fig. 6.11e, which was obtained using a 40/60 partition. Note, however, four things:

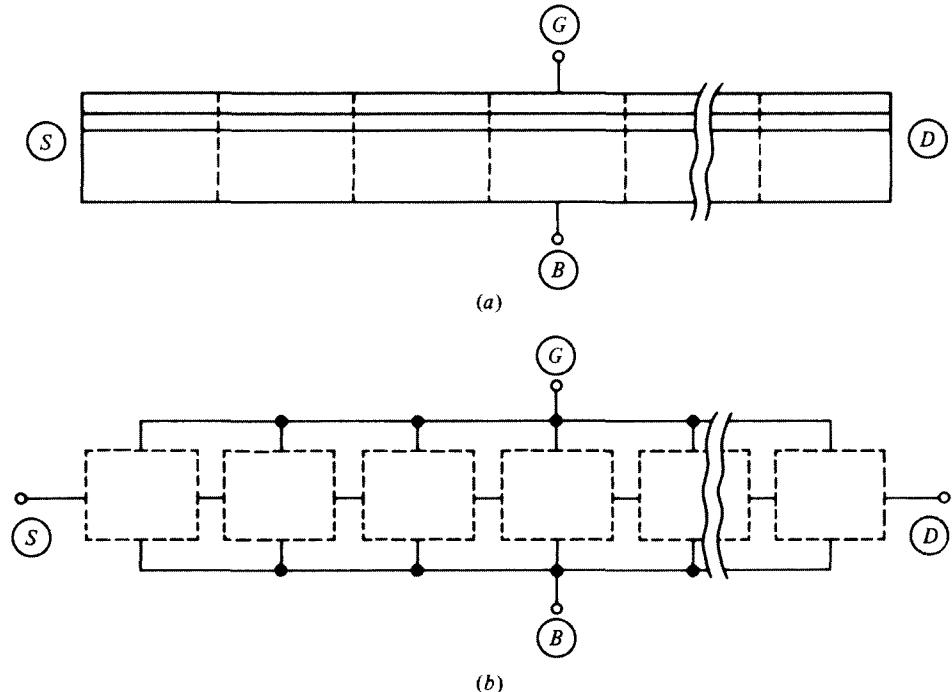
1. It is the negative i_{CA} that makes the positive part of the $i_D(t)$ plot in Fig. 6.11e resemble what is measured in practice (compare with Fig. 6.11f). If q_D , and thus i_{DA} , were zero, i_D would be identical to i_T in Fig. 6.11c, which is not necessarily an improvement. Thus, in Fig. 6.11, which of the two waveforms in c and e is a better approximation to f will depend on the application.
2. Although it is true that in the case of Fig. 6.11a no negative i_D would normally be observed, this would not necessarily be the case if V_{DD} were lower. Indeed, in nonsaturation, negative drain current transients are possible, and this has been verified by measurements.[†] Such currents would not be possible to predict if q_D were taken to be identically zero.
3. A 0/100 choice may ensure that i_D flows in only one direction, but it does not ensure this for i_S .
4. Notice that i_{DA} depends on the rate of change of v_G , as seen in (6.6.2). It can be shown that if (6.6.3) is satisfied, the negative excursion in the i_D plot is limited to only a couple of percent of the eventual, maximum value of i_D . Equation (6.6.2) will only predict a negative current with a large magnitude if dv_G/dt is large, in which (6.6.3) is violated. However, in such cases, the quasi-static model should not be used in the first place!

Thus, to conclude, we can state the following: *for rates of change at which the assumption of quasi-static operation is valid*, (6.3.9) is the correct choice for q_D and q_S . Several facts in support of this claim have been listed following (6.3.12).

Multisegment Models One way to model a transistor at speeds where its quasi-static model breaks down is to view it as consisting of several sections, each section being short enough to be modeled quasi-statically. This idea is shown in Fig. 6.12. Each box in Fig. 6.12b is assumed to be a “subtransistor,” with its own imaginary source and drain *points*. Of course, with the exception of the left and right extremes, these subtransistors are assumed to consist only of intrinsic parts. In other words, no extrinsic source and drain *regions* at intermediate points are assumed to exist.

Other Effects Note that in our previous discussion, we have not considered short-channel effects. The latter can greatly complicate the picture.^{19–21,38,45–53,55,56} Among the phenomena observed in them, but not in truly long-channel devices, are a *transient transport current*²¹ (in addition to the transient charging current) and velocity saturation that renders (6.4.1) and the relations based on it invalid. Two-dimensional

[†] If V_{DD} were smaller, so that the maximum value of v_G corresponded to the *nonsaturation* region, we would have a situation analogous that in Fig. 1.34c. Here, as the piston moved downward fast and stopped at the position shown, fluid would initially flow into the “channel” not only from the source but also from the drain. This would correspond to a *negative* drain current. After things settled, a regular flow toward the right would have to be established, so the drain current would go through zero and then would become positive. These predictions are verified by actual measurements on transistors.

**FIGURE 6.12**

(a) A long-channel transistor separated into several sections; (b) model for (a); each box represents a quasi-static model corresponding to one section in (a).

numerical simulation is a valuable tool in this case.^{19,20} Finally, there is another phenomenon, not related to short-channel effects, that we have not considered in our modeling. This phenomenon is observed during the falling part of a gate voltage waveform. As $v_G(t)$ is decreased, the magnitude of the inversion layer charge must be reduced. To this end, electrons exit through the drain and source terminals. The dynamics of this removal process show that there is a finite removal capability associated with it.⁷ If v_G is decreased too fast, the ensuing “bottlenecks” at either end of the channel result in some electrons being temporarily “trapped” in the channel. The field-induced junction consisting of the inversion layer and the substrate can then become momentarily forward-biased and the electrons can cross into the bulk. There they recombine with holes and cause a substrate current to flow. This phenomenon is called *charge pumping*. It is more pronounced for shorter falling times of the gate voltage waveform, as might be expected intuitively.^{20,59}

We conclude this section by reminding the reader that parasitic elements associated with the *extrinsic* part of the device can significantly alter the behavior of the transistor, compared with that predicted here. Such elements include the gate-source and gate-drain overlap capacitances, and the substrate-source and substrate-drain junction capacitances. In addition, the fact that the substrate is not a perfect conductor can become important. The significant resistance associated with the latter, in conjunction

with the intrinsic and extrinsic substrate capacitances, can affect the dynamic performance of the device. Extrinsic elements will be considered in Sec. 6.8.

6.7 Non-Quasi-Static Modeling

6.7.1 Introduction

In the previous section, it was seen that, for a given channel length, the quasi-static model breaks down if the input changes too fast. It was suggested that one way to extend the validity of quasi-static models in that case is to consider the device as a connection of several shorter devices (Fig. 6.12) and to model each section quasi-statically. Note that, for each section, the current entering one end is, in general, different from the current leaving the other end during transients. This accounts for possible inversion layer charge buildup within each section, just as different magnitudes of drain and source current are encountered because of charge buildup in the single-section quasi-static model. The faster the change of the input, the shorter each section must be, and the greater the number of sections. In the limit, one can let the section length approach zero, thus making the total number of sections approach infinity. The resulting model would then not be subjected to the speed limitations of quasi-static models. We will develop this idea formally in this section.^{1–4,6,9,12,13,17,19,20,22,32,54,62–69,71} Note that, following the previous argument concerning currents, we should allow for the current to be a function of position x along the channel as well as time:

$$i = i(x,t) \quad (6.7.1)$$

Similarly, the inversion layer charge per unit area will be a function of position and time:

$$q'_I = q'_I(x,t) \quad (6.7.2)$$

Although such a generality will allow us to extend the range of speed over which we can obtain useful results, we will assume that the device dimensions are much smaller than the wavelength corresponding to the frequency of operation, and that basic electrostatic laws still hold, only with constant voltages and charges replaced by their time-varying counterparts. All charge-potential relations at each point in the channel are still assumed to be valid. The speeds involved are assumed such that magnetic effects do not have to be taken into account. The device is still considered to be overall charge neutral.[†]

[†] Consider a transistor driven by four ideal voltage sources, as in Fig. 6.3. Kirchhoff's current law is satisfied for this combination of ideal sources, thus leading to (6.3.14) again. Thus, the net current entering the device is assumed to always be zero for reasons external to the device, and the device remains electrically neutral at all times. More general cases, involving driving circuits that are not ideal (e.g., including other transistors, also operating non-quasi-statically), are not considered here. The combined systems, operating at arbitrarily high speeds, can properly be analyzed by involving the full Maxwell's equations. This is not normally done for circuit simulations, in which charge neutrality is assumed to hold for individual devices, as previously indicated.

6.7.2 The Continuity Equation

Consider a section of the inversion layer of very small length Δx , as shown in Fig. 6.13. Let the current entering on the right be in general different from that leaving from the left by a small amount Δi , as shown. The total charge entering on the right in a small time interval Δt is $(i + \Delta i)\Delta t$; the total charge leaving from the left in the same amount of time is $i\Delta t$. Thus, in the interval Δt , the charge inside the chunk must be increasing by $(i + \Delta i)\Delta t - i\Delta t = \Delta i\Delta t$. The corresponding increase $\Delta q'_l$ in the inversion charge *per unit area* will simply be the total charge increase divided by the chunk's area, as seen above. Thus,

$$\Delta q'_l = \frac{\Delta i\Delta t}{W\Delta x} \quad (6.7.3)$$

which can also be written as

$$\frac{\Delta i}{\Delta x} = W \frac{\Delta q'_l}{\Delta t} \quad (6.7.4)$$

We now let the finite differences approach zero. The left-hand side then becomes the partial derivative of $i(x, t)$ with respect to x . Similarly, the fraction on the right becomes the partial derivative of $q'_l(x, t)$ with respect to t . Thus,

$$\frac{\partial i(x, t)}{\partial x} = W \frac{\partial q'_l(x, t)}{\partial t} \quad (6.7.5)$$

This equation is referred to as the *continuity equation*.[†] It is simply a way to express charge conservation for a chunk of infinitesimal length. Note that if q'_l does

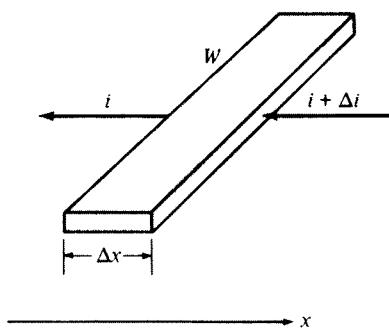


FIGURE 6.13
A chunk of the inversion layer of small length Δx .

[†] The reader may have encountered this equation with i defined in the opposite direction from that in Fig. 6.13, in which case a minus sign would appear in one side of the equation. The above form of the continuity equation is appropriate for our purposes. Note also that, while in the case we are considering only electrons are assumed to be present as usual, in more general cases the continuity equation remains valid if the current and charge in it are taken to include the effect of both electrons and holes. If separate equations are written for each carrier, additional terms must be included to account for carrier generation and recombination.³³

not change with time ($\partial q'_I / \partial t = 0$), the preceding equation gives $\partial i / \partial x = 0$, that is, i then has a constant value independent of position x . This is because there is no charge buildup, and thus the current exiting must be equal to the current entering. That is the case under dc conditions. In fact, the constancy of i was instrumental in developing the dc I_D equations in Chap. 4.

6.7.3 Non-Quasi-Static Analysis

Non-quasi-static analysis of the MOS transistor is a difficult mathematical exercise.^{2,6,12,13,17,22,32,54,62–69,71} We will illustrate it for the special case in which all points in the channel are in *strong inversion*, which simplifies matters considerably. We can relate $q'_I(x, t)$ to the external terminal voltages and the internal effective reverse bias $V_{CB}(x, t)$ by the time-varying version of (4.7.7)–(4.7.8):

$$q'_I(x, t) = -C'_{ox} \left[v_{GB}(t) - V_{FB} - \phi_0 - v_{CB}(x, t) - \gamma \sqrt{\phi_0 + v_{CB}(x, t)} \right] \quad (6.7.6a)$$

The variation of $v_{CB}(x, t)$ with x is the “driving force” for the current flow. This is expressed from (4.7.5) with, of course, I_{DS} replaced by $i(x, t)$:

$$i(x, t) = -\mu W q'_I(x, t) \frac{\partial v_{CB}(x, t)}{\partial x} \quad (6.7.6b)$$

Finally, the continuity equation developed in the previous subsection is

$$\frac{\partial i(x, t)}{\partial x} = W \frac{\partial q'_I(x, t)}{\partial t} \quad (6.7.6c)$$

Equations (6.7.6) constitute a system of three equations in three unknowns: $q'_I(x, t)$, $i(x, t)$, and $v_{CB}(x, t)$.[†] The last two equations express basic facts about current flow. They have both been developed by considering a chunk of material with mobile electrons in it. Apart from the convenient symbols used and the assumption of only drift current in (6.7.6b), the development was independent of whether the material is part of an MOS structure. The MOS transistor physics only enters in the first equation.

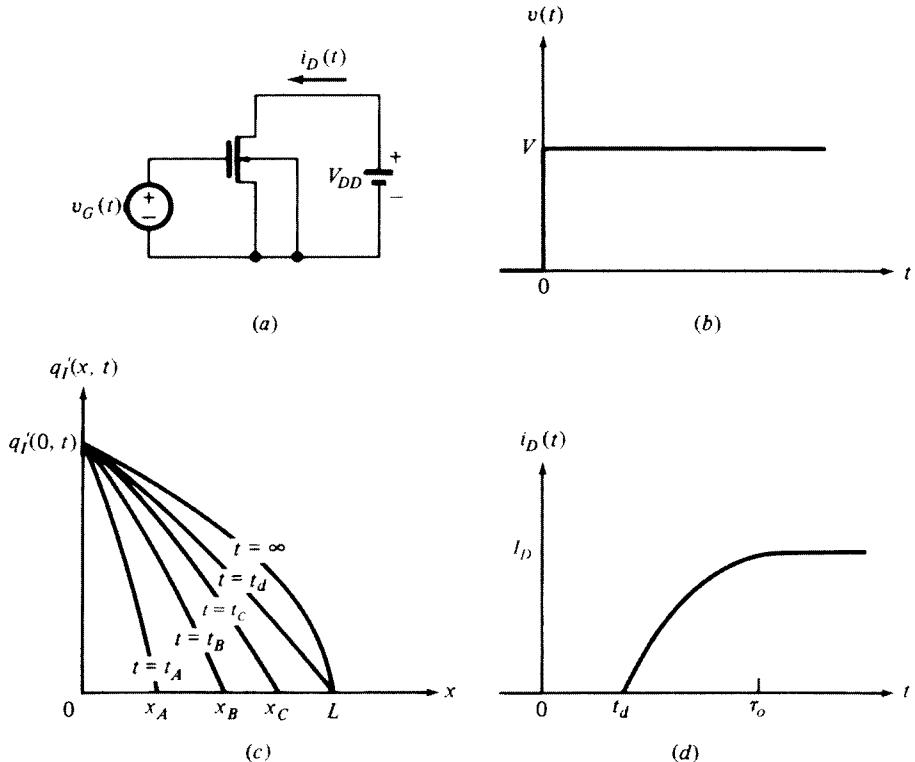
[†] We remind the reader that the analysis presented assumed that the channel is in strong inversion. In the more general case, corresponding to the model of Sec. 4.3, we must allow for diffusion currents. Thus (6.7.6a) must be replaced by [see (4.3.13)]

$$q'_I(x, t) = -C'_{ox} \left[v_{GB}(t) - V_{FB} - \psi_s(x, t) - \gamma \sqrt{\psi_s(x, t)} \right]$$

where $\psi_s(x, t)$ is the surface potential at position x and at time t . Similarly, (6.7.6b) must be replaced by [see (4.3.4)]

$$i(x, t) = -\mu W q'_I(x, t) \frac{\partial \psi_s(x, t)}{\partial x} + \mu W \phi_t \frac{\partial q'_I(x, t)}{\partial x}$$

or, alternatively, (6.7.6b) can be used, with v_{CB} taken to be the “quasi-Fermi potential difference” (Sec. 4.5). Equation (6.7.6c) is general and remains unchanged.

**FIGURE 6.14**

(a) A transistor with a step excitation; (b) gate-source voltage; (c) inversion layer charge per unit area as a function of position along the channel with time as a parameter; (d) drain current as a function of time. A sufficiently large V_{DD} is assumed, so that even the maximum value V of $v_G(t)$ corresponds to operation in the saturation region.

The solution of (6.7.6) requires a set of *initial and boundary conditions*. These will depend on the terminal voltages. As an example, consider the circuit of Fig. 6.14a with the step input shown in Fig. 6.14b.⁶ The device is assumed to have settled in the off condition before the positive step is applied. The values of the step V and of V_{DD} are such that, after a transient period, the device settles in the saturation region. Since the device is initially off, q'_I will be zero at $t = 0$ everywhere in the channel:

$$q'_I(x, 0) = 0 \quad (6.7.7a)$$

With the large value $v_{GB}(t) = V$ applied at $t = 0$, the source end of the channel ($x = 0$) is assumed to reach strong inversion immediately after $t = 0$. Since at that end $v_{CB} = v_{SB} = 0$, (6.7.6a) gives, for all positive time

$$q'_I(0, t) = -C'_{ox} (V - V_{FB} - \phi_0 - \gamma\sqrt{\phi_0}) \quad (6.7.7b)$$

Finally, at the drain end ($x = L$), q'_I is zero whether the device is off or in saturation (“pinchoff” assumption). Thus

$$q'_I(L,t) = 0 \quad (6.7.7c)$$

The system (6.7.6) with the conditions (6.7.7) can now, in principle, be solved by using partial differential equation techniques to provide the distribution of i , v_{CB} , and q'_I with position and time. The drain and source currents can be determined from the solution for $i(x, t)$ by noting that

$$i_D(t) = i(L,t) \quad (6.7.8)$$

$$i_S(t) = -i(0,t) \quad (6.7.9)$$

From the solution for $v_{CB}(x,t)$, one can determine $q'_B(x,t)$ and integrate it with respect to position to find the total instantaneous depletion region charge as in (6.2.3c). The substrate transient current can then be determined as in (6.3.2). The gate transient current can be found in a similar manner. We remind the reader that extrinsic parasitics are not considered here. In a real device, substrate and gate extrinsic resistance (Sec. 6.8) can change the picture.

Unfortunately, the actual details of the solution outlined above are complicated. The results presented in the literature are obtained by using numerical techniques. We will not present these long procedures here,^{6,12,19,20,22,32,54,62–69} but will only summarize the most important results. Solution of (6.7.6) for the circuit of Fig. 6.14a results in $q'_I(x,t)$, as shown in Fig. 6.14c. At $t = 0$, the channel is empty. At $t = t_A$, the electrons coming from the source have reached up to point $x = x_A$; hence, q'_I is zero beyond this point. The wavefront of electrons continues moving to the right as shown, and reaches the drain at $t = t_d$; t_d will be referred to as the *delay time*. At the instant $t = t_d$, the channel charge has not reached steady state yet. Steady state is reached asymptotically and corresponds to the curve marked $t = \infty$. Note that a quasi-static model implicitly assumes instead that this distribution is reached instantly at $t = 0^+$ (“an instant after” $t = 0$).

The drain current as a function of time is shown in Fig. 6.14d. It is zero up to $t = t_d$, at which time the electrons reach the drain. Current then begins to flow and, eventually, builds up to the value I_D , which is the value calculated from dc equations. [In contrast to the drain current, the source current starts flowing immediately at $t = 0^+$, since electrons start filling the channel from the source end as soon as $v(t)$ goes high, as expected by the fluid analog in Fig. 1.34d.]

Assuming q'_I corresponds to the approximate model of Sec. 4.7.3 with $\alpha = 1$, a numerical solution of the above problem gives for the delay time⁶

$$t_d \approx 0.38\tau_o \quad (6.7.10)$$

where

$$\tau_o = \frac{L^2}{\mu(V_{GS} - V_T)}, \quad \alpha = 1 \quad (6.7.11)$$

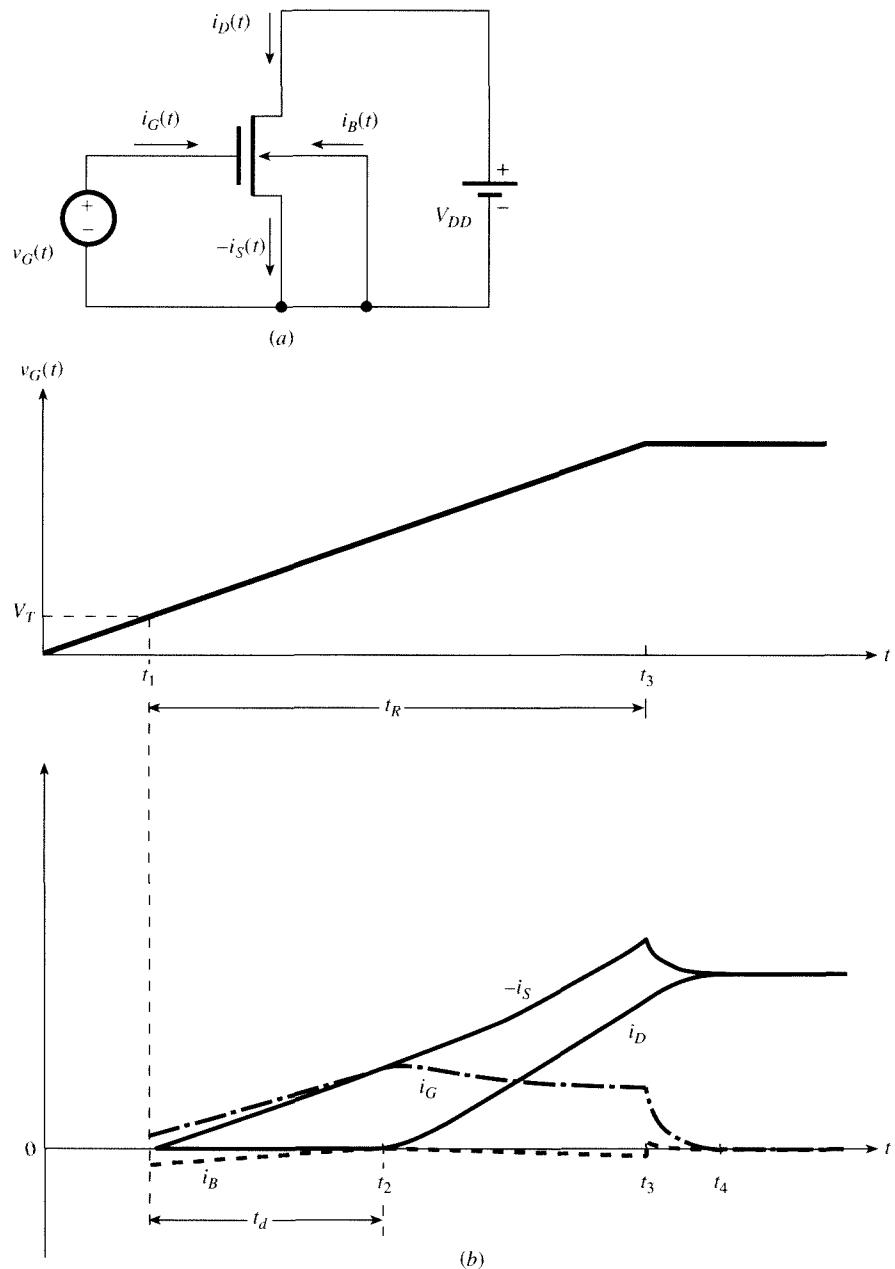
with V_{GS} being the value of the input for $t > 0$. The same numerical solution predicts that at $t = \tau_o$ the current has reached about 98 percent of the final value I_D . Note that the quantity τ_o in the above equation is the same as in (6.5.4), which was derived for dc conditions. However, the fact that the same quantity is involved in the two different types of analysis does not mean that one can casually use the dc transit time to explain directly the non-quasi-static behavior of the transistor. Care is required at this point.

We now briefly consider what happens when the above circuit is driven by an input with significant risetime t_R .^{12,19,20,22,32,54,64,66,68,69,71} In this simplified analysis, we will assume again that the transistor is off until a time $t = t_1$, when $v_{GS} < V_T$, and operates in strong inversion when $v_{GS} > V_T$. The value of V_{DD} is again assumed to be large enough to keep the device in saturation, even when the input attains its maximum value. The definitions of various currents are shown in Fig. 6.15a. In Fig. 6.15b, we show these currents vs. time for an input with t_R equal to several times τ_o . As seen, $-i_S$ (the current leaving the source) starts flowing as soon as the device turns on and is positive; positive current leaving the source is equivalent to negative current entering it; the channel of the device is been “charged” with mobile electrons. The current $-i_S$ soon rises to a value that is almost balanced by the gate current i_G ; the small difference seen is due to the substrate current, which is also shown in the figure. At time t_2 , the electrons reach the drain, and drain current begins to flow. This current approximately balances further increases in $-i_S$, and i_G does not increase anymore. Since the channel must be charged further as v_G rises further, $-i_S$ continues to be larger than i_D . When the input stops rising at $t = t_3$, the channel is nearly charged, but not quite, since the inversion layer charge distribution has not reached steady state yet; thus $-i_S$ continues to be larger than i_D for a little while longer. However, $-i_S$ and i_D both gradually tend to their steady-state value, determined by the maximum value of v_G using dc equations. At $t = t_4$, we have approximately $i_D = -i_S$ and $i_G = i_B = 0$; thus, we have practically reached dc steady state.

In Fig. 6.15c, the input rise time t_R has been reduced to a value of about τ_0 . The waveforms of the currents in this case should be compared with those in Fig. 6.15b. As seen, in Fig. 6.15c there is no time for i_G to taper off. Also, the larger value of dv_G/dt causes large charging currents (as would be the case even for a simple capacitor). Again, initially $-i_S$ rises to a value close to that of i_G , save for a small difference due to i_B . Once the maximum input value is reached, the behavior is qualitatively similar to that in Fig. 6.15b. Here, though, the mobile electrons happen to reach the drain at a time t_2 that is about equal to the time the input reaches its maximum value. Note that t_2 is smaller in Fig. 6.15c, since larger v_G values are available earlier than in Fig. 6.15b, and these help charge the channel faster.

As can be seen from the previous example, the delay time t_d depends on the rise time t_R . As long as t_R is much less than τ_o , the observations given earlier for the case of the step input are approximately valid. However, if t_R is significant (larger than τ_o), the delay time is given approximately by $t_d = \sqrt{t_R \tau_o}$, as found by another numerical solution. If t_R is over $20\tau_o$, the numerical solution of (6.7.6) gives roughly the same results as the quasi-static model. This is the reason why the limit of validity for the latter is expressed as in (6.6.3).

The preceding results have been derived for a long-channel device. If, instead, L is small, velocity saturation can occur, and electrons will be traveling at maximum

**FIGURE 6.15**

(a) A transistor with a varying $v_G(t)$ and a fixed drain voltage (the latter is assumed to be large enough for the device to be in saturation, even with v_G at its maximum value); the currents plotted in the other parts of the figure are defined as shown here. (b) Waveforms for t_R equal to several times τ_o .

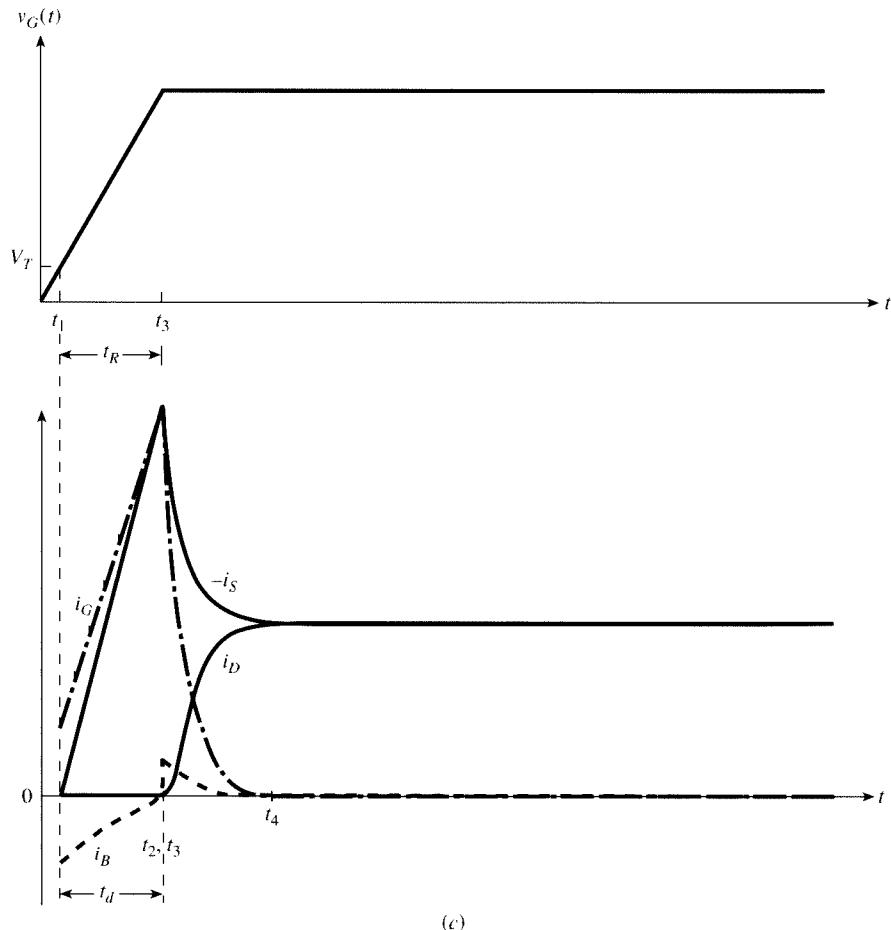


FIGURE 6.15 (Continued)

(c) Waveforms for t_R approximately equal to τ_o . Extrinsic capacitances are assumed zero in order to clearly reveal the intrinsic transient behavior.

speed $|v_d|_{\max}$ toward the drain. If it is assumed that this happens over the whole length of the channel, the delay time will be, for a step input,

$$\tau_d \approx \frac{L}{|v_d|_{\max}} \quad (6.7.12)$$

which can be significantly *larger* than what long-channel theory would predict.^{19,20} On the other hand, if the rise time of the input is significantly larger than the preceding limit, and saturation velocity is no longer the limiting factor, the delay is found to be less than that predicted by long-channel theory. One may attribute this to the fact that in short-channel devices the drain and source act also as gates, owing to two-dimensional effects (Sec. 5.4). These “gates” have already been activated before

$t = 0$ in Fig. 6.14. Hence, the main gate does not have to start building up the whole inversion layer “from scratch.” This heuristic explanation is supported by accurate numerical calculations.²⁰

The preceding analysis emphasized the principles of non-quasi-static analysis. For circuit simulation, the solution of full non-quasi-static equations is computationally complex and inefficient. Thus, simplified approaches have been developed.^{73,77–82} A summary of such approaches is given in Sec. 10.4.

6.8 Extrinsic Parasitics

So far in this chapter, we have dealt with the “intrinsic” part of the transistor, indicated in Fig. 6.1. This is the part responsible for transistor action. However, the rest of the transistor, called the “extrinsic” part, forms an inseparable whole with the intrinsic part, and it has to be taken into account. The extrinsic part is there, to a large extent, in order to connect the intrinsic part to other devices on an integrated circuit. The extrinsic part affects the overall behavior of the device, generally *decreasing* performance. The various components of the extrinsic part, called “extrinsic parasitics” (or just “extrinsics” or “parasitics” for short), include the source-bulk and drain-bulk diodes; series resistance for all of the drain, source, gate, and bulk terminals; and additional capacitances not associated with the intrinsic MOS charges. In general, the relative importance of the deleterious effect of these parasitics on overall device or circuit performance has increased as device dimensions have decreased with the historic evolution of CMOS technologies.

In Fig. 6.16, which is admittedly a bit scary, we show the intrinsic part with a transistor symbol in the center. Everything else forms the extrinsic part. We now discuss the various extrinsic components separately. In what follows, we will use the subscript e to denote extrinsic elements and associated values.

6.8.1 Extrinsic Capacitances

Capacitances The charge storage effects associated with the extrinsic part could be modeled by giving the corresponding charge-voltage expressions, as we have done for the intrinsic part. It is common, though, to model extrinsic charges in terms of the corresponding capacitances. This is not a fundamental distinction; rather, it is a matter of practice. To see this, consider a two-terminal nonlinear charge storage element (e.g., a reverse-biased pn junction) and denote the charge on one of its sides by $q(v)$, where $v = v(t)$ is the voltage across the element. Ignoring junction leakage, the current entering that side will be

$$i(t) = \frac{dq}{dt} = \frac{dq}{dv} \frac{dv}{dt} \quad (6.8.1)$$

If we define

$$C(v) \equiv \frac{dq}{dv} \quad (6.8.2)$$

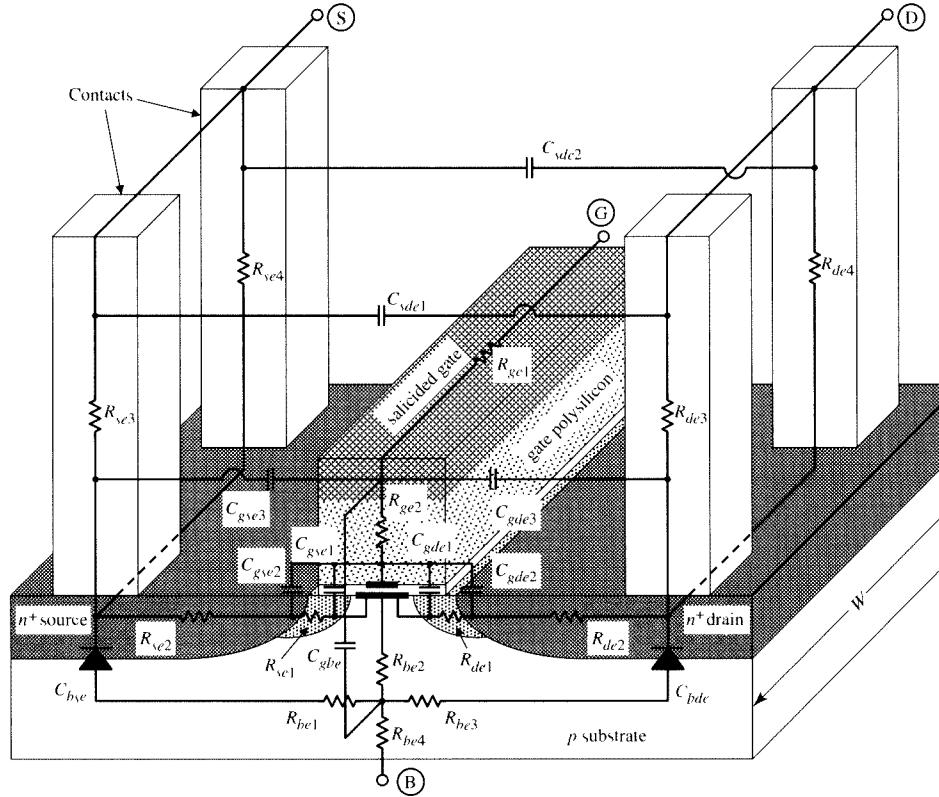


FIGURE 6.16
Extrinsic parasitics.

this relation can be written

$$i(t) = C(v) \frac{dv}{dt} \quad (6.8.3)$$

The quantity representation $C(v)$ is called the “incremental capacitance” of the element. This has the same form as the conventional relation for a linear capacitor, $i(t) = C(dv/dt)$, only here the capacitance, rather than being constant, is a function of the voltage. The last relation makes it clear that, in the general case, the current in a nonlinear capacitor depends both on the rate of change of the voltage (through the second factor on the right-hand side), and on the voltage itself (through the first factor). Clearly, the current can be expressed in terms of charge, as in (6.8.1), or in terms of capacitance, as in (6.8.3). We will use capacitances below.[†]

The various MOS transistor extrinsic capacitances are shown in Fig. 6.16. Each of these can be associated with parts of the physical transistor structure, as discussed

[†] Despite this equivalence in principle, numerical computation in circuit simulators is best done in terms of charges; this is discussed in Sec. 10.4.1.

in the following subsections. We will start from the center of the figure and then continue outward.

Overlap Gate-Source and Gate-Drain Capacitances The capacitances C_{gse1} and C_{gde1} in Fig. 6.16 are due to the overlap of the gate over the source/drain extensions; they are called *overlap* capacitances. Neglecting edge effects on either side of the channel, these dielectric capacitances are proportional to the channel width W , which is indicated on the right in Fig. 6.16. These capacitances have a bias dependence⁸³ because of the modulation of the overlap region charge density by the gate voltage.[†]

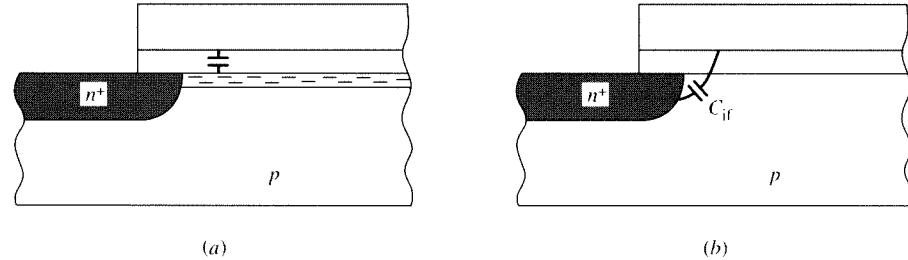
Fringing Gate-Source and Gate-Drain Capacitances The capacitances C_{gse2} and C_{gde2} are due to the fringing field between the source and drain regions and the sides of the gate. Again, neglecting edge effects on the sides of the channel, these capacitances are proportional to the gate width W .

Gate-Body Capacitance The capacitance C_{gbe} is a parasitic capacitance between gate and substrate along the channel length, outside the immediate channel area. This capacitance is proportional to the channel length L , and one must count the contributions of both sides of the channel.

Junction Capacitances and Currents The body-source and body-drain junctions, shown as diodes in Fig. 6.16, exhibit charge storage effects associated with their depletion regions. Each of these junctions has three distinct physical regions: the bottom of the source and drain regions (the “area” component), the side of the junction adjacent to the gate edge (the “gate sidewall” component), and the side of the junction not adjacent to the gate that abuts the isolation region. For careful modeling, a separate junction capacitance component must be included for each of these regions; however, for simplicity, only single components, C_{bse} and C_{bde} are shown in the figure. Details of various junction models are described in Ref. 84. The simplest use components such as (1.5.2) for the junction current and (1.5.27) for the junction capacitance, with each component appropriately scaled by the junction area, the gate edge length (which is W), or the length of the outside perimeter of the source and drain regions. For accurate modeling of junction leakage currents, which can be significant in modern transistors and is especially important for floating body SOI devices, the ideal exponential model (1.5.2) is not sufficient. A detailed junction model is available elsewhere.⁸⁵

Contact-Related Capacitances The tall contacts to the source and drain in Fig. 6.16 have parasitic capacitances as shown. These are between gate and source (C_{gse3}), gate and drain (C_{gde3}), and source and drain (C_{sde1} and C_{sde2}). These capacitances depend on the details of the layout, including the numbers of contacts. Models used for them often come from three-dimensional electromagnetic simulations that are verified against measurements.

[†] The plates of these capacitances are typically in accumulation, and the associated charges can be found along the lines of Sec. 6.4.6, except that L is replaced by the length of the overlap region and the body is replaced by the n^+ regions; thus the role of V_{GB} is played by V_{GS} for the source and V_{GD} for the drain.

**FIGURE 6.17**

Intrinsic capacitances near the source. (a) In the presence of a strong-inversion layer; (b) inner fringing capacitance in the absence of a strong-inversion layer.

Body-Substrate Capacitance If the transistor is inside a well on a CMOS chip (Sec. 1.7), then one must consider also the capacitance due to the pn junction between the well and the common substrate on which the well has been formed. This capacitance, which we will denote by $C_{bb'}$, is not shown in the figure. It is obvious that, if a group of more than one transistor share the same well, $C_{bb'}$ must be included only once in the model for the group.

Inner Fringing Capacitance An effect that can be noticeable in short-channel devices is illustrated in Fig. 6.17,⁸⁶ where the situation near the source is shown. In Fig. 6.17a, a strong-inversion layer is assumed formed, and a capacitance between it and the gate exists; this gives rise to intrinsic charge storage effects, which have already been discussed. In Fig. 6.17b, the device is assumed to be off. Here the absence of an inversion layer allows the gate to “see” the inner wall of the source region directly. This gives rise to an “inner fringing” capacitance. A similar capacitance exists at the drain end, when the inversion layer does not “screen” the gate from the inner wall of the drain, i.e., at cutoff and in saturation. As a device is driven from deep nonsaturation into saturation, the effect of the inversion layer near the drain goes from fully screening the gate from the inner wall, to only partially screening it; thus, the effective inner fringing capacitance is really a function of how far we are from saturation and depends on bias. The inner fringing capacitance exists also in long-channel devices, but is usually small in comparison to the other capacitances. The inner fringing capacitance is often neglected.

Although the preceding capacitances can be estimated using two- or three-dimensional electrostatic simulations, in practice they are often determined by measurements.

6.8.2 Extrinsic Resistances

Source and Drain Series Resistances There are several components to the source and drain series resistances,^{87–91} as shown in Fig. 6.16: contact resistance (R_{se3} , R_{se4} , R_{de3} , R_{de4} , which are assumed to include the resistance at the point of contact between the metal contact material and the silicon source/drain), resistance in the main source and drain regions (R_{se2} and R_{de2}), and resistance in the source-drain extension regions (R_{se1} and R_{de1}). The last component is voltage-dependent. This is

because the gate bias affects the condition at the surface of the n^+ extension regions, which is typically in accumulation. In addition, the body-source and body-drain voltages modulate the width of the associated depletion regions and thus vary the cross-sectional area through which the current must pass in these extension regions.[†] Empirical models are often used for such voltage dependence.

Consider the source series resistance. If the resistance per contact is R_C and there are N_C contacts to the source, the corresponding contact resistances are assumed to be in parallel (which is, strictly speaking, not true, as can be seen in Fig. 6.16), contributing a total resistance R_C/N_C . For the main source/drain regions, if the sheet resistance is $R_{SH,\text{main}}$ and the spacing between the edge of the regions and the source/drain contacts is d_{main} , we have, using (1.3.16), a resistance of $R_{SH,\text{main}}(d_{\text{main}}/W)$. Similarly, for the source/drain extensions, if the corresponding quantities are $R_{SH,\text{ext}}$ and d_{ext} , we have a resistance of $R_{SH,\text{ext}}(d_{\text{ext}}/W)$ (the product $R_{SH,\text{ext}}d_{\text{ext}}$ is often provided as a single value). The total source extrinsic resistance is then modeled as follows⁹²

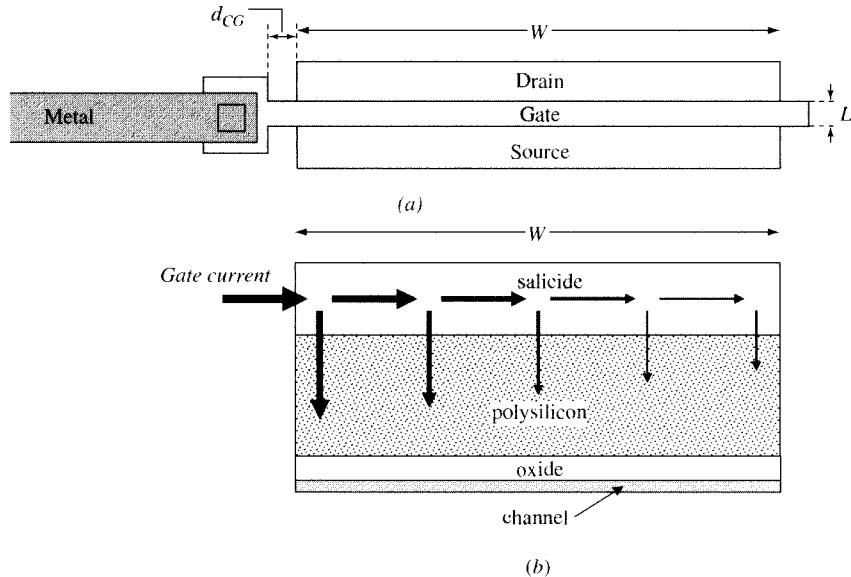
$$R_{Se} = \frac{R_C}{N_C} + R_{SH,\text{main}} \frac{d_{\text{main}}}{W} + R_{SH,\text{ext}} \frac{d_{\text{ext}}}{W} \left(p_B V_{SB} + \frac{1}{1 + p_G(V_{GS} - V_{FB,\text{ext}})} \right) \quad (6.8.4)$$

where p_B and p_G are empirical parameters that model the bulk and gate voltage dependencies of the extension region resistance, and $V_{FB,\text{ext}}$ is the flatband voltage of the extension regions. An analogous form holds for the drain extrinsic resistance R_{De} . It goes without saying that a lot of approximations have been used in the preceding formula, so the result cannot be expected to be very accurate. For example, the effects of current spreading out of the contacts and into the source region are not considered.^{87–91} The preceding formula is a reasonable compromise between accuracy and complexity.

The effect of the source and drain series resistances on the dc characteristics of the transistor has already been discussed in Sec. 4.12. However, these resistances also affect the transient behavior of the transistor, which is why they are discussed in this chapter.

Gate Resistance Modeling of the gate series resistance is important for high-speed digital and RF design (even if there is no or little dc gate current, there can be significant gate transient current flow under large-signal excitation, or ac current under small-signal ac excitation). As can be seen from Fig. 6.16, there are two components to the gate resistance: one from the salicided top region of the gate (R_{ge1}), and one from the vertical portion of the polysilicon gate (R_{ge2}), which is not salicided and therefore has a higher resistivity than the salicided portion.^{93,94} To a first approximation, gate current flows laterally through the salicided portion and vertically through the nonsalicided portion. In reality, the current flow is complicated and proper evaluation of the gate resistance requires three-dimensional simulation. To obtain a simple estimate, approximations must be used, as discussed in the following.

[†] There is a similar dependence of the main source/drain regions on junction bias. This dependence is often neglected because the doping concentration of those regions is typically high, and thus the depletion region width extent in them is small. The doping in the source/drain regions is not uniform but decreases with increased depth into the silicon; therefore, the current flows mainly near the surface and this further reduces the influence of variations in the depletion region width, which occur around the junction.

**FIGURE 6.18**

(a) A simple transistor layout. (b) Current paths for (a).

Assume first that the gate is contacted on only one side, as shown in Fig. 6.18a. The gate current flows initially *horizontally* in this picture, and then vertically, toward the gate-oxide interface, as shown in Fig. 6.18b. Let d_{CG} be the distance from the contact to the edge of the channel, and let the sheet resistance of the salicidized gate (including salicidized and nonsalicidized parts in parallel) be R_{SHG} . The resistance of the part between the gate contact and the edge of the channel will be, from (1.3.16), $R_{SHG}(d_{CG}/L)$. The resistance of the entire main part of the gate will be $R_{SHG}(W/L)$. However, this value is relevant only for the extreme right edge of the device. As illustrated in Fig. 6.18b, at high frequencies, the vertical component of the gate current is shunted to the channel through the oxide capacitance; thus, although parts of the device toward the right see most of the horizontal gate resistance, parts on the left see only a small part of it. Thus, if a single effective gate resistance value is to be used, that value can be expected to be smaller than the total value found previously, and can be expressed in the form $R_{SHG}(W/mL)$, where $m > 1$. To find the proper value for m , one must view the gate resistance and gate oxide capacitance as a distributed RC system, and perform a transmission line analysis. For ac small-signal purposes, such an analysis⁹⁵ results in $m = 3$, and this value is often used, somewhat arbitrarily, also for large signal transient analysis.

Finally, to reach the bottom part of the gate, adjacent to the oxide in Fig. 6.16, the current must flow vertically through unsalicidized polysilicon of area WL . If the current flow could be assumed to be uniform, the resistance would be inversely proportional to the cross-sectional area through which the current flows and would thus be of the form $a_{G,\text{vert}}/(WL)$, where $a_{G,\text{vert}}$ is a constant, numerically equal to the resistance one would have for the vertical portion if WL were equal to one unit of area. Unfortunately, as can be expected from Fig. 6.18b, as we go toward the right, less and less gate current remains to be diverted vertically; as a result, the vertical current

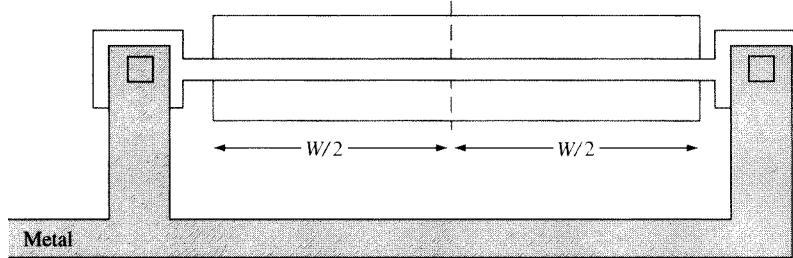


FIGURE 6.19
A transistor with gate contacts on both sides.

density is greater toward the left than toward the right. This is referred to as current crowding, and it complicates the vertical resistance calculation. Nevertheless, for simplicity the vertical resistance is still assumed to be given by the aforementioned formula, only with the value of $a_{G,\text{vert}}$ modified to approximately account for current crowding. Putting the various resistance contributions together, we have

$$R_G = R_{SHG} \left(\frac{d_{CG}}{L} + \frac{W}{mL} \right) + \frac{a_{G,\text{vert}}}{WL} \quad (6.8.5)$$

To reduce the gate resistance, we can contact the gate on both sides, as shown in Fig. 6.19. The effect will be equivalent to having two devices, each with gate width $W/2$ and thus gate resistance of half the total value, in parallel. Thus we will have

$$R_G = \frac{1}{2} \left(R_{SHG} \left(\frac{d_{CG}}{L} + \frac{W/2}{mL} \right) + \frac{a_{G,\text{vert}}}{(W/2)L} \right) = R_{SHG} \left(\frac{d_{CG}}{2L} + \frac{W}{4mL} \right) + \frac{a_{G,\text{vert}}}{WL} \quad (6.8.6)$$

The preceding estimates give an idea of the approximations that must be used to balance complexity and accuracy for modeling extrinsic parasitics. Similar levels of approximations and trade-offs are used in modeling other parasitics, including extrinsic capacitances.

Substrate Resistance The body material offers resistance to capacitive currents, which come mostly from charge storage effects in the depletion regions under the channel and under the source and drain regions. The body also offers resistance to dc currents in the substrate, which primarily come from impact ionization and GIDL currents; see Secs. 5.8 and 5.13.[†] This resistance is obviously distributed; it is approximated by a few lumped elements, as shown in Fig. 6.16, which form a bulk resistance network. Several such networks have been proposed, depending on such factors as the placement of bulk contacts. Readers are referred to Refs. 92, 97, and 98 for details.

[†] The dc effect of substrate resistance is important more for its indirect effect on substrate bias. The voltage drop across the substrate resistance affects the body potential seen by the intrinsic transistor and therefore affects the threshold voltage of the transistor through the body effect. This gives rise to the so-called substrate-current-induced body effect.⁹⁷

6.8.3 Temperature Dependence

The value of extrinsic parasitics can vary with temperature. This is notably true in the case of resistances and is usually modeled using one of two empirical forms. One of these is a power law:

$$R(T) = R(T_r) \left(\frac{T}{T_r} \right)^{k_R} \quad (6.8.7)$$

where T is absolute temperature, T_r is a reference temperature (typically room temperature), and k_R is an empirical coefficient. The second form used is linear:

$$R(T) = R(T_r)[(1 + c_R(T - T_r))] \quad (6.8.8)$$

where c_R is the temperature coefficient (or TC for short) of the resistance. Whichever form is used, the coefficient is determined from experimental data, and different components of parasitic resistance (source/drain, gate, and substrate) have different coefficients.

Junction capacitances vary weakly with temperature and can be modeled using similar relations; details are available in Ref. 85. The parasitic dielectric capacitances do not vary significantly with temperature.

6.8.4 Simplified Models

The extrinsic parasitics model in Fig. 6.16 is certainly an approximation; yet, even this approximation is too complex for hand analysis during circuit design, and even for computer simulation of circuits involving large numbers of transistors. This is because each new node introduced in a model increases the size of the systems of equations that must be solved and slows down the computation. Thus, usually simpler models are used with fewer additional nodes. A common approach is to combine all related elements in Fig. 6.16 into a single element. Thus, for example, the effect of all source-related resistances (with subscripts se in the figure) is represented by a single extrinsic source resistance, R_{se} , and all gate-source-related extrinsic capacitances (with subscripts gse in the figure) are represented by a single extrinsic gate-source capacitance, C_{gse} . An example of such an approach is the model of Fig. 6.20. Variations of this model are also in use; for example, in certain cases the substrate resistance network of Fig. 6.16 is maintained. $C_{bb'}$ is the body-substrate capacitance (Sec. 6.8.1).

It is clear that the simplified model of Fig. 6.20 is not equivalent to that in Fig. 6.16 and thus cannot reveal all of the intricacies of the behavior of the latter. However, such simplified models are widely used and have proven useful in practice. The elements of the model are mostly determined from measurements on the entire transistor, using parameter extraction (Sec. 10.7). The values of the elements thus do not exactly correspond to the values of the physical elements in the transistor structure, but rather are chosen so as to make the behavior of the simplified model best match the measured behavior of a transistor, which of necessity includes the effects of both the intrinsic and extrinsic parts. This “fine adjustment” of the model parameter values helps extend the validity of the simplified model.

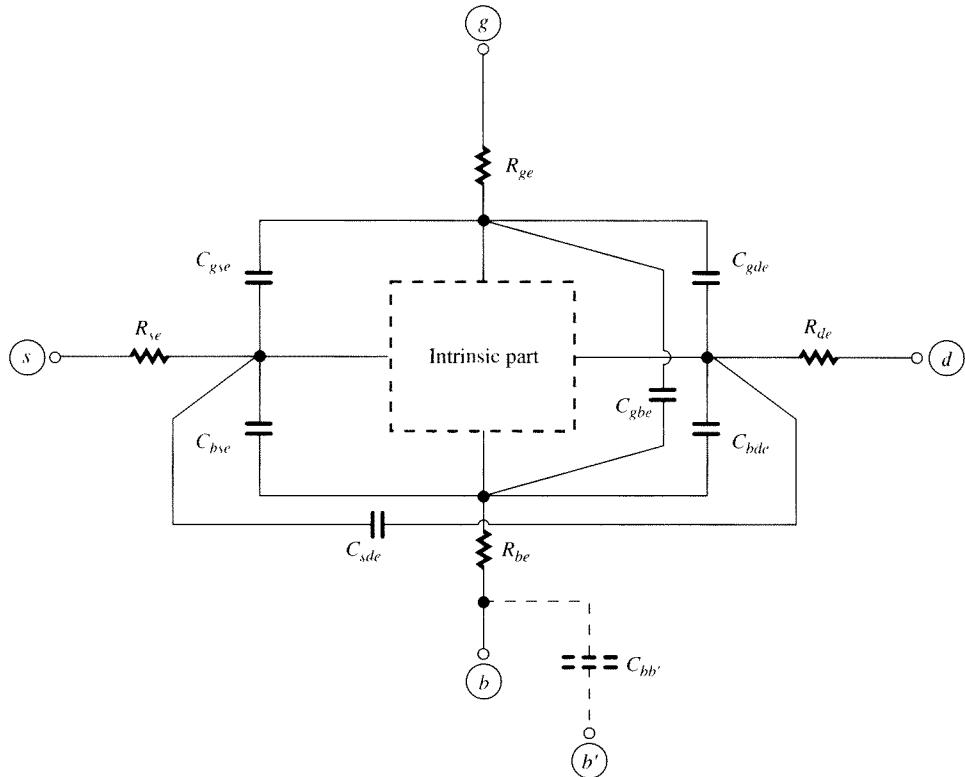


FIGURE 6.20
Simplified equivalent circuit of the MOS transistor, including extrinsic capacitances and resistances.

References

1. T. J. O'Reilly, "The transient response of insulated gate field-effect transistors," *Solid-State Electronics*, vol. 8, pp. 947–956, 1965.
2. Z. S. Gribnikov and Y. A. Tkhonik, "Calculation of the transient process in field triodes with an insulated gate for the saturated mode of operation," *Radio Engineering and Electronic Physics*, vol. 11, pp. 776–781, 1966.
3. A. Möschwitzer, "Zum statischen und dynamischen Grossignalverhalten des MOS-Feldeffekt-Transistors (static and dynamic large signal behavior of MOS field effect transistor)," *NTZ*, vol. 20, pp. 150–154, 1967.
4. M. B. Das, "Switching characteristics of MOS and junction-gate field-effect transistors," *IEE Proceedings*, vol. 114, pp. 1223–1230, 1967.
5. F. A. Lindholm, R. J. Balda, and J. L. Clements, "Characterization of the four-terminal MOS transistor for digital and linear applications," *Proceedings of the International Electronics Conference*, Toronto, 1967, pp. 116–117.
6. J. R. Burns, "Large-signal transit-time effects in the MOS transistor," *RCA Review*, vol. 15, pp. 15–35, 1969.
7. J. S. Brugler and P. G. A. Jespers, "Charge pumping in MOS devices," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 297–302, 1969.

8. R. S. Cobbold, *Theory and Applications of Field-Effect Transistors*, Wiley-Interscience, New York, 1970.
9. K. Gocer, "Einschaltzeiten und Umladungsvorgänge bei MOS-Transistoren," *AEU*, vol. 24, pp. 21–28, 1970.
10. D. J. Hamilton, F. A. Lindholm, and A. H. Marshak, *Principles and Applications of Semiconductor Device Modeling*, Holt, Rinehart, and Winston, New York, 1971.
11. J. E. Meyer, "MOS models and circuit simulation," *RCA Review*, vol. 32, pp. 42–43, 1971.
12. M. E. Zahn, "Calculation of the turn-on behavior of MOST," *Solid-State Electronics*, vol. 17, pp. 843–854, 1974.
13. R. M. Swanson, "Complementary MOS transistors in micropower circuits," Technical Report 4963-1, Integrated Circuits Laboratory, Stanford University, Stanford, Calif. 1974.
14. F. M. Klaassen, "A MOS model for computer-aided design," *Philips Research Reports*, vol. 31, pp. 71–83, 1976.
15. J. I. Arreola, "Equivalent circuit modeling of the large signal transient response of four-terminal MOS field-effect transistors," doctoral dissertation, University of Florida, 1978.
16. D. E. Ward and R. W. Dutton, "A charge-oriented model for MOS transistor capacitances," *IEEE Journal of Solid-State Circuits*, vol. SC-13, pp. 703–707, 1978.
17. P. E. Cottrell and E. Buturla, "Two-dimensional static and transient simulation of mobile carrier transport in a semiconductor," *Proceedings of NASECODE I*, Dublin, June 1979.
18. J. A. Robinson, Y. A. El-Mansy, and A. R. Boothroyd, "A general four-terminal charging-current model for the insulated-gate field effect transistor," parts I and II, *Solid-State Electronics*, vol. 23, pp. 405–414, 1980.
19. S. Y. Oh, D. E. Ward, and R. W. Dutton, "Transient analysis of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. SC-15, pp. 636–643, 1980.
20. S. Y. Oh, "A simplified two-dimensional numerical analysis of MOS devices including transient phenomena," Technical Report G201-10, Integrated Circuits Laboratory, Stanford University, Calif. June 1981.
21. D. E. Ward, "Charge-based modeling of capacitance in MOS transistors," Technical Report G201-11, Integrated Circuits Laboratory, Stanford University, Calif. June 1981.
22. M. S. Mock, "A time-dependent numerical model of the insulated-gate field-effect transistor," *Solid-State Electronics*, vol. 24, pp. 959–966, 1981.
23. Y. Ikawa, W. R. Eisenstadt, and R. W. Dutton, "Modeling of high-speed, large-signal transistor switching transients from s-parameter measurements," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 669–675, 1982.
24. G. W. Taylor, W. Fichtner, and J. G. Simmons, "A description of MOS internodal capacitances for transient simulations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-1, pp. 150–156, 1982.
25. R. Conilogue and E. Viswanathan, "A complete large and small signal charge model for a M.O.S. transistor," *Proceedings of the International Electron Devices Meeting*, San Francisco, 1982, pp. 654–657.
26. K. Y. Tong, "A model for MOS transistors from transient current computations," *IEE Proceedings*, vol. 130, part I, pp. 33–36, 1983.
27. P. Yang, B. D. Epler, and P. K. Chatterjee, "An investigation of the charge conservation problem for MOSFET circuit simulation," *IEEE Journal of Solid-State Circuits*, vol. SC-18, pp. 128–138, 1983.
28. C. Turchetti, G. Masetti, and Y. Tsividis, "On the small-signal behavior of the MOS transistor in quasi-static operation," *Solid-State Electronics*, vol. 26, pp. 941–949, 1983.
29. B. J. Sheu, D. L. Scharfetter, C. Hu, and D. O. Pederson, "A compact IGFET charge model," *IEEE Transactions on Circuits and Systems*, vol. CAS-31, pp. 745–749, 1984.
30. J. J. Paulos, "Measurement and modeling of small-geometry MOS transistor capacitances," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, September 1984.
31. C. Turchetti, P. Prioretti, G. Masetti, E. Profumo, and J. Vanzi, "A Meyer-like approach for the transient analysis of digital MOS ICs," *IEEE Transactions on CAD*, vol. CAD-5, pp. 499–507, 1986.
32. C. Turchetti, P. Mancini, and G. Masetti, "A CAD-oriented non quasi-static approach for the transient analysis of MOS ICs," *IEEE Journal of Solid-State Circuits*, vol. SC-21, pp. 827–836, 1986.

33. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
34. J. G. Fossum, H. Jeong, and S. Veeraraghavan, "Significance of the channel-charge partition in the transient MOSFET model," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 1621–1623, 1986.
35. M. F. Sevat, "On the channel charge division in MOSFET modeling," *Proceedings of the IEEE International Conference on Computer-Aided Design*, pp. 204–207, 1987.
36. M. A. Cirit, "The Meyer model revisited: Why is charge not conserved?" *IEEE Transactions on Computer-Aided Design*, vol. CAD-8, pp. 1033–1037, 1989.
37. I. W. Smith, H. Statz, H. A. Haus, and R. A. Pucel, "On charge nonconservation in FETs," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 2565–2568, 1987.
38. K. A. Sakallah, Y.-T. Yen, and S. S. Greenberg, "A first-order charge conserving MOS capacitance model," *IEEE Transactions on Computer-Aided Design*, vol. 9, pp. 99–108, 1990.
39. A. Afzali-Kushaa and M. El-Nokali, "Modeling subthreshold capacitances of MOS transistors," *Solid-State Electronics*, vol. 35, pp. 45–49, 1992.
40. M. Anderson and P. Kuivalainen, "Transit-time model for short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, p. 830, 1993.
41. H. K. Lim and J. Fossum, "A charge-based large-signal model for thin-film SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 446–457, 1985.
42. J. Katzenelson and A. Unikovski, "A network charge-oriented MOS transistor model," *International Journal of High Speed Electronics and Systems*, vol. 6, pp. 285–316, 1995.
43. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
44. N. Arora, *MOSFET Models for VLSI Circuit Simulation—Theory and Practice*, Springer-Verlag, Vienna, 1993.
45. K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, N.J., 1993.
46. G. I. Serhan and S.-Y. Yu, "A simple charge-based model for MOS transistor capacitances: A new production tool," *IEEE Transactions on Computer-Aided Design*, vol. CAD-2, p. 48, 1983.
47. M. A. Maher and C. A. Mead, "A physical charge-controlled model for MOS transistors," pp. 211–229, P. Leslesben (editor), *Advanced Research in VLSI*, The MIT Press, Cambridge, Mass., 1987.
48. B. J. Sheu and P.-K. Ko, "Measurement and modeling of short-channel MOS transistor gate capacitances," *IEEE Journal of Solid-State Circuits*, vol. SC-22, pp. 464–472, 1987.
49. H. Masuda, Y. Aoki, J. Mano, and O. Yamashiro, "MOSTSM: A physically based charge conservative MOSFET model," *IEEE Transactions on Computer-Aided Design*, vol. 7, pp. 1229–1236, 1988.
50. B. J. Sheu, W.-J. Hsu, and P. K. Ko, "An MOS transistor charge model for VLSI design," *IEEE Transactions on Computer-Aided Design*, vol. 7, pp. 520–527, 1988.
51. R. Gharabagi and M. E. Nokali, "A model for the intrinsic gate capacitances of short channel MOSFETs," *Solid-State Electronics*, vol. 32, pp. 57–63, 1989.
52. W. Budde and W. H. Lamfried, "A charge-sheet capacitance model based on drain current modeling," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1678–1687, 1990.
53. R. Gharabagi and M. A. El-Nokali, "A charge-based model for short-channel MOS transistor capacitances," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1064–1073, 1990.
54. H.-J. Park, P. K. Ko, and C. Hu, "A charge conserving non-quasi-static (NQS) MOSFET model for SPICE transient analysis," *IEEE Transactions on Computer-Aided Design*, vol. 10, pp. 629–641, 1991.
55. K.-M. Rho, K. Lee, M. Shur, and T. A. Fjeldly, "Unified quasi-static MOSFET capacitance model," *IEEE Transactions on Electron Devices*, vol. 40, pp. 131–135, 1993.
56. D.-H. Cho, S.-M. Kang, K.-H. Kim, and S.-H. Lee, "An accurate intrinsic capacitance modeling for deep submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. 42, pp. 540–548, 1995.
57. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An explicit physical model for the long-channel MOS transistor including small-signal parameters," *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
58. C. C. Enz, "The EVK model: A MOST model dedicated to low-current and low-voltage analogue circuit design and simulation," chap. 7, *Low-Power HF Microelectronics: A Unified Approach*, G. A. S. Machado (editor), IEE Circuits and Systems Series No. 8, IEE Book Publishing, London 1996.

59. H.-H. Li, Y.-L. Chu, and C.-Y. Wu, "A new simplified charge pumping current model and its model parameter extraction," *IEEE Transactions on Electron Devices*, vol. 43, pp. 1857–1863, 1996.
60. A. I. A. Cunha, O. C. Gouveia-Filho, M. C. Schneider, and C. Galup-Montoro, "A current-based model for the MOS transistor," *Proceedings of the 1997 International Symposium on Circuits and Systems*, Hong Kong, June 1997, pp. 1608–1611.
61. A. I. A. Cunha, *Um Modelo do Transistor MOS para Projeto de Circuitos Integrados*, Ph.D. thesis, Universidade Federal de Santa Catarina, December 1996 (in Portuguese).
62. P. Mancini, C. Turchetti, and G. Masetti, "A non-quasi-static analysis of the transient behavior of the long-channel MOST valid in all regions of operation," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 325–334, 1987.
63. K.-W. Chai and J. J. Paulos, "Unified non-quasi-static modeling of the long-channel four-terminal MOSFET for large- and small-signal analyses in all operating regimes," *IEEE Transactions on Electron Devices*, vol. 36, pp. 2513–2520, 1989.
64. H. J. Park, P. K. Ko, and C. Hu, "A non-quasi-static MOSFET model for SPICE-transient analysis," *IEEE Transactions on Electron Devices*, vol. 36, pp. 561–576, 1989.
65. L.-J. Pu and Y. Tsividis, "Harmonic distortion of the four-terminal MOSFET in non-quasi-static operation," *IEE Proceedings*, vol. 137, pp. 325–332, 1990.
66. P. Roblin, S. C. Kang, and W.-R. Liou, "Improved small-signal equivalent circuit model and large-signal state equations for the MOSFET/MODFET wave equation," *IEEE Transactions on Electron Devices*, vol. 38, pp. 1706–1718, 1991.
67. M. Chan, K. Hui, R. Neff, C. Hu, and P. K. Ko, "A relaxation time approach to model the non-quasi-static transient effects in MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 169–172, December 1994.
68. E. Dubois and E. Robilliart, "Efficient non-quasi-static MOSFET's model for circuit simulation," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 945–948, December 1995.
69. W. Liu, C. Bowen, and M.-C. Chang, "A CAD-compatible non-quasi-static MOSFET model," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 151–154, December 1996.
70. D. Foty, *MOSFET Modeling with SPICE—Principles and Practice*, Prentice Hall, Upper Saddle River, N.J., 1997.
71. D. Murray, J. J. Sanchez, and T. A. Demassa, "A non-quasi-static semi-empirical model for small geometry MOSFETs," *Solid-State Electronics*, vol. 41, pp. 1309–1319, 1997.
72. C. C. McAndrew and J. Victory, "Accuracy of approximations in MOSFET charge models," *IEEE Transactions on Electron Devices*, vol. 49, no. 1, pp. 72–81, 2002.
73. H. Wang, T.-L. Chen, and G. Gildenblat, "Quasi-static and non-quasi-static compact MOSFET models based on symmetric linearization of the bulk and inversion charges," *IEEE Transactions on Electron Devices*, vol. 50, no. 11, pp. 2262–2272, 2003.
74. G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, 2006.
75. M. Bucher, *Analytical MOS Transistor Modelling for Analog Circuit Simulation*, EPFL Ph.D. Thesis no. 2114, Lausanne, 1999.
76. C. Galup-Montoro and M. Schneider, *MOSFET Modeling for Circuit Analysis and Design*, World Scientific, River Edge, N.J., 2007.
77. M. Chan, K. Hui, R. Neff, C. Hu, and P. K. Ko, "A relaxation time approach to model the non-quasi-static transient effects in MOSFETs," *Proceedings of the International Electron Devices Meeting*, pp. 169–172, December 1994.
78. K. Machida, D. Navarro, M. Miyake, R. Inagaki, N. Sadachika, T. Ezaki, H. J. Mattausch, and M. Miura-Mattausch, "Efficient non-quasi-static MOSFET model for both time-domain and frequency-domain analysis," *Proceedings of the Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems*, pp. 73–76, January 2006.
79. P. Mancini, C. Turchetti, and G. Masetti, "A non-quasi-static analysis of the transient behavior of the long-channel MOST valid in all regions of operation," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 2, pp. 325–334, 1987.
80. D. H. Kwon, Y. S. Yu, K. H. Kim, T. W. Yoon, and S. W. Hwang, "A new CAD-compatible non-quasi-static MOS transient model with direct inclusion of depletion charge variations," *Journal of the Korean Physical Society*, vol. 33, pp. 192–195, 1998.

81. S. W. Hwang, T.-W. Yoon, D. H. Kwon, Y. S. Yu, and K. H. Kim, "A physics-based, SPICE (Simulation Program with Integrated Circuit Emphasis)-compatible non-quasi-static MOS (metal-oxide-semiconductor) transient model based on the collocation method," *Japanese Journal of Applied Physics*, vol. 37, pp. L119–L121, 1998.
82. H. Wang, X. Li, W. Wu, G. Gildenblat, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "A unified non-quasi-static MOSFET model for large-signal and small-signal simulations," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2035–2043, 2006.
83. G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, 2006.
84. P. Antognetti and G. Massobrio, *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York, 1993.
85. A. J. Scholten, G. D. J. Smit, M. Durand, R. van Langevelde, and D. B. M. Klaassen, "The physical background of JUNCAP2," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2098–2107, 2006.
86. H. Iwai, M. R. Pinto, C. S. Rafferty, J. E. Orlitian, and R. W. Dutton, "Analysis of velocity saturation and other effects on short-channel MOS transistor capacitances," *IEEE Transactions on Computer-Aided Design*, vol. CAD-6, pp. 173–184, 1987.
87. H. Murrman and D. Widmann, "Current crowding on metal contacts to planar devices," *IEEE Transactions Electron Devices*, vol. ED-16, pp. 1022–1026, 1969.
88. H. H. Berger, "Models for contacts to planar devices," *Solid-State Electronics*, vol. 15, pp. 145–158, 1972.
89. G. Baccarani and G. A. Sai-Halasz, "Spreading resistance in submicron MOSFETs," *IEEE Electron Device Letters*, vol. EDL-4, pp. 27–29, 1983.
90. K. K. Ng, R. J. Bayruns, and S. C. Fang, "The spreading resistance of MOSFETs," *IEEE Electron Device Letters*, vol. EDL-6, pp. 195–197, 1985.
91. K. K. Ng and W. T. Lynch, "Analysis of the Series Resistance of MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 965–972, 1986.
92. X. Xi, M. Dunga, J. He, W. Liu, K. M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu, *BSIM4.5.0 MOSFET Model—User's Manual*, University of California, Berkeley, 2004.
93. A. Litwin, "Overlooked interfacial silicide-polysilicon gate resistance in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 48, no. 9, pp. 2179–2181, 2001.
94. A. J. Scholten, L. F. Tiemeijer, R. van Langevelde, R. J. Havens, A. T. A. Zegers-van Duijnoven, R. de Kort, and D. B. M. Klaassen, "Compact modelling of noise of RF CMOS circuit design," *IEE Proceedings—Circuits, Devices and Systems*, vol. 151, no. 2, pp. 167–174, 2004.
95. B. Razavi, R.-H. Yan, and K. F. Lee, "Impact of distributed gate resistance on the performance of MOS devices," *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, vol. 41, no. 11, pp. 750–754, 1994.
96. BSIM website. Available at <http://www-device.eecs.berkeley.edu/~bsim3/>. Accessed Oct. 1, 2009.
97. A. J. Scholten, L. F. Tiemeijer, P. W. H. De Vreede, and D. B. M. Klaassen, "A large signal non-quasi-static MOS model for RF circuit simulation," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 163–166, December 1999.
98. C. C. Enz and Y. Cheng, "MOS transistor modeling for RF IC design," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, pp. 186–201, 2000.

Problems

- 6.1.** Show that (6.4.4) and (6.4.5) are equivalent.
- 6.2.** Derive in detail the expressions for the total charges Q_I , Q_B , Q_G , Q_D , and Q_S in (6.4.14), (6.4.15), (6.4.17), (6.4.19), and (6.4.20).
- 6.3.** Consider a transistor with $N_A = 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2 \text{ nm}$, $Q'_o = 2 \times 10^{-9} \text{ C/cm}^2$, $\phi_{MS} = -1 \text{ V}$, $W = L = 1 \mu\text{m}$, with $V_{SB} = 0 \text{ V}$. Plot the total charges Q_G , Q_B , Q_I , Q_D , and Q_S vs. V_{DS} , with V_{DS} between 0 and 1.2 V, for $V_{GS} = 0.7$ and 1.2 V. Use the results of Sec. 6.4.2.

- 6.4.** Find expressions given for Q_B , Q_I , and Q_G corresponding to the complete strong-inversion model of Sec. 4.7.1.
- 6.5.** Compare the expressions for Q_B , Q_I , and Q_G in (6.4.15), (6.4.14) and (6.4.17) to those found in Prob. 6.4. For quantitative comparisons, use the parameters of Prob. 6.3.
- 6.6.** Using the parameters of Prob. 6.3, compare the following expressions for the total inversion layer charge Q_I :
- Expression corresponding to the simplified model (6.4.14).
 - Expression corresponding to the complete strong-inversion model found in Prob. 6.4.
- 6.7.** Prove the expressions for Q_I , Q_D , and Q_S in weak inversion given in Sec. 6.4.4.
- 6.8.** Using the method outlined in Sec. 6.4.5, develop expressions for Q_I , Q_B , Q_G , Q_D , and Q_S in terms of ψ_{s0} and ψ_{sL} corresponding to the all-region model in Sec. 4.3 (this is a long problem).
- 6.9.** Repeat Prob. 6.8 by using a first-degree polynomial approximation for Q'_B as in (4.4.3); use the linearization involving (4.4.7).
- 6.10.** For the device of Prob. 6.3, plot Q_G , Q_B , Q_D , and Q_S vs. V_{GS} , with V_{GS} from -1.2 to 1.2 V, and for $V_{DS} = 0.2$ and 0.8 V.
- 6.11.** Assume a fixed parasitic extrinsic capacitance C_{gde} exists between gate and drain in the device of Fig. 6.11a (such a capacitance can be caused by the physical overlap between gate and drain, as discussed in Sec. 6.8.1). Show that the effect of such a capacitance will be a downward shift of the current waveforms in Fig. 6.11d and e, between $t = 0$ and $t = t_3$.
- 6.12.** For the transistor of Prob. 6.3, assume a constant effective mobility $\mu = 600 \text{ cm}^2/(\text{V}\cdot\text{s})$. The transistor is connected as in Fig. 6.11a with $V_{DD} = 1.2 \text{ V}$; $v_G(t)$ is as shown in Fig. 6.11b, with $t_3 = 4 \text{ ns}$ and $v_{G,\max} = 1.2 \text{ V}$. Plot $i_D(t)$, $i_G(t)$, $i_S(t)$, and $i_B(t)$ as functions of time.
- 6.13.** Consider the situation shown in Fig. P6.1. Assume that the transistor is always in strong inversion and operates quasi-statically. Neglect all extrinsic parasitic elements. Derive expressions for, and plot, $i_S(t)$ and $i_D(t)$.
- 6.14.** The transistor of Prob. 6.3 is biased with $V_G = 1.2 \text{ V}$ and $V_S = V_D = 0.3 \text{ V}$, all with respect to ground. The body is connected to a linear voltage ramp that varies from 0 V to 0.3 V in 20 ns. The mobility is $600 \text{ cm}^2/(\text{V}\cdot\text{s})$. Consider only intrinsic effects. Plot the drain current vs. time.

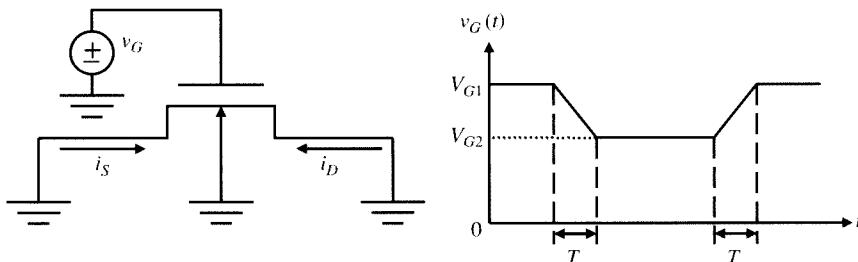


FIGURE P6.1

Small-Signal Modeling for Low and Medium Frequencies

7.1 Introduction

In the previous chapter we considered the MOS transistor with terminal voltages undergoing variations with time. No restrictions were placed on the magnitude of these variations. In this chapter we will consider the case where the terminal voltage variations are sufficiently small so that the resulting small current variations can be expressed in terms of them using linear relations. We will derive such linear relations and develop linear circuits to represent them. These circuits will be called *small-signal equivalent circuits*. When excited by voltages equal to the *small variations* of the actual terminal voltages, these circuits will produce currents equal to the *small variations* of the actual transistor currents. Such models find wide use in analog circuit design and have been studied since the early days of MOS transistors.¹⁻⁵

In most of this chapter, we will concentrate on the *intrinsic* part of the transistor (Fig. 6.1). We will first develop a small-signal model valid when the voltage and current variations are so slow that charge storage effects can be neglected. Then we will develop a small-signal model valid at medium speeds, assuming quasi-static operation. Such models¹⁻⁴⁵ are sufficient for many applications. (More advanced models will be presented in Chap. 8.) It is convenient to discuss modeling by assuming that the voltage and current *variations* are sinusoidal. In that case, one often talks of frequency rather than “speed,” and we will adopt this convention from now on. Following the discussion of the models for the intrinsic part of the transistor, we will combine such models with those for the extrinsic part. We will then consider the noise generated within a MOS transistor.

As in the previous chapter, *n*-channel devices with long and wide channels on a uniform substrate and with constant mobility will be assumed, unless noted otherwise.

We begin by considering the *intrinsic* part of the device.

7.2 A Low-Frequency Small-Signal Model for the Intrinsic Part

7.2.1 Introduction

In developing a small-signal model for the dc behavior of the MOS transistor, we should recall that, in the general case, the drain-to-source current I_{DS} is not the only current that flows. As we have seen in Sec. 5.12, the use of very thin oxides in modern processes results in a gate current I_G . In addition, as we have seen in Sec. 5.8, high electric fields near the drain can result in impact ionization, which can cause a body current I_B . We will build our small-signal model gradually; we will first consider only the drain-to-source current and develop a small-signal model for it, then consider how this model must be augmented to account for gate current, and finally how it must be augmented further to account for body current.

7.2.2 Small-Signal Model for the Drain-to-Source Current

In order to concentrate on the drain-to-source current, we initially consider the case where the gate and body currents are negligible. We assume

$$I_G = 0 \quad (7.2.1)$$

$$I_B = 0 \quad (7.2.2)$$

These assumptions will be revised in Secs. 7.2.3 and 7.2.4. We assume that the drain-to-source current is expressed as a function of V_{GS} , V_{BS} , and V_{DS} . This is the case, for example, with the source-referenced model of Sec. 4.7.3; body-referenced models such as those of Secs. 4.7.1 or 4.7.2 can be expressed as functions of the preceding voltages using the simple transformations $V_{GB} = V_{GS} - V_{BS}$ and $V_{DB} = V_{DS} - V_{BS}$. Finally, the general models of Sec. 4.3 can be considered as implicit functions of the preceding voltages, through the surface potentials. Thus, we will assume that the drain-to-source current is of the form

$$I_{DS} = I_{DS}(V_{GS}, V_{BS}, V_{DS}) \quad (7.2.3)$$

Let us consider an *n*-channel MOS transistor biased with V_{GS} , V_{SB} , and V_{DS} fixed at values V_{GS0} , V_{SB0} , and V_{DS0} , respectively, as shown in Fig. 7.1a (the subscript 0 indicates “quiescent” values). Let I_{DS0} be the resulting value of I_{DS} . We can study the effect of *very small* changes of the bias voltages on I_{DS} by varying these voltages *one at a time*, as shown in Fig. 7.1b, c, and d. We are for now interested only in the change of the *dc steady-state* value of I_{DS} ; that is, we assume that the voltages are

constant before and after each change and that I_{DS} has reached dc steady state in both cases. We then consider the change ΔI_{DS} between the two dc steady-state values. We can relate cause and effect by using three conductance parameters, which can be measured as shown next to each figure. These parameters are

1. The *small-signal gate transconductance* g_m , often referred to simply as “transconductance.” Mathematically, it is defined by the following relation, corresponding to the situation in Fig. 7.1b:

$$g_m \equiv \left. \frac{\partial I_{DS}}{\partial V_{GS}} \right|_{V_{BS}, V_{DS}} \quad (7.2.4)$$

where to the right of the vertical line we show the voltages being held fixed.[†]

2. The *small-signal substrate transconductance* g_{mb} . Corresponding to the situation in Fig. 7.1c, we have

$$g_{mb} \equiv \left. \frac{\partial I_{DS}}{\partial V_{BS}} \right|_{V_{GS}, V_{DS}} \quad (7.2.5)$$

Increasing V_{BS} as shown in Fig. 7.1c *decreases* V_{SB} . A consideration of the results of the body effect on I_{DS} shows that I_{DS} increases. Thus, ΔV_{BS} has on I_{DS} qualitatively the same effect as ΔV_{GS} has on Fig. 7.1b. The substrate acts, in this sense, as a second gate and is often referred to as the *back gate*.

3. The *small-signal source-drain conductance* g_{sd} . Corresponding to the measurement in Fig. 7.1d, we define

$$g_{sd} \equiv \left. \frac{\partial I_{DS}}{\partial V_{DS}} \right|_{V_{GS}, V_{BS}} \quad (7.2.6)$$

The approximate equality signs in Fig 7.1b to d become “equal” signs as ΔV_{GS} , ΔV_{BS} , and ΔV_{DS} approach zero. All three parameters have units of conductance and are expressed in siemens [(S): 1 S = 1 A/V].

Let us now consider the general case in which all three voltages are changed simultaneously, as shown in Fig. 7.1e. The corresponding total change in the drain-to-source current will be

$$\Delta I_{DS} \approx \left(\left. \frac{\partial I_{DS}}{\partial V_{GS}} \right|_{V_{BS}, V_{DS}} \right) \Delta V_{GS} + \left(\left. \frac{\partial I_{DS}}{\partial V_{BS}} \right|_{V_{GS}, V_{DS}} \right) \Delta V_{BS} + \left(\left. \frac{\partial I_{DS}}{\partial V_{DS}} \right|_{V_{GS}, V_{BS}} \right) \Delta V_{DS} \quad (7.2.7)$$

[†] Strictly speaking, this “reminder” is redundant, since it is implied by the definition of a partial derivative, provided we remember that I_{DS} is expressed as a function of V_{GS} , V_{BS} , and V_{DS} . If, however, I_{DS} were expressed as a function of other independent variables (e.g., V_{GS} , V_{GB} , and V_{GD}), then $\partial I_D / \partial V_{GS}$ would *not* mean the same as (7.2.4). So, for clarity, it doesn’t hurt to remind ourselves which voltages are being held constant.

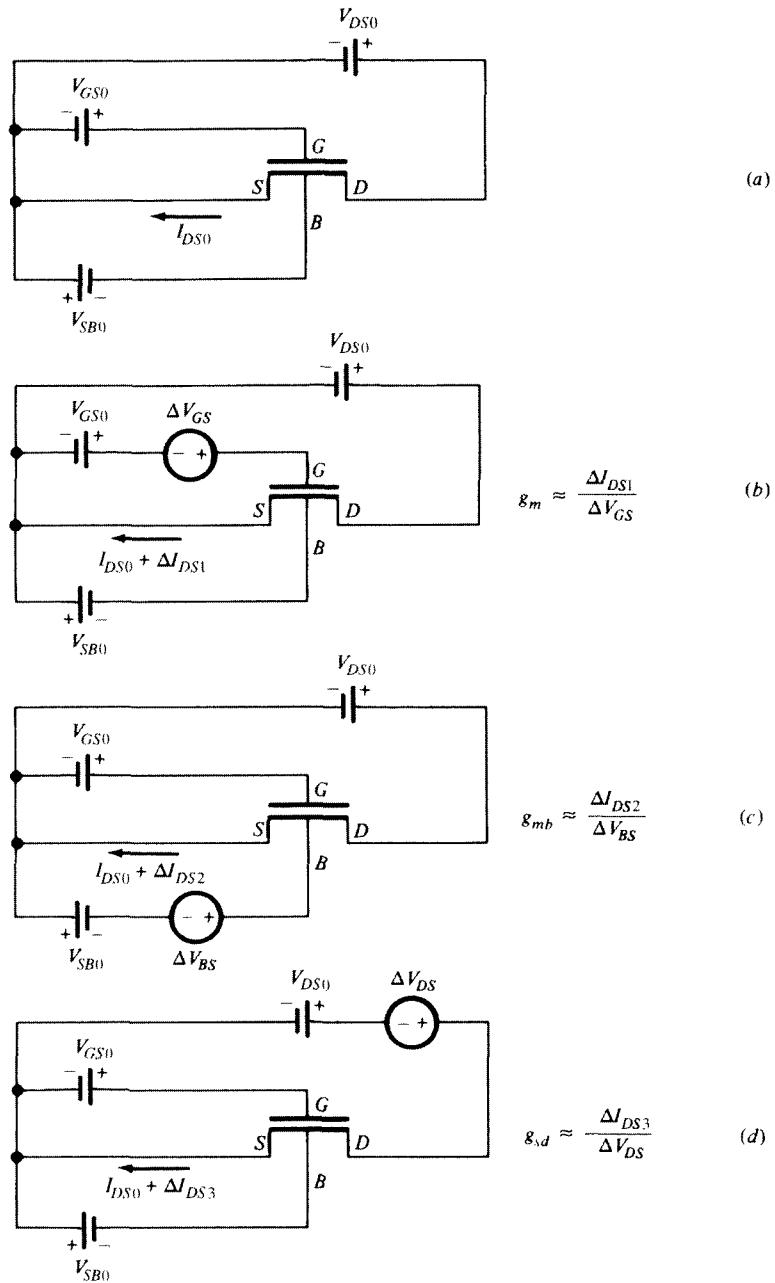


FIGURE 7.1

(a) A MOS transistor with dc voltages at a certain operating point. The rest of the figure shows the same circuit with very small changes (denoted by Δ) in the terminal voltages, and specifically (b) a small change in V_{GS} , (c) a small change in V_{BS} , (d) a small change in V_{DS} , and

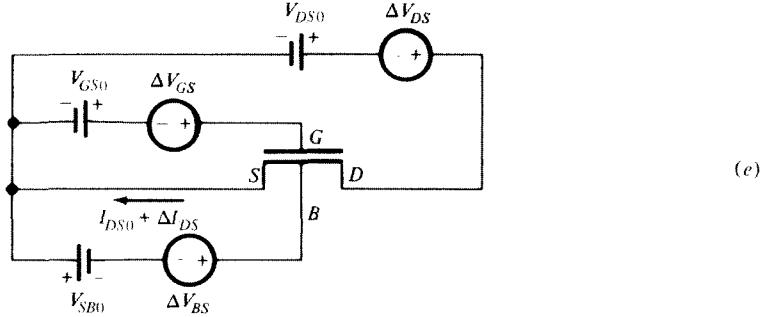


FIGURE 7.1 (Continued).
(c) small changes in V_{GS} , V_{BS} , and V_{DS} simultaneously.

From our above definitions, and assuming that the finite differences go to zero in the limit, this becomes

$$\Delta I_{DS} = g_m \Delta V_{GS} + g_{mb} \Delta V_{BS} + g_{sd} \Delta V_{DS} \quad (7.2.8)$$

In addition, (7.2.1) and (7.2.2) give

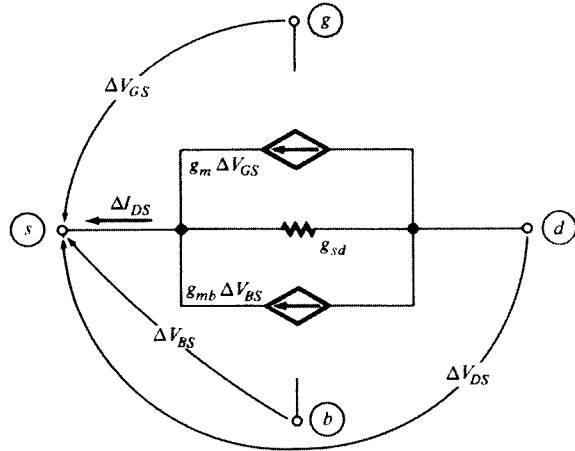
$$\Delta I_G = 0 \quad (7.2.9)$$

and

$$\Delta I_B = 0 \quad (7.2.10)$$

The above three equations relate small-signal quantities and can be represented by the small-signal equivalent circuit in Fig. 7.2, where the rhombic symbols represent controlled current sources. It is easy to verify this circuit by writing Kirchhoff's current law for the left-hand node. As seen, we were able to represent the last term in (7.2.8) by a resistor of conductance g_{sd} , since ΔV_{DS} is the voltage across that resistor and thus $g_{sd} \Delta V_{DS}$ is the current through it. However, the other two terms in (7.2.8) have controlling voltages defined at nodes other than the node into which the current flows; thus, these terms are represented by voltage-controlled current sources.

In the preceding model derivation, the changes ΔV_{GS} , ΔV_{BS} , and ΔV_{DS} represented differences between two dc steady-state values of the terminal voltages V_{GS} , V_{BS} , and V_{DS} respectively. However, the model derived will be valid for representing the effects of gate and substrate on the drain current, even if the changes are continuously varying with time, as long as the variations are slow enough so that capacitive effects can be neglected. This will be understood better, and will be made more quantitative, after more complete models have been considered. It will then be shown that, as the frequency of variation is decreased, more complete models reduce to the one in Fig. 7.2.

**FIGURE 7.2**

A low-frequency small-signal equivalent circuit for the channel of a MOS transistor. The drain-to-substrate path is not included in this partial model.

In the above definitions, we followed the common convention of referring the gate, drain, and substrate potentials to the source. An alternative approach, in which all *four* terminal voltages are taken with respect to an arbitrary reference, is considered in Prob. 7.4.

Source Conductance In body-referenced models, source conductance is a quantity used often,^{19,20} measured as shown in Fig. 7.3, and defined as follows:

$$g_{ss} = \left. \frac{\partial I_S}{\partial V_{SB}} \right|_{V_{GB}, V_{DB}} \quad (7.2.11)$$

where $I_S = -I_{DS}$. The quantity g_{ss} is seen to be the small-signal source conductance. Applying the small-signal equivalent circuit of Fig. 7.2 to the situation in Fig. 7.3, it is easy to show that (Prob. 7.2)

$$g_{ss} = g_m + g_{mb} + g_{sd} \quad (7.2.12)$$

7.2.3 Small-Signal Model for the Gate and Body Currents

We now relax the assumption in (7.2.1). The use of very thin oxides results in a gate current, I_G . From the discussion in Sec. 5.12, we know that in general this current depends on the potential of the gate with respect to the other three terminals, so we assume that it can be expressed in the form

$$I_G = I_G(V_{GS}, V_{GB}, V_{GD}) \quad (7.2.13)$$

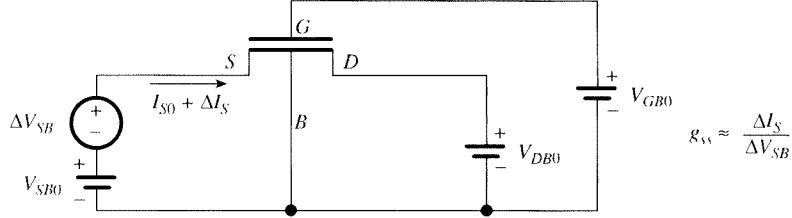


FIGURE 7.3
Measurement of source-to-drain transconductance.

One can follow an approach analogous to the one above, to model the behavior of the gate current for small signals. We have, in analogy to (7.2.8):

$$\Delta I_G = g_{gs} \Delta V_{GS} + g_{gb} \Delta V_{GB} + g_{gd} \Delta V_{GD} \quad (7.2.14)$$

where

$$g_{gs} \equiv \left. \frac{\partial I_G}{\partial V_{GS}} \right|_{V_{GB}, V_{GD}} \quad (7.2.15)$$

$$g_{gb} \equiv \left. \frac{\partial I_G}{\partial V_{GB}} \right|_{V_{GS}, V_{GD}} \quad (7.2.16)$$

$$g_{gd} \equiv \left. \frac{\partial I_G}{\partial V_{GD}} \right|_{V_{GS}, V_{GB}} \quad (7.2.17)$$

Equation (7.2.14) can be modeled simply as shown in Fig. 7.4. Note that this circuit models the terminal description (7.2.14), which is completely general; it is best not to associate individual elements in it with particular current paths inside the device. See also Sec. 7.2.4.

Finally, we relax the assumption in (7.2.2), to allow for body current I_B . Consistent with the convention adopted in Chap. 6, and to be used in later chapters as well, we define this current as going *into* the body terminal, which may result in a negative value for this current. For example, for a *n*MOS device, we know that a positive current I_{DB} due to impact ionization goes *out* from the body terminal (Fig. 5.18); thus, assuming this is the dominant body current component, we will have $I_B = -I_{DB}$, and I_B will be negative. This in no way influences the general results to follow.

We will express the body current in a manner similar to (7.2.13):

$$I_B = I_B(V_{BS}, V_{BG}, V_{BD}) \quad (7.2.18)$$

In analogy with (7.2.14) we will have:

$$\Delta I_B = g_{bs} \Delta V_{BS} + g_{bg} \Delta V_{BG} + g_{bd} \Delta V_{BD} \quad (7.2.19)$$

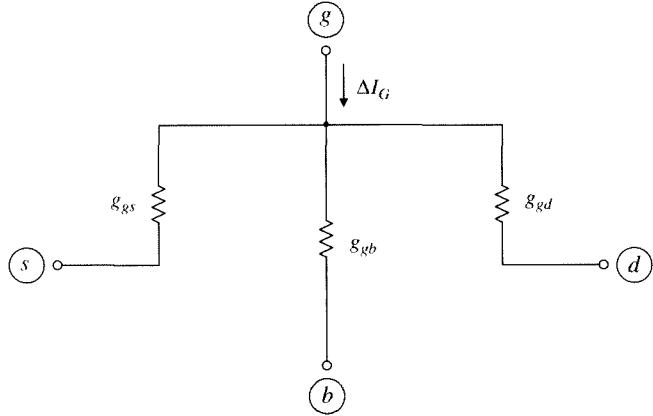


FIGURE 7.4
Small-signal model for gate leakage current.

where

$$g_{bs} = \left. \frac{\partial I_B}{\partial V_{BS}} \right|_{V_{BG}, V_{BD}} \quad (7.2.20)$$

$$g_{bg} = \left. \frac{\partial I_B}{\partial V_{BG}} \right|_{V_{BS}, V_{BD}} \quad (7.2.21)$$

$$g_{bd} = \left. \frac{\partial I_B}{\partial V_{BD}} \right|_{V_{BS}, V_{BG}} \quad (7.2.22)$$

We will now augment the model of Fig. 7.4 to include the body current. According to (7.2.19), we can use a small-signal conductance g_{hs} between body and source, and another small-signal conductance g_{hd} between body and drain. However, when we attempt to add a small-signal conductance g_{bg} between body and gate, a problem becomes apparent: this element will interfere with the modeling of the gate current, which has already been taken care of by g_{gb} in Fig. 7.4. To get around this problem, we can write the middle term in (7.2.19) as

$$g_{bg} \Delta V_{BG} = g_{gb} \Delta V_{BG} + (g_{bg} - g_{gb}) \Delta V_{BG} = g_{gb} \Delta V_{BG} + (g_{gb} - g_{bg}) \Delta V_{GB}$$

where in the last term we have used $\Delta V_{GB} = -\Delta V_{BG}$. Thus, (7.2.19) can be written as follows:

$$\Delta I_B = g_{bs} \Delta V_{BS} + g_{gb} \Delta V_{BG} + g_{mx} \Delta V_{GB} + g_{bd} \Delta V_{BD} \quad (7.2.23)$$

where

$$g_{mx} \equiv g_{gb} - g_{bg} \quad (7.2.24)$$

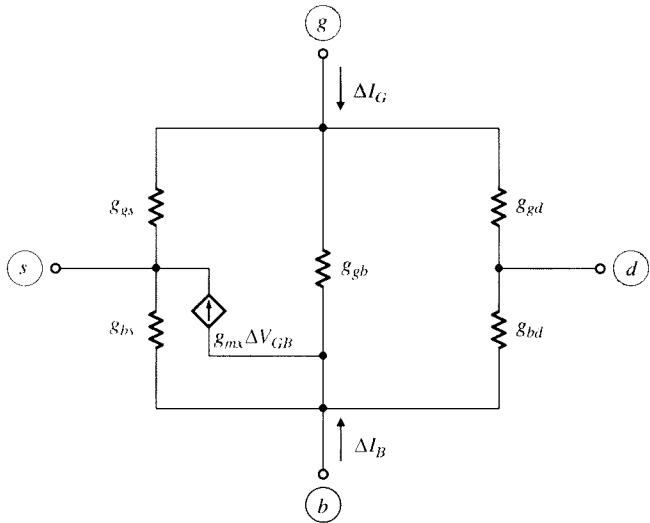


FIGURE 7.5
Adding the small-signal model for body current to the model of Fig. 7.4.

This equation can be modeled by a simple circuit that, when added to that of Fig. 7.4, gives the circuit in Fig. 7.5. As seen, we have used the existing small-signal conductance g_{gb} for part of the body current, and we used an extra term, $g_{mx} \Delta V_{GB}$, to take care of the difference, so that the total body current is equal to that given by (7.2.19). The voltage-controlled current source diverts this difference to the source, so that it does not interfere with the gate current. It is easy to verify that the combined circuit in Fig. 7.5 gives both the gate and the body small-signal currents exactly.

7.2.4 Complete Low-Frequency Small-Signal Model for the Intrinsic Part

If the model of Fig. 7.5 is added to that of Fig. 7.2, we obtain the complete model of Fig. 7.6. This model is a valid representation of the small-signal behavior of the entire MOS transistor with drain-to-source, gate, and body currents. This model relates only *small changes* of voltages to *small changes* of currents. The values of its parameters can be determined from the values of the total (bias) voltages, as will be discussed in the next several sections.

It is reasonable to ask whether the parts of the model, having been obtained separately, interfere with each other when they are combined. The reader is encouraged to write equations for ΔI_{DS} , ΔI_G , and ΔI_B for the circuit of Fig. 7.6, to verify that the equations for these currents are equivalent to the expressions obtained, and thus no interference is present. The current ΔI_{DS} is marked on Fig. 7.6.

Note that the small-signal drain current, ΔI_D , is *not* equal to the drain-to-source current ΔI_{DS} , but rather contains gate current and body current components. This is as it should be, as can be deduced from Secs. 5.8 and 5.12. The current I_{DS} is not externally observable, but expressions for it exist, as we have seen in previous chapters, and these can be used in definitions (7.2.4)–(7.2.6).

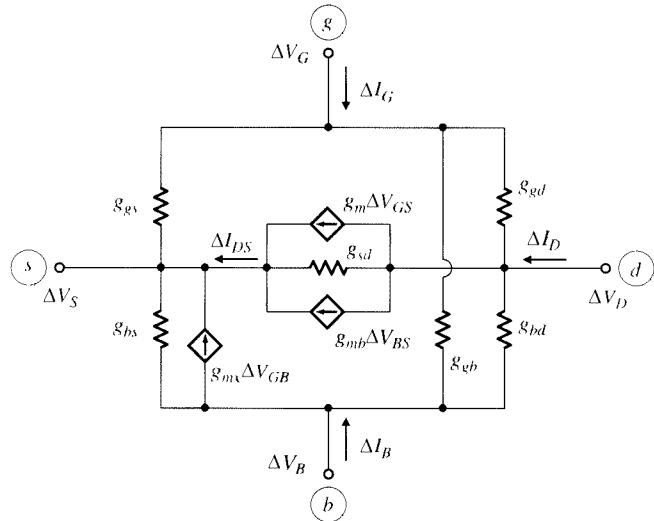


FIGURE 7.6
Complete low-frequency small-signal model for the intrinsic part.

The values of all nine small-signal parameters in Fig. 7.6 can be deduced from measurements. This is rather obvious for the parameters related to the gate and body currents; all one has to do is replace the differentials in the corresponding definitions by small finite differences, which are measurable. In the case of g_m , g_{mb} , and g_{ds} , however, this is not so obvious, as definitions (7.2.4)–(7.2.6) involve I_{DS} rather than the externally measurable I_D . Nevertheless, it can be shown that these parameters, too, can be determined from external measurements; the approach is sketched in Prob. 7.4b.

Although obtained gradually so as to make evident the physical meaning of the various elements, it turns out that the model of Fig. 7.6 is very general; in fact, it can be shown that this model can represent any four-terminal device at low frequencies as long as the parameter values are chosen correctly, and in fact, this representation becomes exact as the frequency is reduced to zero, in which case capacitive effects (see Sec. 7.3) can be neglected. The model needs no more, and no less, than nine parameters for this completeness to be possible. These statements will be proven in Sec. 8.2.1.

Alternative Definitions Although we have defined the small-signal parameters in the model of Fig. 7.6 in terms of derivatives with respect to terminal voltage differences, one can define these quantities by using instead terminal voltages V_S , V_G , V_D , and V_B with respect to an arbitrary reference. One can derive the new definitions mathematically, by expressing the voltages in the definitions in terms of these terminal voltages, and keeping three of these voltages constant at a time (Prob. 7.4). For example, if we keep V_S , V_G , and V_B constant and vary V_D , we have $\Delta V_{DS} = \Delta V_D$, and in (7.2.6), the partial derivative can be with respect to V_D rather than V_{DS} .

One can accomplish the same thing by using the small-signal equivalent circuit of Fig. 7.6. As an example, assume that V_S , V_G , and V_B are held constant, while V_D is varied. This means that the changes ΔV_S , ΔV_G , and ΔV_B in Fig. 7.6 are zero, and only

ΔV_D is nonzero. It can be seen that the only elements that carry currents are g_{sd} , g_{gd} , and g_{bd} ; all other conductances have zero voltage across them, and all transconductance sources are controlled by zero voltages and can be removed. From the remaining circuit, we have $\Delta I_{DS} = g_{sd} \Delta V_D$, $-\Delta I_G = g_{gd} \Delta V_D$, and $-\Delta I_B = g_{bd} \Delta V_D$. From these equations, we find $g_{sd} = \Delta I_{DS}/\Delta V_D$, $g_{gd} = -\Delta I_G/\Delta V_D$, and $g_{bd} = -\Delta I_B/\Delta V_D$. Letting the finite differences approach zero, we obtain the corresponding partial derivatives. We can repeat this with the other terminals, activating one of them at a time (Prob. 7.4). This results in the following definitions:

$$g_m = \left. \frac{\partial I_{DS}}{\partial V_G} \right|_{V_S, V_D, V_B} \quad (7.2.25a)$$

$$g_{mb} = \left. \frac{\partial I_{DS}}{\partial V_B} \right|_{V_S, V_G, V_D} \quad (7.2.25b)$$

$$g_{sd} = \left. \frac{\partial I_{DS}}{\partial V_D} \right|_{V_S, V_G, V_B} \quad (7.2.25c)$$

$$g_{gs} = - \left. \frac{\partial I_G}{\partial V_S} \right|_{V_G, V_D, V_B} \quad (7.2.26a)$$

$$g_{gb} = - \left. \frac{\partial I_G}{\partial V_B} \right|_{V_S, V_G, V_D} \quad (7.2.26b)$$

$$g_{gd} = - \left. \frac{\partial I_G}{\partial V_D} \right|_{V_S, V_G, V_B} \quad (7.2.26c)$$

$$g_{bs} = - \left. \frac{\partial I_B}{\partial V_S} \right|_{V_G, V_D, V_B} \quad (7.2.27a)$$

$$g_{bg} = - \left. \frac{\partial I_B}{\partial V_G} \right|_{V_S, V_D, V_B} \quad (7.2.27b)$$

$$g_{bd} = - \left. \frac{\partial I_B}{\partial V_D} \right|_{V_S, V_G, V_B} \quad (7.2.27c)$$

We will now derive expressions for the various small-signal parameters, considering each region of inversion separately. (Models valid in all regions of inversion will be considered in Sec. 7.2.8.) For brevity, the words *small signal* may be omitted when referring to a small-signal parameter when there is no chance for confusion, e.g., g_{sd} will be referred to as the *source-drain conductance*. Unless noted otherwise, in the derivations that follow no impact ionization or gate current will be assumed to be present.

7.2.5 Strong Inversion

Gate Transconductance Consider a long-channel device with constant effective mobility. Using either the complete strong-inversion model of Sec. 4.7.1 with $V_{GB} = V_{GS} + V_{SB}$, or the simplified model of Sec. 4.7.3 in the definition for g_m (7.2.4), we find

$$g_m = \frac{W}{L} \mu C'_{ox} V_{DS}, \quad V_{DS} \leq V'_{DS} \quad (7.2.28a)$$

$$= \frac{W}{L} \mu C'_{ox} V'_{DS}, \quad V_{DS} > V'_{DS} \quad (7.2.28b)$$

In the saturation region, the simplified model (4.7.24b) gives for the current $I_{DS} = \frac{1}{2}(W/L)\mu C'_{ox}(V_{GS} - V_T)^2/\alpha$. This, along with (7.2.4) produces the following equivalent forms[†] for g_m :

$$g_m = \frac{W}{L} \frac{\mu C'_{ox}}{\alpha} (V_{GS} - V_T), \quad V_{DS} > V'_{DS} \quad (7.2.29a)$$

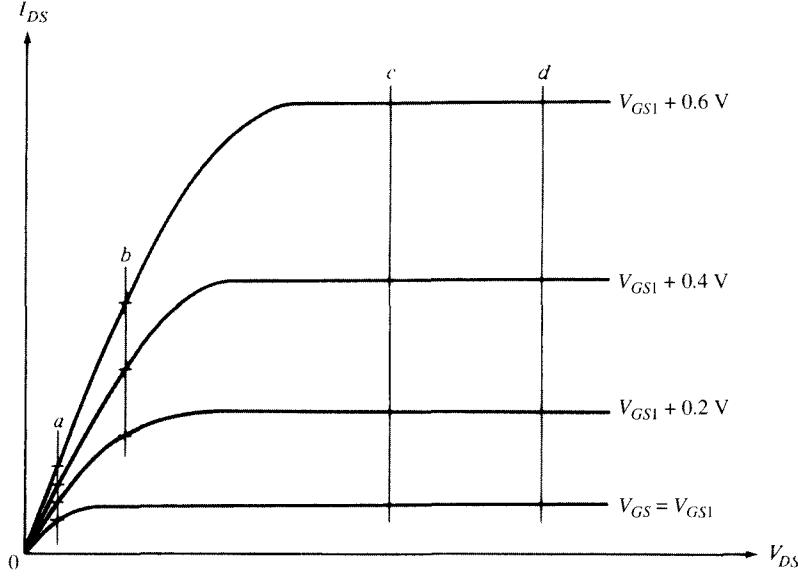
$$= \sqrt{2 \frac{W}{L} \frac{\mu C'_{ox}}{\alpha} I_{DS}}, \quad V_{DS} > V'_{DS} \quad (7.2.29b)$$

$$= \frac{2I_{DS}}{V_{GS} - V_T}, \quad V_{DS} > V'_{DS} \quad (7.2.29c)$$

As noted from (7.2.28a), in nonsaturation g_m is independent of V_{GS} . This is illustrated graphically in Fig. 7.7. We assume that the V_{GS} step used to obtain successive curves was fixed, so that g_m can be estimated from the spacing of the curves if the V_{GS} steps are small. As seen, for lines *a* and *b* the spacing is independent of V_{GS} but depends on V_{DS} , as expected from (7.2.28a). In saturation, the situation is reversed, as seen from (7.2.28b) and illustrated by lines *c* and *d*. Now g_m is independent of V_{DS} but depends on V_{GS} (through V'_{DS}).

Various corrections can be applied to the previous equations to increase their accuracy in the presence of higher order effects. For example, if the effective mobility is not constant with V_{GS} , differentiating I_{DS} to produce g_m will lead to an additional term (Prob. 7.6). An analogous correction should be applied if channel length modulation cannot be neglected. For example, a first-order correction results when obtaining g_m by differentiating (5.3.9).

[†] It is interesting to note that, at $V_{GS} = V_T$, at which point, some nonzero moderate-inversion current I_{DS} flows, these three equivalent expressions give g_m values of zero, $\sqrt{2(W/L)(\mu C'_{ox}/\alpha)I_{DS}}$, and infinity! This is a good example of the absurd results that can be obtained if modeling expressions are used outside their intended region of validity.

**FIGURE 7.7**

A family of I_{DS} - V_{DS} curves obtained for fixed V_{GS} increment. In nonsaturation, spacing is proportional to V_{DS} but independent of V_{GS} (lines *a*, *b*); in saturation, spacing is independent of V_{DS} but depends on V_{GS} (lines *c*, *d*).

Short-channel effects can strongly affect the value of all small-signal parameters. An extreme example is what happens to g_m in the presence of velocity saturation. The saturation drain current is then roughly given by (5.2.16b). Differentiating that equation with respect to V_{GS} gives

$$g_m \approx WC'_{\text{ox}}\mu\mathcal{E}_c, \quad \text{velocity saturation} \quad (7.2.30a)$$

$$\approx WC'_{\text{ox}}|v_d|_{\max}, \quad \text{velocity saturation} \quad (7.2.30b)$$

Substrate Transconductance Using the definition of g_{mb} in (7.2.5) with the complete strong inversion model of Sec. 4.7.1, with $V_{DB} = V_{DS} + V_{SB}$, we get

$$g_{mb} = b_g g_m \quad (7.2.31)$$

where

$$b_g = \frac{\gamma}{\sqrt{\phi_0 + V_{SB}} + \sqrt{\phi_0 + V_{SB} + V_{DS}}}, \quad V_{DS} \leq V'_{DS} \quad (7.2.32a)$$

$$= \frac{\gamma}{\sqrt{\phi_0 + V_{SB}} + \sqrt{\phi_0 + V_{SB} + V'_{DS}}}, \quad V_{DS} > V'_{DS} \quad (7.2.32b)$$

which can be approximated well by the following empirical expression

$$b_g \approx \frac{\gamma}{2\sqrt{\phi_0 + V_{SB} + k_g(1-\eta)V'_{DS}}} \quad (7.2.33)$$

where, for compactness, we have used the parameter η , defined in (4.7.25) and plotted in Fig. 4.25. With $k_g \approx 0.4$, it can be verified by plotting that this expression approximates (7.2.32) very well.

If V_{DS} is small, it can be neglected in (7.2.32a). Similarly, if V_{GS} is small, V'_{DS} will also be small, so V'_{DS} can be neglected in (7.2.32b) or (7.2.33). We thus obtain the following useful approximation:

$$b_g = \frac{g_{mb}}{g_m} \approx \frac{\gamma}{2\sqrt{V_{SB} + \phi_0}} = \frac{dV_T}{dV_{SB}} = \alpha_1 - 1 \approx n - 1, \quad \text{low } V_{DS} \text{ or } V_{GS} \quad (7.2.34)$$

where, to obtain the last parts, we used (4.7.20), (4.7.27), and (4.8.14). Using the expression for γ from (3.4.3), as well as (2.2.4) and (3.4.7), it is easy to show that the above relation gives

$$b_g = \frac{g_{mb}}{g_m} \approx \frac{\epsilon_s}{\epsilon_{ox}} \frac{t_{ox}}{d_{Bm}}, \quad \text{low } V_{DS} \text{ or } V_{GS} \quad (7.2.35)$$

where d_{Bm} is the depth of the depletion region, which is nearly uniform if V_{DS} is small (or if V_{GS} , and thus V'_{DS} , is small). The ratio g_{mb}/g_m can be thought of as a measure of the relative control of the “back gate” (substrate) and the “front” gate. As seen in (7.2.35), the thinner the oxide, the smaller the value of g_{mb}/g_m . This makes sense, since then the “front gate” is close to the channel and its control on it is strong. Thus the control of the “back gate” is smaller in a relative sense. The opposite is true if d_{Bm} is small. Then the edge of the depletion region is close to the channel and the substrate’s relative influence is large. In addition, the permittivities enter the picture. Large oxide permittivity means large control of the gate on the channel through the oxide, hence a small relative influence of the “back gate” (small g_{mb}/g_m). The opposite is true for large ϵ_s , which implies strong substrate control. The quantity g_{mb}/g_m is typically 0.1 to 0.3, but can rise significantly above those values for devices with large γ , operated with low V_{SB} .

It is not useful to derive expressions for g_{mb}/g_m using the simplified model of (4.7.24), since the dependence of α on V_{SB} , discussed in Sec. 4.7.3, is *approximate*. It was developed to provide reasonable accuracy for I_{DS} , with no regard to the derivative of I_{DS} with respect to V_{BS} . “Reasonable accuracy” for a function does not necessarily imply reasonable accuracy for its derivatives. Such problems are common when differentiation of empirical or semiempirical expressions is attempted, and one must exercise caution. We will thus avoid obtaining g_{mb} by direct differentiation of the simplified drain current expression[†] and will rely instead on (7.2.31)–(7.2.33).

[†] Such direct differentiation leads to (7.2.31) with b_g from (7.2.33) with $k_i = 0$. It thus fails to predict the variation of b_g with V_{DS} .

In the presence of effects not considered in the preceding discussion, such as substrate nonuniformity (Sec. 9.3) or small dimension effects (Chap. 5), the previous expression cannot be expected to be accurate. For example, g_{mb}/g_m can become very small in short-channel devices, because the trapezoid in Fig. 5.7 becomes close to a triangle, cutting off substrate control of the channel. The determination of accurate values of small-signal parameters in such cases requires using current expressions that include such effects, in the small-signal parameter definitions, e.g., (7.2.4) and (7.2.5). The resulting expressions are usually very complicated and thus of limited value for hand calculations. One usually resorts to numerical differentiation of accurate current expressions. In circuit simulators, small-signal analysis is usually based on complete, but complex, expressions for the small-signal parameters, sometimes generated automatically using symbolic differentiation programs.

Source-Drain Conductance The small-signal source-drain conductance g_{sd} ^{4,10-18,21-44} is the slope of the I_{DS} - V_{DS} characteristics with V_{GS} and V_{SB} held constant. Applying the definition (7.2.6) of g_{sd} to the nonsaturation complete model [(4.7.2b), with $V_{DB} = V_{DS} + V_{SB}$] gives

$$g_{sd} = \frac{W}{L} \mu C'_{ox} \left(V_{GS} - V_{DS} - V_{FB} - \phi_0 - \gamma \sqrt{V_{DS} + V_{SB} + \phi_0} \right), \quad V_{DS} \leq V'_{DS} \quad (7.2.36)$$

The simplified model (4.7.24) gives

$$g_{sd} = \frac{W}{L} \mu C'_{ox} [V_{GS} - V_T - \alpha V_{DS}], \quad V_{DS} \leq V'_{DS} \quad (7.2.37)$$

The two expressions are equivalent at $V_{DS} = 0$.

In the saturation region, simple long-channel drain current models are totally inadequate for use in the definition of g_{sd} in (7.2.6). This is illustrated in Fig. 7.8.⁴⁵ In *a*, the solid line represents a measured characteristic. The broken line represents a model, the parameters of which have been adjusted so that I_{DS} is predicted relatively accurately. Indeed, as shown, the error in predicting I_{DS} values is at most a few percent. In *b*, the *slopes* of the two curves are shown—the slope g_{sd} error exceeds 50 percent! This can have very serious consequences in analog circuit design.

Thus, in order to predict g_{sd} accurately in saturation, one needs a model that tracks the nuances of I_{DS} in that region. Unfortunately, this means that “second-order” effects, such as channel length modulation and drain-induced barrier lowering (all discussed in Chap. 5) become first order when it comes to evaluating g_{sd} . Attempting to develop reasonably simple expressions for g_{sd} in the presence of all these effects simultaneously is futile. In what follows, we will consider some of these phenomena separately, to provide some intuition through the use of simple relations.

Let us begin with channel-length modulation (CLM), discussed in Sec. 5.3. This effect can be dominant compared to other second-order effects in long-channel

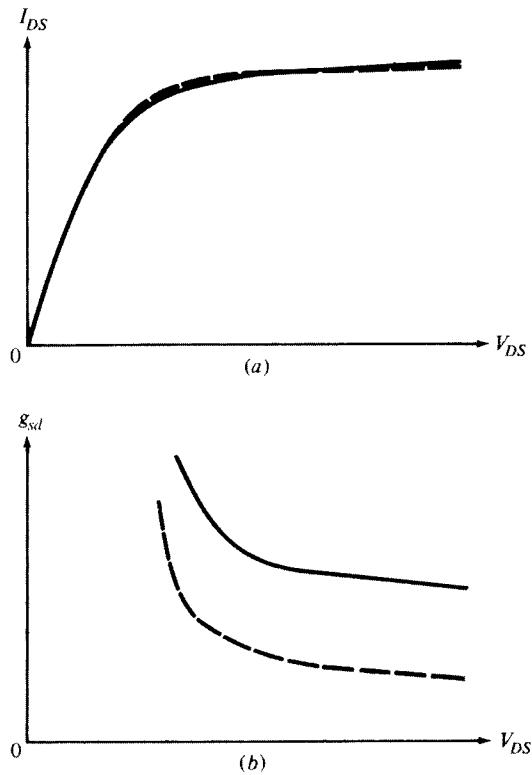


FIGURE 7.8
(a) I_{DS} - V_{DS} characteristics; solid line: measured; broken line: model; (b) slopes obtained from (a)⁴⁵
© 1984 by IEEE).

devices, in which the length of the “pinchoff” region, l_p , is much smaller than the channel length L . We can then use (5.3.2) in (7.2.6) to obtain

$$g_{sd} \approx I'_{DS} \frac{1}{L} \frac{\partial l_p}{\partial V_{DS}}, \quad V_{DS} > V'_{DS}, \quad \text{CLM only, long channels} \quad (7.2.38)$$

The expression that results from this depends on the model used for l_p . Models for this quantity resulting from one-dimensional analysis (Prob. 7.10) cannot be expected to result in accurate expressions because, as discussed in Sec. 5.3, the situation is two-dimensional in the pinchoff region. If, instead, we use the pseudo-two-dimensional analysis result (5.3.8) in (7.2.38), we obtain

$$g_{sd} \approx \frac{I'_{DS}}{V_A(V_{DS})} \quad V_{DS} > V'_{DS}, \quad \text{CLM only, long channels} \quad (7.2.39)$$

where

$$V_A(V_{DS}) = \frac{L}{l_a} [V_E + (V_{DS} - V'_{DS})] \quad (7.2.40)$$

where l_a is given by (5.3.6) and depends on oxide thickness and junction depth, and V_E is a quantity usually treated as a fitting parameter, with typical values of less than 1 V, best determined from measurements. Equation (7.2.39) has been put in the form often used in hand analysis for circuit design; however, note that in contrast to what is often assumed in circuits texts, the quantity V_A in the denominator (often called the Early voltage) is not constant, but rather is a strong function of V_{DS} .

Let us now consider the effect of drain-induced barrier lowering, or DIBL (Sec. 5.5), on g_{sd} . In order to reveal the main effects of this phenomenon, we will neglect CLM in the following discussion. We will assume that no velocity saturation occurs, and that the device can be described in saturation by the simple square-law equation

$$I_{DS} = \frac{W}{L} \frac{\mu C'_ox}{2\alpha} [V_{GS} - \hat{V}_T(V_{DS})]^2, \quad V_{DS} > V'_{DS} \quad (7.2.41)$$

where $\hat{V}_T(V_{DS}) = V_T + \Delta V_{T,DIBL}(V_{DS})$ is the effective threshold voltage incorporating the DIBL effect (Sec. 5.5). It was seen that \hat{V}_T decreases as V_{DS} is raised, even in the saturation region. This gives a positive slope to the I_{DS} vs. V_{DS} plot, i.e., a positive g_{sd} . Differentiating (7.2.41) with respect to V_{DS} , we obtain

$$g_{sd} = \frac{W}{L} \frac{\mu C'_ox}{\alpha} [V_{GS} - \hat{V}_T(V_{DS})] \left(-\frac{\partial \hat{V}_T}{\partial V_{DS}} \right) \quad (7.2.42)$$

We will relate this result to the gate transconductance g_m . Using (7.2.41) in (7.2.4), we find that g_m is given by (7.2.29a), with V_T replaced by $\hat{V}_T(V_{DS})$. From this and (7.2.42), we obtain

$$g_{sd} = g_m \left(-\frac{\partial \hat{V}_T}{\partial V_{DS}} \right), \quad V_{DS} > V'_{DS}, \quad \text{DIBL only} \quad (7.2.43)$$

For example, using $\hat{V}_T = V_T + \Delta V_{T,DIBL}$ with $\Delta V_{T,DIBL}$ from (5.5.1) and (5.5.2), we obtain

$$\frac{g_{sd}}{g_m} \approx \exp \left[-\sqrt{\beta_3 \frac{\epsilon_{ox}}{\epsilon_s}} \frac{L}{\sqrt{t_{ox} d_B}} \right] \quad (7.2.44)$$

Qualitatively, the behavior predicted by this equation makes sense. As follows from the discussion in Sec. 5.5, the drain acts as a “somewhat inefficient gate”; increasing the voltage on this “gate” increases the current and gives rise to a nonzero g_{sd} . The quantity g_{sd}/g_m in (7.2.44) is a measure of the “competition” between this “gate” and the normal gate (just like g_{mb}/g_m played a similar role as a measure of the relative

competition between the substrate and the normal gate previously). The smaller the L , the closer the whole of the inversion layer is to the drain region, and the stronger the influence of the latter; g_{sd}/g_m can be large in this case. The smaller t_{ox} , the closer the actual gate is to the inversion layer, and the more this gate wins out in the competition; then g_{sd}/g_m will be small. The relative influence of the two gates also depends on the permittivity of the media separating them from the inversion layer. Large permittivity implies a large effective capacitance, and hence, a strong influence. We note here that another parameter that affects g_{sd}/g_m is the junction depth, although this effect is not explicitly apparent in (7.2.44) (it would influence the empirical parameter β_3). The deeper the drain junction, the stronger will be the drain's "presence" as far as the inversion layer is concerned. Hence, g_{sd} , and g_{sd}/g_m , can be expected to *increase* with increasing drain junction depth.

Although the effects of CLM and DIBL were considered separately in the preceding discussion, in a real device they are present simultaneously. In general, DIBL can be dominant for V_{GS} close to V_T , with CLM becoming dominant at higher V_{GS} values. For a complete description of g_{sd} , one would have to differentiate a complete expression for I_{DS} , including all short-channel effects. This can lead to very complicated expressions involving all device geometrical parameters, substrate doping, saturation velocity, and even inversion layer thickness.¹⁷ The resulting expression is very much dependent on the form of the model used to derive it.

The transition from nonsaturation to saturation is especially difficult to model (i.e., for values of V_{DS} close to V'_{DS}). As has been remarked in Sec. 4.7.1, strong-inversion models are not accurate in that region, since the drain end of the channel is then moderately inverted. Thus, if strong-inversion expressions are used in that region, errors in g_{sd} can result.

It should be emphasized that, if separate I_{DS} expressions are used to derive g_{sd} for different regions, it should be made sure that the resulting expressions for g_{sd} are continuous at the region boundaries, and that at least the first derivatives of g_{sd} with respect to each terminal voltage are also continuous (see also Chap. 10). In some computer simulator models, "smoothing functions" are used to produce a single I_{DS} expression out of the separate expressions for each region.⁴⁶

Figure 7.9 shows the small-signal conductance parameters g_m , g_{mb} , and g_{sd} plotted vs. V_{DS} for a fixed V_{GS} , using simple nonsaturation expressions derived above up to $V_{DS} = V'_{DS}$, and assuming that the parameters are constant above that point. The sharp corners are artificial and are the result of the simplifications involved in the modeling process. More accurate models would provide smooth curves. For example, if one models correctly the transition region between nonsaturation and saturation, and includes CLM in saturation, the g_{sd} plot would be of the form shown by the dashed line in the figure.

It is interesting to note that, as follows from Fig. 7.9,

$$g_{sd}|_{V_{DS}=0} \approx g_m|_{V_{DS}>V'_{DS}}, \quad \text{if } \alpha \approx 1 \quad (7.2.45)$$

assuming the same V_{GS} value.

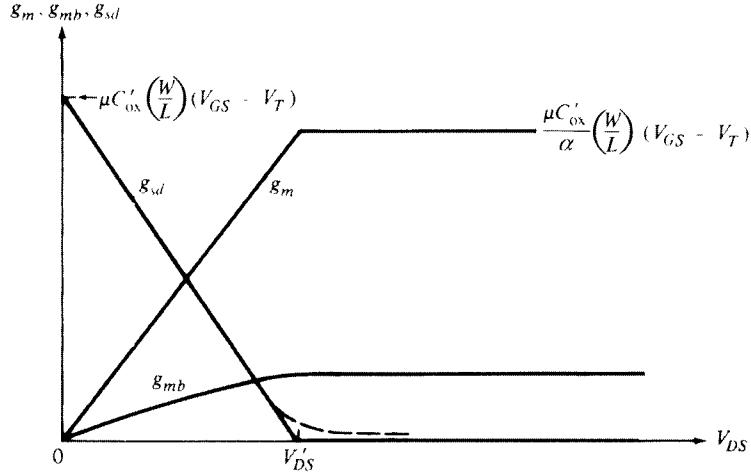


FIGURE 7.9
Small-signal gate transconductance, substrate transconductance, and drain conductance vs. V_{DS} for fixed V_{GS} and V_{SB} , as predicted by the simplified model (solid lines). The broken line represents a more accurate model for g_{sd} .

Other Conductances The rest of the parameters in Fig. 7.6, modeling the small-signal gate and body currents, can similarly be obtained if adequate expressions for them are known, by using the definitions in Sec. 7.2.3. Unfortunately, such expressions are complicated for the most part. Plots of these parameters will be given in Sec. 7.2.8.

An approximate expression can be obtained for g_{bd} in saturation, a parameter that will be seen to play an important role in what follows. In the definition of (7.2.22), the current I_B , defined as going into the body terminal, is the negative of the drain-to-body impact ionization current discussed in Sec. 5.8 and approximately given by (5.8.4). Using this expression and (7.2.27c), it can be shown that g_{bd} is approximately given by

$$g_{bd} \approx \frac{I_{DB}V_i}{(V_{DS} - V_{DS}')^2} \quad (7.2.46)$$

where, as discussed in Sec. 5.8, V_i is typically 10 to 30 V, and I_{DB} is normally several orders of magnitude below I_{DS} , especially for low-voltage processes. The preceding formula is very approximate and should be used only for rough calculations.

Output Conductance Consider a transistor connected as shown in Fig. 7.10a. Its *total* drain current, I_D , is shown vs. V_{DS} in Fig. 7.10b. Let us define the slope of this plot as the small-signal “output conductance” g_o :

$$g_o = \left. \frac{\partial I_D}{\partial V_{DS}} \right|_{V_{GS}, V_{SB}} \quad (7.2.47)$$

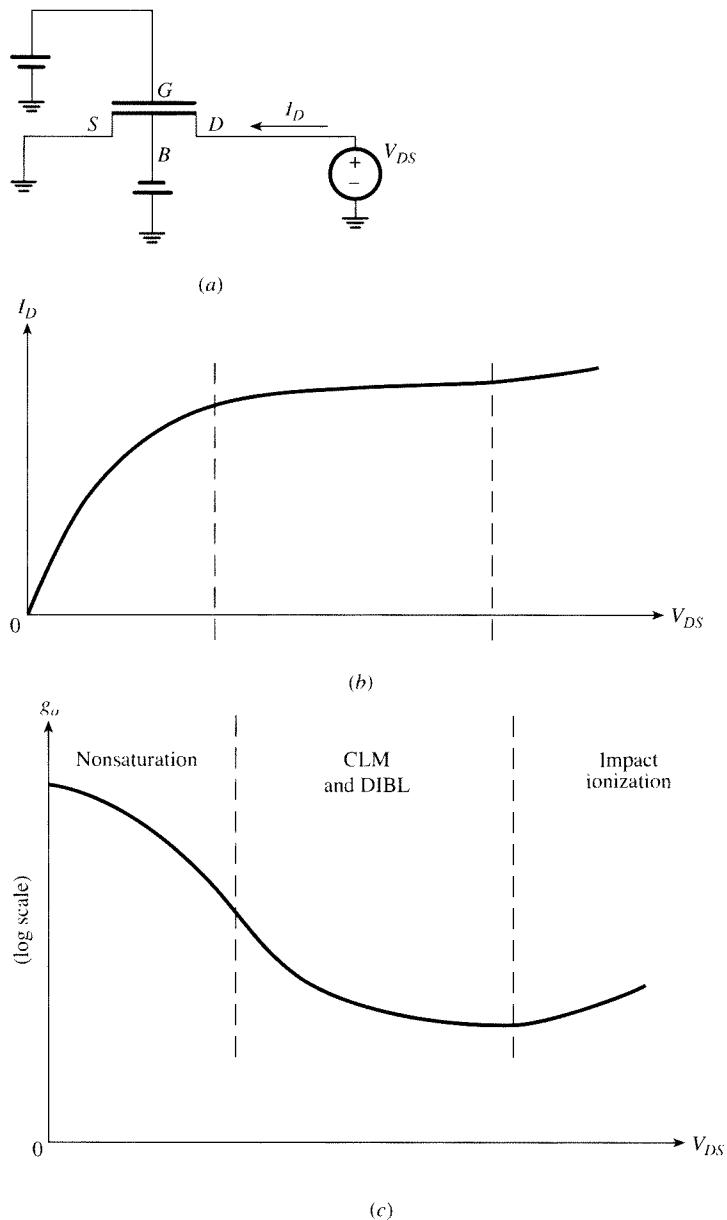


FIGURE 7.10
 (a) Transistor with fixed V_{GS} and V_{SB} , and a variable V_{DS} . (b) Total drain current I_D vs. V_{DS} for the connection in (a). (c) Output conductance [slope of the plot in (b)], with dominant effects indicated.

This quantity is shown plotted vs. V_{DS} in Fig. 7.10c. The main effects responsible for the behavior of this plot are indicated on the figure. The rising part toward the right is due to impact ionization. Limiting attention to the intrinsic device, it is seen

that the output conductance can be found from the small-signal equivalent circuit of Fig. 7.6, using in it $\Delta V_S = \Delta V_G = \Delta V_B = 0$, and a nonzero ΔV_D , corresponding to Fig. 7.10a. This gives $g_o = g_{vd} + g_{gd} + g_{bd}$. In addition, the externally observed output conductance can be affected by the extrinsic body resistance, as will be explained shortly. Small-signal output conductance is one of the most difficult quantities to model accurately.

The quantity g_o is shown vs. V_{DS} , with V_{GS} as a parameter, in Fig. 7.11, for a short- and long-channel device.³⁵ For the former, the presence of strong CLM and DIBL effects keeps g_{sd} large, no matter what bias values are used. In contrast, the long-channel device g_{sd} can vary over a wide range and can be made very small for V_{DS} of a couple of volts. This helps reveal the beginning of impact ionization effects at the right. To capture such effects, which can have consequences in analog circuit design, it is important to include g_{bd} in the small-signal equivalent circuit of the intrinsic device. Fortunately, the low supply voltages used in modern processes usually prevent such effects from seriously affecting the output conductance.

Effect of Extrinsic Resistances The preceding discussion has been limited to intrinsic effects. However, the behavior seen at the device terminals includes the effects of extrinsic elements, which sometimes can be significant. Two examples will be given here.

Consider first the extrinsic source resistance, discussed in Sec. 6.8.2. Let the total value of this resistance (assumed voltage-independent for simplicity) be R_{se} . The drain-source current will cause across this resistance a voltage drop $R_{se}I_{DS}$, and thus the intrinsic part will not see the externally applied gate-source voltage V_{GS} , but rather a reduced effective voltage $V_{GS} - R_{se}I_{DS}$. This reduction will cause a decrease in I_{DS} , and a decrease in the gate transconductance g_m , compared with the values that would have been expected if R_{se} were zero. Working with the small-signal model of Fig. 7.2, and neglecting g_{mb} and g_{sd} , it can be shown that the effective transconductance, as observed externally, will be (Prob. 7.13):

$$g_{m,\text{eff}} = \frac{g_m}{1 + g_m R_{se}} \quad (7.2.48)$$

A second effect⁴⁷ is present due to the extrinsic resistance of the body material (Sec. 6.8.2). Assume that the body offers a resistance R_{be} to impact ionization current I_{DB} in Fig. 5.18. This causes a voltage drop $R_{be}I_{DB}$, which opposes the source-substrate voltage V_{SB} ; i.e., the intrinsic transistor now sees an effective voltage $V_{SB} - R_{be}I_{DB}$ between source and internal body. As an increasing drain voltage increases I_{DB} , the preceding effective voltage decreases, which decreases the threshold voltage due to the body effect. This causes an increase in I_{DS} , thus causing an increase in I_D over that predicted directly by I_{DB} . This effect is sometimes called the “substrate current induced body effect.” Using the small-signal model of Fig. 7.6, but inserting a resistance R_{be} between body and ground, and assuming that g_{gb} , g_{bs} , g_{mx} , and g_{bd} are all $\ll 1/R_{be}$, it can be shown that the effective output conductance, as

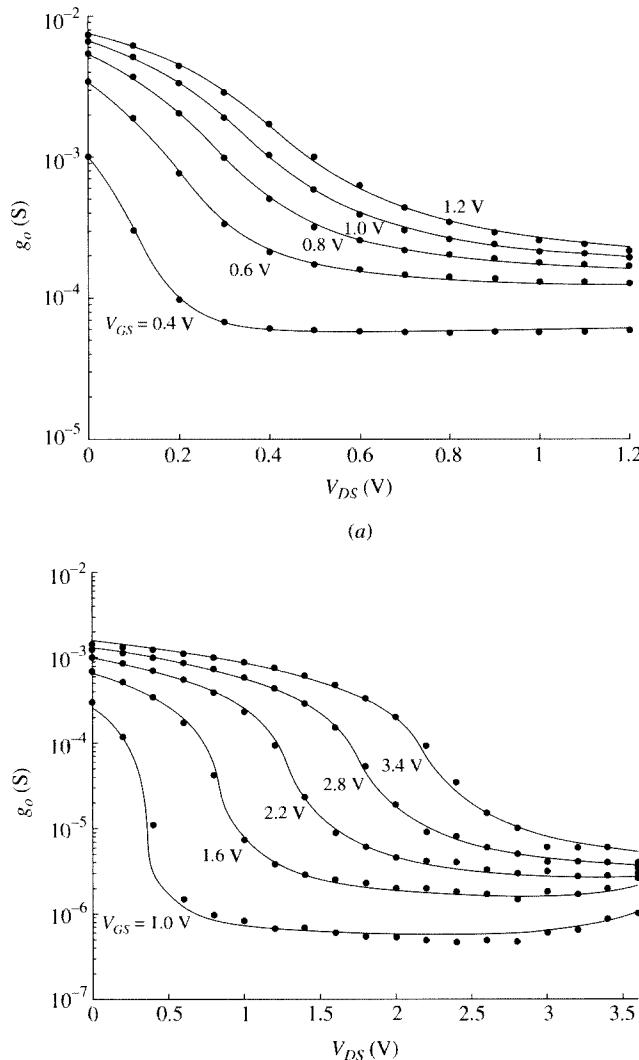


FIGURE 7.11
Measured output conductance g_o (on a log scale) vs. V_{DS} for (a) a short-channel device ($W = 2.5 \mu\text{m}$, $L = 0.09 \mu\text{m}$, $t_{ox} = 2 \text{ nm}$, $N_A = 7 \times 10^{17} \text{ cm}^{-3}$) and (b) a long-channel device ($W = 25 \mu\text{m}$, $L = 5 \mu\text{m}$, $t_{ox} = 10.5 \text{ nm}$, $N_A = 1.2 \times 10^{17} \text{ cm}^{-3}$).

observed externally, is given by (Prob. 7.13):

$$g_{o,\text{eff}} \approx g_{sd} + g_{gd} + g_{bd}(1 + g_{mb}R_{be}) \quad (7.2.49)$$

This quantity is difficult to calculate accurately, as the path over which I_{DB} flows changes with bias, thus changing the effective R_{be} . For devices built on a low-resistivity epitaxial layer, R_{be} is small and so is the effect.

The body extrinsic resistance can also affect the effective transconductance. When V_{GS} is increased, I_{DB} initially increases, as shown in Fig. 5.19. This decreases the effective source-substrate bias $V_{SB} - R_{be} I_{DB}$, and thus decreases the effective threshold, causing an excess increase in I_{DS} , above what would be expected in the absence of R_{be} . If the origin of this effect is not known, the increase in $\partial I_D / \partial V_{GS}$ might be interpreted as a “ g_m overshoot.” However, this would be misleading; as explained, this effect has nothing to do with g_m , as defined in (7.2.4) for the *intrinsic* device.[†]

As can be expected from this discussion, in analog design it is important to consider the effects of extrinsic resistances on small-signal behavior. Instead of having to derive “effective” values as described previously, it is best to just include R_{se} and R_{be} in the model for the complete device (Sec. 7.4); this will take care of the effects just discussed. It is important to be aware of such effects in order to be able to interpret measurement or simulation results. It should be noted that the values reported by circuit simulators for g_m , g_o , and other small-signal conductances are often the intrinsic values, not the effective values discussed previously.

7.2.6 Weak Inversion

Using the weak-inversion-current equation (4.8.11) in the definition for g_m (7.2.4), we find

$$g_m = \frac{1}{n} \frac{I_{DS}}{\phi_t} \quad (7.2.50)$$

where n is given by (4.8.14), assuming the effect of interface traps is negligible, a valid assumption for many modern devices (Sec. 2.7). If this effect is not negligible, n can have a larger value, which is best determined experimentally.

Equation (7.2.50) predicts that in weak inversion g_m/I_{DS} is constant and independent of the value of I_{DS} . This is the result of assuming that $I_{DS}(V_{GS})$ is exactly exponential, as in (4.8.11). Actually, though, the dependence is not exactly exponential, as can be seen from (4.8.6) and (4.8.7). Thus, some dependence of g_m/I_{DS} on I_{DS} can be expected in this region. For simplicity, in what follows, we will neglect this dependence.

Due to the approximately exponential $I_{DS}(V_{GS})$, g_m is known once the current is known, as seen from (7.2.50) (assuming n and the temperature are known), whereas in strong inversion this is not the case. This property in weak inversion is qualitatively the same as for a bipolar transistor. For the latter, the current in the forward active region is given by $I_C \approx I_O e^{V_{BE}/\phi_i}$, where V_{BE} is the base-emitter voltage and I_O is a characteristic current independent of V_{BE} . Thus, $g_m = dI_C/dV_{BE} = I_C/\phi_i$, or $g_m/I_C = 1/\phi_i$. Note that the corresponding quantity for the MOS transistor, $g_m/I_{DS} = 1/(n\phi_i)$, is always smaller since n is larger than unity. The value $1/\phi_i$, reached by the transconductance-to-current

[†] A g_m overshoot is possible in very small devices for an entirely different reason: the velocity overshoot discussed in Sec. 5.9.

ratio for the bipolar transistor but not for the MOS transistor, is sometimes called the *Boltzmann limit*.

The substrate transconductance will not be obtained from (4.8.11) since I'_M , V_M , and n depend on V_{SB} in a complicated manner. We will use instead (4.8.6) and (4.8.7) with $V_{GB} = V_{GS} - V_{BS}$ and $V_{DB} = V_{DS} - V_{BS}$. Then from the definition of g_{mb} , (7.2.5), we obtain (Prob. 7.14)

$$g_{mb} \approx \frac{n-1}{n} \frac{I_{DS}}{\phi_t} \quad (7.2.51)$$

Thus the ratio g_{mb}/g_m is given by

$$\frac{g_{mb}}{g_m} \approx n-1 \quad (7.2.52a)$$

$$\approx \frac{\gamma}{2\sqrt{V_{SB} + 2\phi_F}} \quad (7.2.52b)$$

$$\approx \frac{\epsilon_s t_{ox}}{\epsilon_{ox} d_B} \quad (7.2.52c)$$

where we have assumed negligible interface trap density, and where in the last expression d_B is the approximate width of the depletion region in weak inversion. This relation is of the same form as (7.2.35).

For g_{sd} we find, using the definition (7.2.6) in (4.8.11),

$$g_{sd} = \frac{e^{-V_{DS}/\phi_t}}{1 - e^{-V_{DS}/\phi_t}} \frac{I_{DS}}{\phi_t}, \quad \text{nonsaturation} \quad (7.2.53)$$

This equation predicts a g_{sd} that rapidly goes to zero with increasing V_{DS} . However, this neglects the influence of DIBL (Sec. 5.5). To see how DIBL affects g_{sd} , consider (4.8.11). In this equation, V_M stays close to the threshold voltage and through DIBL, decreases with increasing V_{DS} and causes an increase in the current. Working with this equation and assuming that V_M tracks V_T , it is easy to show that (7.2.43) and thus (7.2.44) are valid in weak inversion. Using in the latter (7.2.50), it can then be seen that g_{sd} is of the form

$$g_{sd} = \frac{I'_{DS}}{V_{AW}}, \quad V_{DS} > 5\phi_t \quad (7.2.54)$$

where the condition on the right has been imposed to ensure operation in the “flat” part of the curves in Fig. 4.30, and V_{AW} is given by

$$V_{AW} \approx n\phi_t \exp\left(\sqrt{\beta_3 \frac{\epsilon_{ox}}{\epsilon_s}} \frac{L}{\sqrt{t_{ox} d_B}}\right) \quad (7.2.55)$$

For short-channel devices, this quantity can be smaller than the corresponding quantity V_A in strong inversion.

7.2.7 Moderate Inversion

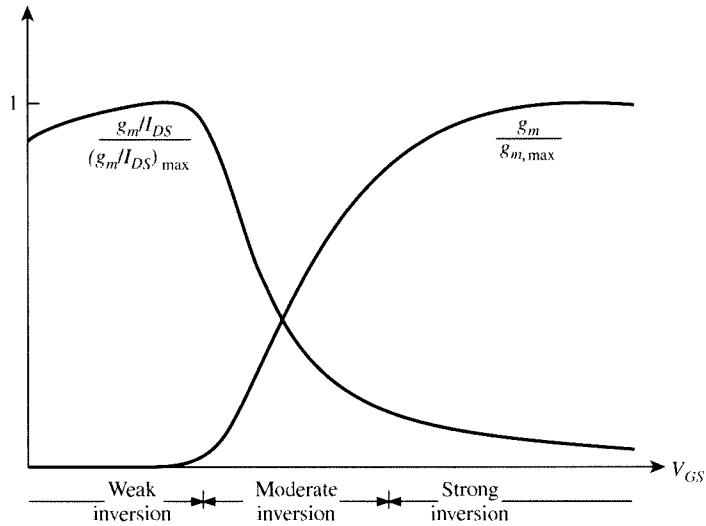
In moderate inversion, the behavior of small-signal parameters can be very different from that in weak and in strong inversion. To illustrate this, we will use g_m as an example and will consider the behavior of this parameter over a wide range of V_{GS} for a fixed V_{DS} . The results for g_m and for g_m/I_{DS} in saturation are shown in Fig. 7.12.⁴⁸ These results were obtained by using the all-region model of Sec. 4.3, which is valid in all regions of operation (see also Sec. 7.2.8). As seen, in strong inversion g_m is approximately constant.[†] In weak inversion, g_m/I_{DS} , instead, is approximately constant. This is expected from (7.2.50) and is a consequence of the practically exponential dependence of I_{DS} on V_{GS} . In moderate inversion, neither g_m nor g_m/I_{DS} can be considered approximately constant, as seen in the figure. Yet, circuit textbooks often use models that neglect the moderate-inversion region, thus taking weak- and strong-inversion regions to be adjacent. From Fig. 7.12, it is clear that this is a very poor approximation. Thus, using such models in analog circuit applications and for V_{GS} values in the vicinity of V_T can lead to *very* wrong predictions.

Unfortunately, there are as yet no analytically derived closed-form expressions for the small-signal conductances in the moderate-inversion region. Thus, one must resort to all-region models, discussed in the following subsection.

7.2.8 All-Region Models

In Sec. 4.3, it was shown that general expressions can be developed for I_{DS} that are valid in all regions of operation for a long-channel device (neglecting CLM). Such expressions can, in principle, be used to provide expressions for the small-signal conductances, which will be valid in all regions including moderate inversion and for the transition from nonsaturation to saturation in strong inversion.^{50–53} The difficulty here will be complexity. The drain current expression contains V_{GB} explicitly in (4.3.14), and also implicitly through ψ_{s0} and ψ_{sL} [see (4.3.15)]. As expected, the results are algebraically complicated. A different approach is now suggested. It can be shown using (4.5.2) that the quantities g_{sd} and g_{ss} , defined in (7.2.6) and (7.2.11), respectively, are given by⁵⁰

[†] More precisely, in strong inversion g_m reaches a maximum and then decreases with increasing V_{GS} due to the influence of V_{GS} on the effective mobility⁴⁹ (Sec. 4.11).

**FIGURE 7.12**

Saturation transconductance and transconductance-to-current ratio, both normalized to their maximum values, as a function of V_{GS} .

$$g_{sd} = \mu \frac{W}{L} (-Q'_{IL}) \quad (7.2.56)$$

$$g_{ss} = \mu \frac{W}{L} (-Q'_{I0}) \quad (7.2.57)$$

where Q'_{IL} and Q'_{I0} are the inversion layer charges per unit area at the drain end and at the source end of the channel, respectively. If desired, these can be expressed in terms of the surface potential by using (4.3.13). For example, g_{sd} becomes

$$g_{sd} = \frac{W}{L} \mu C_{ox}' \left(V_{GB} - V_{FB} - \psi_{sL} - \gamma \sqrt{\psi_{sL}} \right) \quad (7.2.58)$$

As a check, consider the nonsaturation region. Substituting ψ_{sL} by $V_{DB} + \phi_0 = V_{DS} + V_{SB} + \phi_0$ in the preceding equation, produces (7.2.36). In the transition region from nonsaturation to saturation, the use of precise values for ψ_{sL} in (7.2.58) can provide accurate values for g_{sd} in long-channel devices. Of course, this equation should not be used in saturation, as it ignores CLM, velocity saturation and DIBL.

Simpler results are possible if the preceding relations are applied to the simplified all-region models discussed in Sec. 4.4.4. As an example, we will assume operation

in saturation and will develop expressions valid in all regions of inversion. In saturation, Q'_{IL} is negligible. Using $Q'_{IL} = 0$ in (4.4.22b), solving this equation for Q'_{I0} in terms of I_{DS} , and using the result in (7.2.57) gives^{52,53} (Prob. 7.15):

$$g_{ss} = \frac{I_{DS}}{\phi_t} \frac{2}{1 + \sqrt{4 \frac{I_{DS}}{I_Z} + 1}}, \quad \text{saturation} \quad (7.2.59)$$

where

$$I_Z = \frac{W}{L} \mu C'_{ox} (2n\phi_t^2) \quad (7.2.60)$$

is a characteristic current.⁵⁴⁻⁵⁷ Further manipulation of the equations of the simplified model produces^{52,53}

$$g_m \approx \frac{g_{ss}}{n}, \quad \text{saturation} \quad (7.2.61)$$

Since, in the saturation region, g_{sd} is negligible compared with g_m and g_{mb} (we assume no short-channel effects), (7.2.12) gives

$$g_m + g_{mb} \approx g_{ss}, \quad \text{saturation} \quad (7.2.62)$$

Using (7.2.61) and (7.2.62), we obtain

$$g_{mb} \approx \frac{n-1}{n} g_{ss}, \quad \text{saturation} \quad (7.2.63)$$

As a check, consider $I_{DS} \ll I_Z$; then the preceding expressions for g_m and g_{mb} reduce to those obtained for weak inversion in (7.2.50) and (7.2.51). Or consider $I_{DS} \gg I_Z$; then g_m can be easily seen to reduce to the strong-inversion expression (7.2.29b), with n replacing α (the two quantities are of similar magnitude); similarly, g_{mb} approaches the value given by (7.2.34). Similar results have been produced earlier using interpolation techniques.⁵⁴⁻⁵⁷

These results are plotted, in normalized form, in Fig. 7.13.^{53,55,57,58} In the same plot, we show g_m plots from our earlier weak- and strong-inversion expressions. The point $I_{DS}/I_Z = 1$ falls in the moderate-inversion region. Clearly, the weak- and strong-inversion expressions fail for a range of about two orders of magnitude centered around this point. Instead, (7.2.59) has been extensively compared with measurements and has been found to be very accurate⁵³ throughout, for a variety of processes. This means⁵³ that, for a given temperature, g_{ss}/I_{DS} is fully determined by

the ratio I_{DS}/I_Z . Note that the maximum possible value of g_{ss} is I_{DS}/ϕ_t (the Boltzmann limit) and that of g_m is $I_{DS}/(n\phi_t)$.

Thus, in saturation, for a given bias current in any region of inversion, the determination of the transconductances becomes trivial! One only needs to find g_{ss} from (7.2.59) (or Fig. 7.13), and then find what part of g_{ss} is apportioned to each of g_m and g_{mb} , by using (7.2.61) and (7.2.63). In a similar manner, other small-signal parameters can also be expressed in terms of I_{DS}/I_Z .^{55,57} An example is shown in Table 7.1 at the end of this chapter.

In Fig. 7.14, $\log g_m$ is plotted vs. $\log I_{DS}$ for various values of W/L and for operation in the flat part of the I_{DS} - V_{DS} characteristics; a fixed V_{SB} is assumed. For each W/L value, the straight-line segment on the *right* corresponds to strong inversion and the straight-line segment on the *left* corresponds to weak inversion. The curved part corresponds to moderate inversion. In strong inversion, for a given value of I_{DS} , one can obtain a larger g_m by using a larger value for W/L , as expected from (7.2.29b). However, this is not true in weak inversion, as can be seen both in the figure and from (7.2.50). There, once I_{DS} is fixed, g_m is known (assuming a fixed value for n). This behavior is compared with that of a bipolar device in Fig. 7.14 (see comments on the latter in Sec. 7.2.6).

Plots of g_m , g_{mb} , and g_{sd} vs. V_{GS} for a fixed, large V_{DS} have the form shown in Fig. 7.15, assuming a constant mobility. Such plots show these quantities varying continuously over all regions of inversion. The order in which the various regions

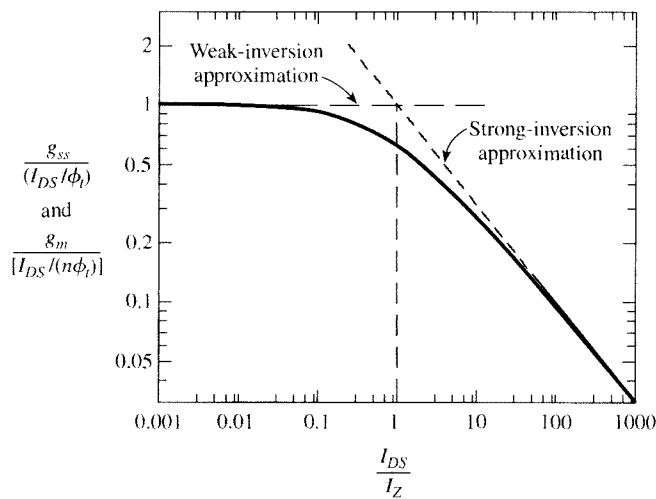
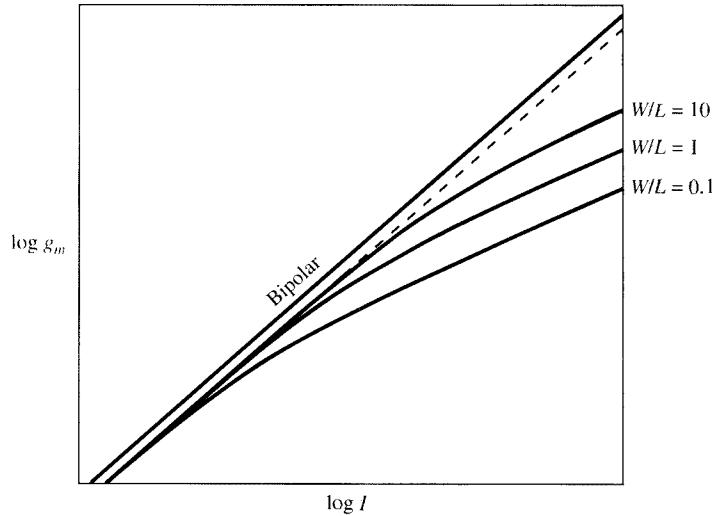
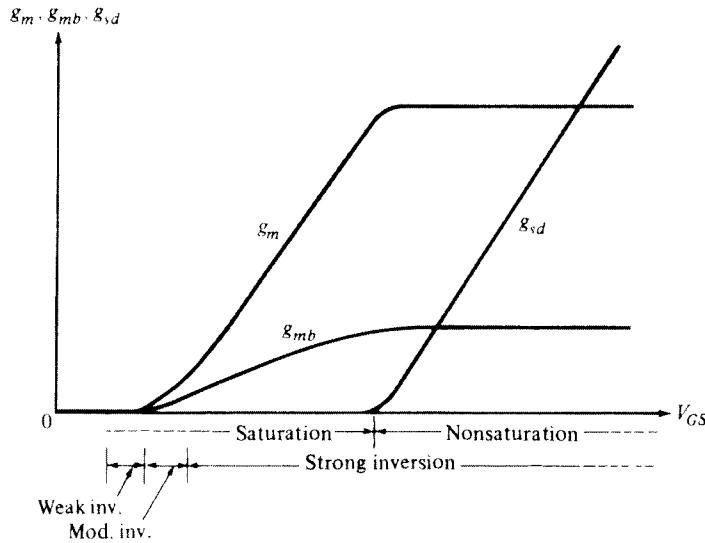


FIGURE 7.13

Source conductance and gate transconductance, normalized to their maximum values, vs. normalized drain-to-source current.^{53,55,57,58}

**FIGURE 7.14**

Comparison of $\log g_m$ vs. $\log I$ behavior for a bipolar transistor and for MOS transistors with various W/L values, with fixed V_{SB} .

**FIGURE 7.15**

Small-signal gate transconductance, body transconductance, and drain conductance as a function of V_{GS} for fixed V_{DS} and V_{SB} as obtained from the all-region model, for a long-channel device. A constant mobility is assumed.

are encountered as V_{GS} is increased has been discussed in Sec. 6.4.7 in conjunction with Fig. 6.8.

Figure 7.16 shows, on a logarithmic axis, all conductance and transconductance parameters in the model of Fig. 7.6, as they result from numerical differentiation. The wide range of values encountered is noted. Mobility dependence on gate field is included; the effect of this can be seen, for example, in the reduction of g_m at large V_{GS} values.

7.3 A Medium-Frequency Small-Signal Model for the Intrinsic Part

7.3.1 Introduction

When the small-signal voltages applied to a MOS transistor are varying fast, the small-signal terminal currents can be very different from those found using the circuit of Fig. 7.6. We now introduce a simple model that will predict such behavior as long as the frequency is not too high. The model achieves this by taking into account charge storage effects, assuming quasi-static operation.^{1,2,5,8,13,15,16,28,29,42–44,51–98} Not all such effects are included; additional charge storage effects will be considered in Chap. 8. Nevertheless, the model we are about to introduce here is important in its own right, and we present it in a self-contained manner for two reasons:

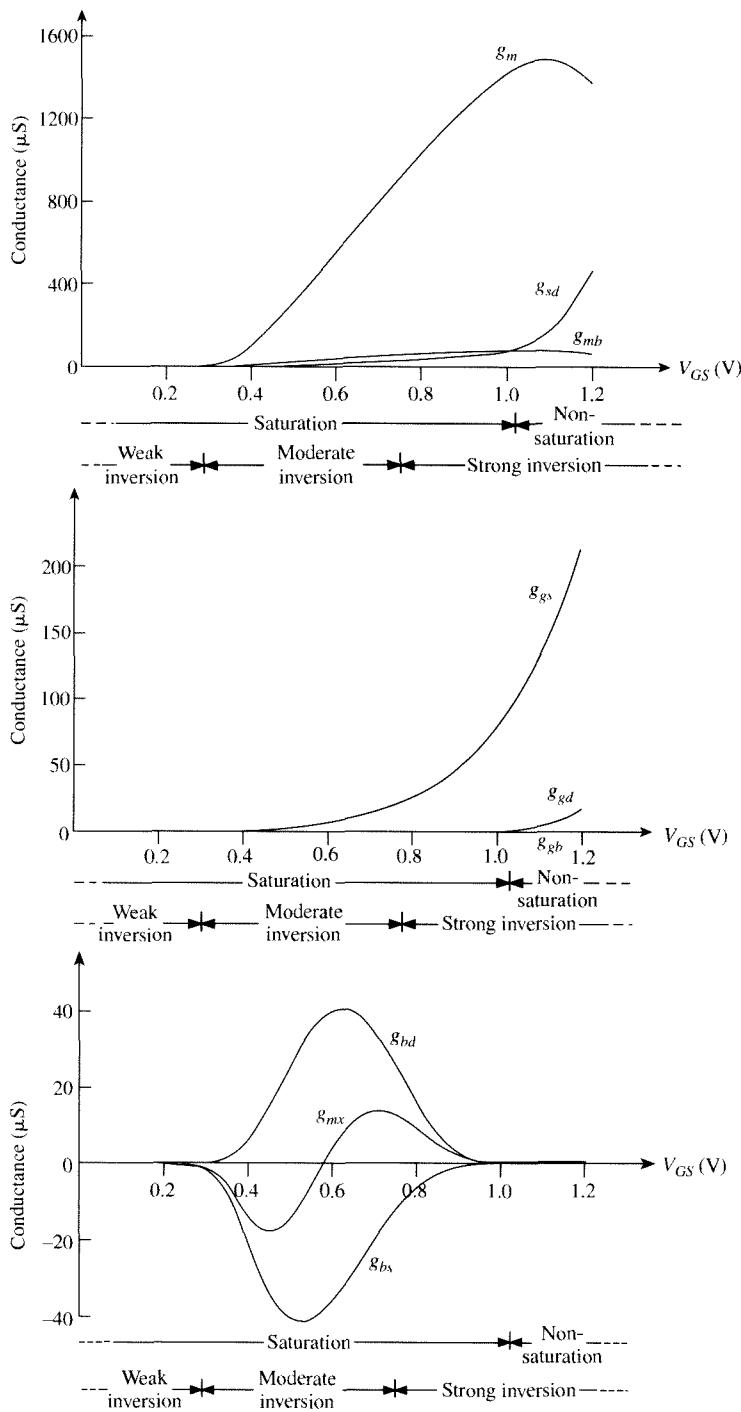
1. It is widely in use, because it achieves a reasonable balance between accuracy and complexity. Some readers might find it adequate for their purposes, so it was felt that they should not have to be subjected to the general treatment of Chap. 8 before they encountered a useful model.
2. The material in this section will be a useful foundation for the general treatment in Chap. 8. In fact, the more complete model in that chapter can be produced just by adding a few extra elements to the model we will develop here.

As before, we will consider only the *intrinsic* part of the transistor in this section. This fact will be understood without being indicated explicitly. Extrinsic device modeling will be introduced in Sec. 7.4. No impact ionization is assumed.

In this section, we will emphasize intuition. For more rigor, the reader will have to wait until Chap. 8.

7.3.2 Intrinsic Capacitances

The intrinsic part of a long-channel transistor (neglecting the two-dimensional effects very close to the source and drain regions) can be considered by itself by viewing it as a fictitious device in which the length of the source and drain regions has shrunk to zero, as shown in Fig. 7.17a. Our discussion will be more convenient if we define the voltages at the four terminals with respect to an arbitrary reference, as shown by the “ground” symbol in the figure. A subscript 0 is used to denote the “quiescent” values of the voltages and charges in Fig. 7.17a. Let us consider the effects of small changes of V_S on the gate and depletion region charges (Fig. 7.17b), the effect of small changes of V_D on the gate and depletion region charges (Fig. 7.17c), and the effect of small

**FIGURE 7.16**

Small-signal conductance and transconductance parameters (used in the model of Fig. 7.6) vs. V_{GS} .
 $t_{ox} = 1.5 \text{ nm}$, $N_A = 3 \times 10^{18} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 2 \mu\text{m}$, $V_{DS} = 0.6 \text{ V}$, $V_{SB} = 0 \text{ V}$.

416 OPERATION AND MODELING OF THE MOS TRANSISTOR

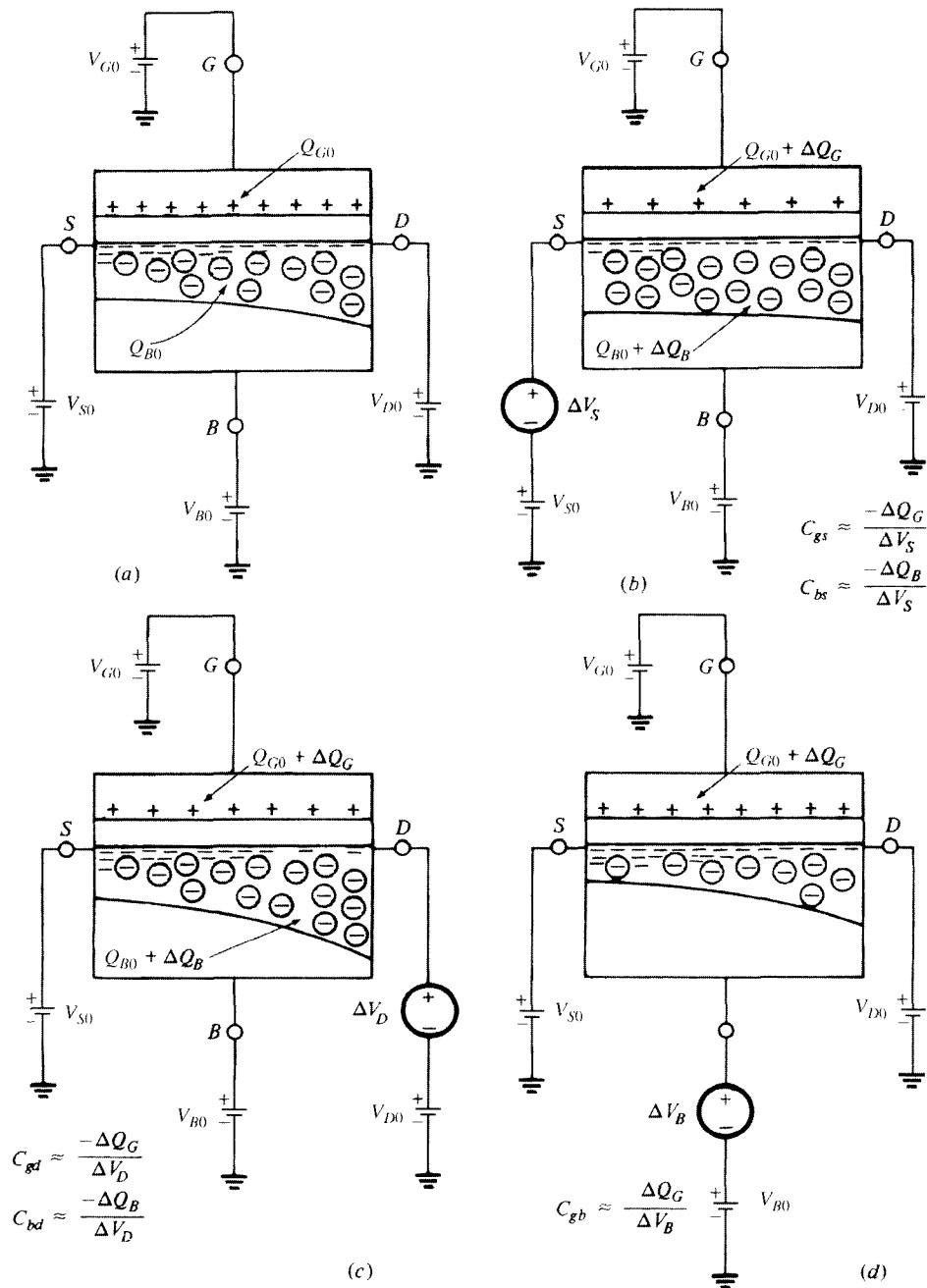


FIGURE 7.17

Measurement of intrinsic capacitances (in principle). The figure shows the intrinsic part of a transistor. (a) Transistor biased with four dc voltages at a certain operating point; (b) effect of a small increase in the source voltage; (c) effect of a small increase in the drain voltage; (d) effect of a small increase in the body voltage. The relations in the figures give the corresponding capacitance values.

changes of V_B on the gate charge (Fig. 7.17d). The details in these figures and the definitions next to them will be explained shortly. In order to discuss a concrete case, and to be able to draw the figures, we have assumed that the device operates in strong inversion, and that ΔV_S , ΔV_D , and ΔV_B represent increases. However, the definitions given are general; they apply in all regions and are independent of the sign of the voltage increments. All voltages are assumed constant before and after the change, and all cases in Fig. 7.17 are assumed to be in dc steady state. Thus, for example, ΔQ_G in Fig. 7.17b is the difference between the total steady-state gate charge when the source voltage is held fixed at $V_S + \Delta V_S$ and the steady-state gate charge when the source voltage is held fixed at V_S . For simplicity, the interface charge Q_o is not shown.

We now consider the experiment of Fig. 7.17b in detail. The potential across the oxide at various points decreases relative to that in Fig. 7.17a and, hence, the gate charge decreases too; thus ΔQ_G is negative. The relation between the cause (ΔV_S) and the effect (ΔQ_G) can be represented by the small-signal equivalent circuit of Fig. 7.18. In this circuit the voltage ΔV_S places a charge $C_{gs}\Delta V_S$ on the bottom plate and a charge $-C_{gs}\Delta V_S$ on the top. For the latter charge to represent the change ΔQ_G in Fig. 7.17b, we must have $-C_{gs}\Delta V_S = \Delta Q_G$. Hence $C_{gs} = -\Delta Q_G/\Delta V_S$, as stated next to Fig. 7.17b. More formally, we have

$$C_{gs} \equiv -\left. \frac{\partial Q_G}{\partial V_S} \right|_{V_G, V_D, V_B} \quad (7.3.1)$$

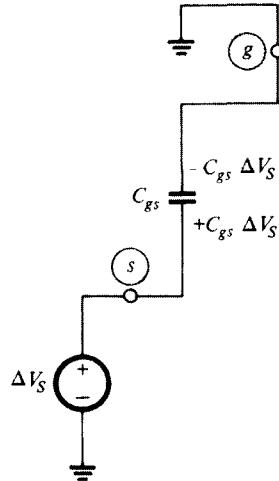
where to the right of the vertical line we show the voltages being held constant. The reader is urged to carefully consider the rationale given for the minus sign in the preceding definition.

It is important *not* to associate C_{gs} with any parallel plate structure in Fig. 7.17. C_{gs} is simply the value that the fictitious capacitor in Fig. 7.18 should have in order for the charge on its top plate to be the same as the charge *change* ΔQ_G in Fig. 7.17b. Note that, since ΔQ_G is negative when ΔV_S is positive, the value of C_{gs} is positive.[†]

Figure 7.17b also illustrates the capacitive effect of the source on the substrate. Increasing V_S increases the width of the depletion region, thus making the total charge there more negative, and hence ΔQ_B is negative. This is accomplished by more acceptor atoms being uncovered (in comparison with Fig. 7.17a), which means that holes must leave the bottom of the depletion region, which causes a total charge $|Q_B|$ to leave through the substrate terminal. This is equivalent to saying that a negative charge ΔQ_B enters the device from the substrate terminal. Reasoning as before, we define

$$C_{bs} \equiv -\left. \frac{\partial Q_B}{\partial V_S} \right|_{V_G, V_D, V_B} \quad (7.3.2)$$

[†] Throughout this section, all capacitances and charges are for the intrinsic part of the device only. For example, C_{gs} is the *intrinsic* gate-source capacitance and *not* the total gate-source capacitance of the device. It will be seen in Sec. 7.4 that an extrinsic capacitance caused by the physical proximity of the gate and the source must be added to C_{gs} to arrive at the total gate-source capacitance.

**FIGURE 7.18**

Meaning of C_{gs} , the small-signal intrinsic gate-source capacitance.

The effects of increasing V_D are entirely analogous to that of increasing V_S and are illustrated in Fig. 7.17c. We define

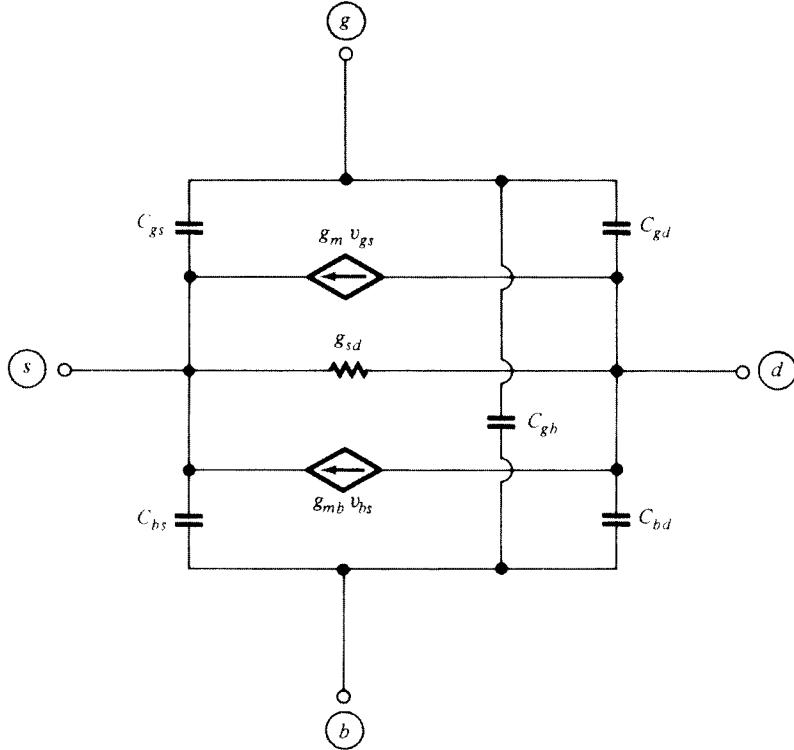
$$C_{gd} \equiv -\left. \frac{\partial Q_G}{\partial V_D} \right|_{V_G, V_S, V_B} \quad (7.3.3)$$

$$C_{bd} \equiv -\left. \frac{\partial Q_B}{\partial V_D} \right|_{V_G, V_S, V_B} \quad (7.3.4)$$

Finally, the effect of increasing V_B is illustrated in Fig. 7.17d. The increasing V_B causes positive charges to flow into the substrate terminal, which are partly balanced by a decrease in Q_G . This is equivalent to a negative ΔQ_G flowing into the gate. This is easier to understand in weak inversion, where the inversion layer charge is negligible. Then practically all the positive charge change ΔQ_B must be balanced by an opposite change in the gate charge; hence ΔQ_G will be negative. We define

$$C_{gb} \equiv -\left. \frac{\partial Q_G}{\partial V_B} \right|_{V_G, V_S, V_D} \quad (7.3.5)$$

All the previous five effects can be modeled in a manner analogous to that of Fig. 7.18. These effects can be included in a small-signal equivalent circuit by adding to it five capacitors. For example, if we do this with the circuit of Fig. 7.2, we obtain the circuit of Fig. 7.19. A more complete equivalent circuit would include the elements modeling the small-signal effects related to the gate and body currents, as in Fig. 7.6. These elements will not be shown explicitly in the equivalent circuits in this and the next chapter, to minimize clutter; but they will be understood to be there. It

**FIGURE 7.19**

Simple small-signal equivalent circuit for the intrinsic part of a MOS transistor. The elements associated with gate and body conductive currents (Fig. 7.6) are not shown for simplicity.

is important to note that the various elements in the resulting model *do not interfere with each other*. For example, keeping terminals g , d , and b at ground and applying ΔV_s at terminal s will cause a charge $-C_{gs}\Delta V_s$ to enter terminal g , thus modeling correctly the effect of the source on the gate. C_{gd} and C_{gb} , although connected to the gate, will *not* interfere since they will act as open circuits (the voltage across them is fixed). The reader can easily check that each of the remaining four capacitance effects is also modeled correctly by the corresponding capacitors and with no interference from other elements (Prob. 7.16). This “*noninterference*” is a very important property that should be carefully checked every time several elements—each meant to simulate a separate effect—are connected together to form a circuit model.

Since C_{gs} , C_{gd} , and C_{gb} represent the effect on the gate of each of the other three terminals, the circuit of Fig. 7.19 can model accurately the quasi-static small-signal effects one “sees looking into” the gate of a MOS transistor. Notice that for now we cannot prove a similar statement for the other terminals. (This subject will be discussed further in Chap. 8.) At this point, we will only say that, as it turns out by comparison to more complete models (Chap. 8), the model topology of Fig. 7.19 is satisfactory (as far as looking into any terminal, and as far as the effect of any terminal on any other is concerned), even if the terminal voltages vary continuously with

time, as long as their variation is slow enough. For sinusoidal small variations, this can be quantified by establishing an upper frequency limit of validity. The figure one comes up with depends on the accuracy desired, the region of operation, which terminals are being driven with signals, which terminal currents are of interest, whether the magnitude or the phase of those currents is of interest,[†] etc. Thus, it is not easy to give a single number. Some general indication can, nevertheless, be given by comparison to more sophisticated model topologies (Chap. 8). Thus, in strong inversion, whatever the criterion being used, the upper frequency limit of validity turns out to be proportional (but *not* equal) to the quantity

$$\boxed{\omega_o = \frac{\mu(V_{GS} - V_T)}{\alpha L^2}} \quad (7.3.6)$$

where the reason for the presence of the quantities in the right-hand side can be understood after higher-order models are discussed in Chap. 8. Thus, assume that satisfactory values have been obtained for all model elements at low frequencies. Then, without touching these element values, the model will continue to be valid up to about $0.1\omega_o$ (a conservative limit, for very critical applications) or even $0.5\omega_o$ (for noncritical applications). Such performance is sufficient in many cases. The degradation of the model is very gradual as the frequency is increased, so no “sharp” deterioration is observed at any particular frequency. When, at high frequencies, the model eventually becomes unacceptable, it will not be because the element values are not right, but rather because the *nature* of the model (Fig. 7.19) is inadequate for such frequencies. The only way to achieve satisfactory modeling in this case is to use more sophisticated model topologies, as discussed in Chap. 8.

The five capacitances previously defined are strongly dependent on the “bias values” of the terminal voltages, i.e., the values around which the small-signal voltages occur. These bias values are denoted by V_{D0} , V_{G0} , V_{B0} , and V_{S0} in Fig. 7.17. For simplicity, from now on we will denote them simply as V_D , V_G , V_B , and V_S . We now present expressions for the capacitances in terms of the bias voltages. We will consider each region of operation separately.

Strong Inversion To derive expressions for the capacitances in strong inversions, we will use the charges corresponding to the source-referenced simplified model in the preceding definitions. A good compromise between simplicity and accuracy can be obtained based on the following two assumptions. (1) The value of α in the model is given by (4.7.27), repeated here:

$$\alpha_I = 1 + \frac{\gamma}{2\sqrt{\phi_0 + V_{SB}}} = 1 + \frac{dV_T}{dV_{SB}} \quad (7.3.7)$$

[†] As the frequency is raised, both the magnitude and the phase as predicted by the model will be increasingly in error. For many applications, the phase error becomes intolerable first.

and (2) the derivative of α_1 with respect to V_S and V_B will be assumed to be negligible; i.e., in our differentiation of the charge expressions, the quantity α_1 will be treated as a *constant*.[†]

The gate-related capacitances, C_{gs} and C_{gd} , can be found by using the gate charge, (6.4.17), in the definitions (7.3.1) and (7.3.3). After much algebra, we obtain (Prob. 7.18)

$$C_{gs} = C_{\text{ox}} \frac{2(1+2\eta)}{3(1+\eta)^2} \quad (7.3.8)$$

and

$$C_{gd} = C_{\text{ox}} \frac{2(\eta^2 + 2\eta)}{3(1+\eta)^2} \quad (7.3.9)$$

where C_{ox} is the total oxide capacitance in the *intrinsic* part:

$$C_{\text{ox}} = WLC'_{\text{ox}} \quad (7.3.10)$$

and η is the parameter defined in (4.7.25) and plotted in Fig. 4.25. These results turn out to be accurate when compared with more complicated results from general models.

We now consider the body-related capacitances, C_{bs} and C_{bd} . We first define some quantities that will be needed for their evaluation. From the “back gate” point of view of the body, one would expect for these capacitances a qualitatively similar behavior as for the corresponding gate capacitances just discussed. Instead of C'_{ox} , the relevant quantity here is the body-channel capacitance per unit area, $C'_{bc}(V_{CB})$, where V_{CB} is the channel-to-body voltage. This capacitance can be found in a manner similar to the junction capacitance (1.5.24), only here we have a “field-induced junction,” with reverse bias V_{CB} and built-in potential ϕ_0 (see Sec. 3.4.2)

$$C'_{bc}(V_{CB}) = \frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\phi_0 + V_{CB}}} \quad (7.3.11)$$

Because the depletion region depth varies along the channel (Fig. 7.17), this capacitance varies along the channel, too. To determine the total channel-to-body capacitance, one needs to integrate the preceding capacitance per unit area over the

[†] These assumptions simplify the calculations, because every time a charge expression containing V_I is differentiated with respect to V_S or V_B it will produce a term containing α_1 , as can be seen from (7.3.7). This term can then be grouped together with other terms containing α_1 , resulting in simple expressions. This approach can be expected to be more accurate for devices with low body effect.

area of the channel. Although this can be done formally,[†] a simple approach will be adopted here. In analogy with (7.3.10), we can write

$$C_{bc} = WL C'_{bc}(V_{CB}^*) \quad (7.3.12)$$

where $C'_{bc}(V_{CB}^*)$, evaluated at an appropriate value of V_{CB}^* (discussed shortly), represents the average body-to-channel capacitance per unit area. The channel-body voltage, V_{CB} , varies from V_{SB} at the source to $V_{SB} + V_{DS}$ at the drain in nonsaturation, or to $V_{SB} + V'_{DS}$ in saturation. Thus, V_{CB}^* should lie between these two extremes and should increase as the voltage across the channel is increased. A simple empirical expression, which exhibits this behavior, turns out to be adequate for our purposes:

$$V_{CB}^* = V_{SB} + k_c(1 - \eta)V'_{DS} \quad (7.3.13)$$

where k_c is an empirical parameter, with a typical value of 0.1 to 0.2.

To determine C_{bs} and C_{bd} , one can use the body charge (6.4.15) in the definitions (7.3.2) and (7.3.4). Using again the simplifications mentioned previously, we obtain

$$C_{bs} = C_{bc} \frac{2(1+2\eta)}{3(1+\eta)^2} \quad (7.3.14)$$

$$C_{bd} = C_{bc} \frac{2(\eta^2+2\eta)}{3(1+\eta)^2} \quad (7.3.15)$$

where C_{bc} is given by (7.3.12) with (7.3.11), with $V_{CB}^* = V_{SB}$; this is the result of the simplifying assumptions we have made, and it can result in poor accuracy. Using (7.3.13) instead improves matters considerably, while keeping the results simple. If V_{GS} is small, V'_{DS} will be small, too, and the average channel-body voltage can be approximated by its value at the source: $V_{CB}^* \approx V_{SB}$. However, if V_{GS} is large, then with large V_{DS} , the channel-body voltage can vary considerably over the channel, and V_{CB}^* can be considerably larger than V_{SB} .

Comparing (7.3.14) and (7.3.15) with (7.3.8) and (7.3.9), we see that the body capacitances can be expressed in the form

$$C_{bs} = b_c C_{gs} \quad (7.3.16)$$

$$C_{bd} = b_c C_{gd} \quad (7.3.17)$$

[†] We have, for the total capacitance, $C_{bc} = \int_0^L C'_{bc} W dx = WL \bar{C}_{bc}$, where $\bar{C}_{bc} = \frac{1}{L} \int_0^L C'_{bc} dx$. This integral can be evaluated using (7.3.11) and a change of variables from (4.7.33). However, the approach suggested in the text is sufficient for our purposes.

where

$$b_c = \frac{C_{bc}}{C_{ox}} \quad (7.3.18)$$

is the ratio of the total body-channel capacitance to the total gate-channel capacitance. The form of (7.3.16) and (7.3.17) is reminiscent of the equation relating body and gate transconductances in (7.2.31).

Consider now the special case of a small V_{DS} , or a small V_{GS} (in the latter case the maximum voltage across the channel, V'_{DS} , will be small). From the preceding relations, along with (7.2.31) and (7.2.32) or (7.2.33), we get $b_c \approx b_g$. This leads to an interesting approximate relation:⁵⁹

$$\frac{C_{bs}}{C_{gs}} \approx \frac{C_{bd}}{C_{gd}} \approx \frac{C_{bc}}{C_{ox}} \approx \frac{g_{mb}}{g_m} \approx \frac{dV_T}{dV_{SB}} \approx \frac{\gamma}{2\sqrt{\phi_0 + V_{SB}}} = \alpha_1 - 1, \quad \text{small } V_{DS} \text{ or small } V_{GS} \quad (7.3.19)$$

This result will be made intuitively plausible shortly.

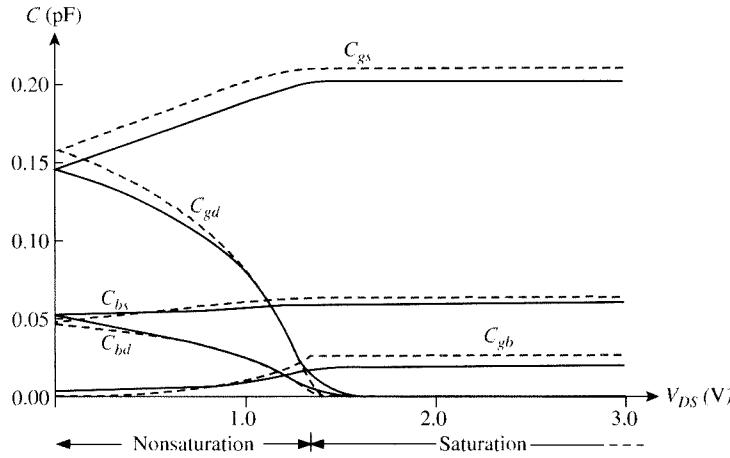
The remaining capacitance, C_{gb} , can be found by using (6.4.17) in the definition (7.3.5). This gives

$$C_{gb} = \frac{C_{ox} \times C_{bc}}{C_{ox} + C_{bc}} \frac{1}{3} \left(\frac{1-\eta}{1+\eta} \right)^2 \quad (7.3.20)$$

In this expression, the term on the right-hand side containing the capacitances corresponds to the series combination of the oxide and depletion region capacitances (per unit area). The rest of the expression keeps C_{gb} below this value, which makes sense because the channel acts as a shield between body and gate (at least partially).†

The five capacitances calculated previously are plotted vs. V_{DS} using broken lines in Fig. 7.20. The solid lines are for an accurate surface potential-based model (involving very complicated expressions). As seen, for the accurate model the transition from nonsaturation to saturation is gradual, as expected (notice, in particular, the behavior of C_{gd}). It is also seen that the strong-inversion approximations overestimate the gate-source and gate-drain capacitances to some extent. This is because the assumption behind strong-inversion expressions is that the inversion layer forms a perfect “plate” below the oxide. This is not exactly true; as we know from Sec. 4.3, if the gate voltage is raised, the surface potential increases somewhat, even if we are in strong inversion. Thus the increase in the oxide voltage is less than the increase in the gate voltage. This results in a smaller gate charge increase than what the strong-inversion approximation would predict, corresponding to a somewhat smaller capacitance. The same reason was responsible for the fact that the gate-body capacitance for the two-terminal structure, shown in Fig. 2.25, is somewhat less than the oxide capacitance, even in strong inversion. The accuracy of our simple expressions

† When using (7.3.13) in association with this equation, the value of k_c may be chosen different from that used for the calculation of C_b and C_{bd} (typically, a somewhat higher value may result in better accuracy).

**FIGURE 7.20**

Small-signal intrinsic capacitances vs. V_{DS} in strong inversion, for a large-gate device with a thick oxide. Dashed lines are for the approximate strong-inversion model discussed in the text; solid lines are for an accurate surface potential-based model. $t_{ox} = 11 \text{ nm}$, $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 10 \mu\text{m}$, $V_{GS} = 2.5 \text{ V}$, $V_{SB} = 0 \text{ V}$.

is seen to be good in all other respects. The qualitative behavior seen in Fig. 7.20 will be discussed intuitively shortly.

The previous expressions for the capacitances and the relations (7.3.19) are approximate, but turn out to be exact at $V_{DS} = 0$, assuming the usual strong-inversion simplifications (e.g., a surface potential “pinned” to a value ϕ_0) are valid; see also below. Expressions corresponding to the complete strong inversion model of Sec. 4.7.1 are discussed elsewhere.^{13.15}

We now consider the capacitances for two cases of interest.

Nonsaturation with $V_{DS} = 0$. With $V_{DS} = 0$ ($\eta = 1$), the previous capacitance expressions give

$$C_{gs} = C_{gd} = \frac{1}{2}C_{ox} \quad (7.3.21)$$

$$C_{bs} = C_{bd} = \frac{1}{2}C_{bc,0} \quad (7.3.22)$$

$$C_{gb} = 0 \quad (7.3.23)$$

where $C_{bc,0}$ denotes the value of C_{bc} at $V_{DS} = 0$, found from (7.3.12) and (7.3.13) with $\eta = 1$.

It is easy to make these results intuitively plausible. With $V_{DS} = 0$, the channel is as shown in Fig. 7.21. Practically all the gate field lines terminate on the inversion

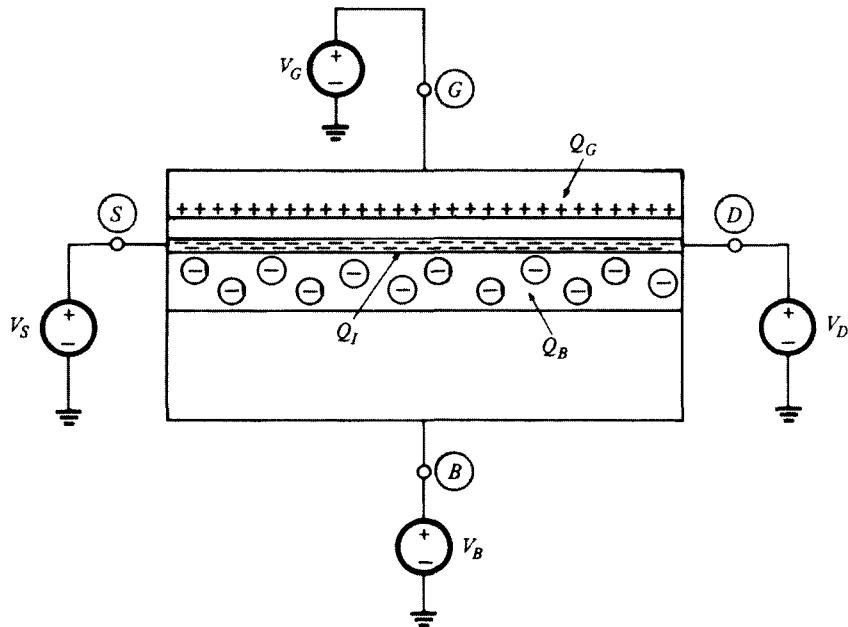


FIGURE 7.21
A MOS transistor with $V_D = V_S$.

layer, which is assumed strongly inverted throughout. If V_G and V_B are kept fixed, and both V_S and V_D are increased by the same amount ΔV , the potential across the oxide will decrease at every point by ΔV . Hence, the corresponding gate charge decrease will be $C_{ox}\Delta V$. If, instead, only V_S is increased by ΔV whereas V_D is kept fixed, the potential *change* across the oxide will vary from ΔV at the source end to zero at the drain end. This variation will be linear for a very small ΔV . It is easy to see that now the decrease in the gate charge will only be half as much as before, i.e., $\frac{1}{2}C_{ox}\Delta V$ (Prob. 7.19). Thus, to properly model this effect, $\frac{1}{2}C_{ox}$ must be the value of C_{gs} . A similar argument can be given for C_{gd} , thus verifying (7.3.21).

The values of C_{bs} and C_{bd} can be discussed in a similar manner. Here the arguments used previously apply with C_{ox} replaced by $C'_{bc}(V_{SB})WL$, which, for the case we are considering, is seen to be the total capacitance of the uniform reverse-biased “field-induced junction” formed by the inversion layer and the substrate in Fig. 7.21. Thus (7.3.22) also makes sense.

We will now attempt to make the relation to transconductances, shown in (7.3.19), intuitively plausible. Consider a nonzero, very small V_{DS} , so that we can have a small nonzero current I_{DS} (g_m and g_{mb} would vanish if V_{DS} were zero). As long as V_{DS} is very small, it will not upset appreciably the uniformity in the channel, and the picture will be practically as shown in Fig. 7.21. Now, keeping V_B , V_S , and V_D fixed, let V_G be changed by ΔV_G . This will cause an inversion layer charge change of magnitude $|\Delta Q_I| = C_{ox}\Delta V_G$ and a corresponding current change $\Delta I_{Dl} = g_m\Delta V_G$.

Repeating this experiment by using the back gate (the body) instead of the front gate, we obtain $|\Delta Q_{l2}| = C_{bc,0} \Delta V_B$ and $\Delta I_{D2} = g_{mb} \Delta V_B$. Because the inversion layer is uniform, the current is proportional to Q_l . Thus $\Delta I_{D2}/\Delta I_{D1} = |\Delta Q_{l2}/\Delta Q_{l1}|$. From these relations we obtain $g_{mb}/g_m = C_{bc,0}/C_{ox}$, which is in agreement with (7.3.19).

We now consider C_{gb} . At $V_{DS} = 0$, the strongly inverted electron layer, which is “connected” to the fixed voltages $V_S = V_D$, keeps the voltage across the oxide fixed at all points, even if V_B is varied. Hence the gate does not feel the variation, and $\Delta Q_G = 0$. This is modeled by [see (7.3.5)]

$$C_{gb} = 0, \quad V_{DS} = 0 \quad (7.3.24)$$

In other words, the strong-inversion layer acts as a “shield” throughout the channel, protecting the gate from the influence of the substrate.

Saturation. In the saturation region ($\eta = 0$), we have from (7.3.8) to (7.3.15) and (7.3.20)

$$C_{gs} = \frac{2}{3} C_{ox} \quad (7.3.25)$$

$$C_{gd} = 0 \quad (7.3.26)$$

$$C_{bs} = \frac{2}{3} C_{bc,sat} \quad (7.3.27)$$

$$C_{bd} = 0 \quad (7.3.28)$$

$$C_{gb} = \frac{1}{3} \frac{C_{ox} \times C_{bc,sat}}{C_{ox} + C_{bc,sat}} \quad (7.3.29)$$

where $C_{bc,sat}$ is the value of C_{bc} in saturation, found from (7.3.12) and (7.3.13) with $\eta = 0$.

The results for C_{gd} and C_{bd} can be easily seen to make sense as follows. In the saturation region, communication from the drain to the rest of the device is cut off owing to pinchoff (neglecting CLM). Hence, when V_D is varied, the intrinsic device is not affected and all charges in it remain the same. In particular, $\Delta Q_G = 0$ and $\Delta Q_B = 0$. This, from (7.3.3) and (7.3.4) implies that $C_{gd} = 0$ and $C_{bd} = 0$, which is in agreement with (7.3.26) and (7.3.28).

Consider now C_{gs} in (7.3.25). This value is accurate, as found by comparisons with accurate models and measurements (see below). The fact that C_{gs} is found to be less than C_{ox} often comes as a surprise, because one tends to think of the inversion layer as the lower of two “parallel plates” of a capacitor, the other plate being the

gate. If we really did have a parallel metal plate all under the oxide (neglecting the pinchoff region near the drain), and if it were attached to the source and cut off from the drain, then we would, in fact, obtain a capacitance equal to C_{ox} . This is because, if the source potential were changed by ΔV_S , we would have $|\Delta\psi_{ox}(x)| = |\Delta V_S|$ for any position x along the channel, with ψ_{ox} the oxide potential. However, this is *not* what happens in the transistor. Using the analysis in Sec. 4.7.3 and (2.3.1), one can show that we will have $|\Delta\psi_{ox}| = |\Delta V_S|$ *only* at the source end of the channel, and that $|\Delta\psi_{ox}(x)|$ will be smaller and smaller as one goes toward the drain (in fact, for the complete model of Sec. 4.7.1, one can show that, in saturation, ψ_{ox} at the *drain* end remains fixed at a value corresponding to the *gate-drain* threshold voltage, which is independent of V_S). Thus, points further away from the source feel the change in the source potential less and contribute less to the change in the gate charge. This is why C_{gs} is less than C_{ox} .[†]

Let us now consider C_{bs} in saturation. Notice that (7.3.27) is of the form of (7.3.25). One can obtain an intuitive feeling for this result for the case of large V_{SB} and small V'_{DS} . In such cases, the effective reverse bias of the channel with respect to the substrate does not vary much along the channel. Thus, the depletion region width is roughly uniform, just as was the case with $V_{DS} = 0$, and one can argue as we did previously for C_{gs} , with the role of C_{ox} played now by C_{bc} . Thus, the result in (7.3.27) is not surprising. If the preceding conditions on bias voltages are not satisfied, though, the previous intuition holds only approximately. If V_{GS} and, thus, V'_{DS} are large, the depletion region width will increase significantly as we go toward the drain, making the influence of the channel on the substrate weaker (as V_S is varied). For the same reasons, if V_{GS} is small, C_{bs} tends to increase with V_{DS} , whereas if V_{GS} is large, C_{bs} tends to decrease with V_{DS} .

Finally, we consider C_{gb} . Assume a change ΔV_B of V_B in Fig. 7.17d. Working as explained in the preceding paragraph, we find that here the resulting change in the oxide potential, $\Delta\psi_{ox}$, although zero at the source end, becomes nonzero at other points, and in fact $|\Delta\psi_{ox}(x)|$ increases toward the drain. (For example, for the complete model of Sec. 4.7.1 in saturation, ψ_{ox} at the drain is maintained at a value corresponding to the *gate-drain* threshold, as already mentioned; when V_B is changed, this value changes due to the body effect at that point.) The nonzero changes in $\psi_{ox}(x)$ contribute to a change in the gate charge. Thus C_{gb} is nonzero in saturation, as predicted by (7.3.29).

Intrinsic Transition Frequency Consider a transistor in the connection shown in Fig. 7.22a, where the bias is assumed such that operation is in the saturation region. The voltage $\epsilon \sin \omega t$ is a sinusoidal small signal of angular frequency ω (in rad/s). The small-signal equivalent circuit for this connection can be derived by substituting

[†] In circuits literature, the value $\frac{2}{3}C_{ox}$ in (7.3.25) is sometimes justified by saying that the length of the pinchoff region is $\frac{1}{3}L$, and thus the effective channel length is $\frac{2}{3}L$. This explanation is not correct. The factor $\frac{2}{3}$ is the result of a mathematical derivation that has nothing to do with the length of the pinchoff region. This length, in fact, must be assumed zero in order to arrive at (7.3.8) and, thus, at (7.3.25). Besides, the length of the pinchoff region is not $\frac{1}{3}L$ in general; in fact it is predicted from first-order theory to be independent of L from (5.3.3) or (5.3.5). For long-channel devices, the pinchoff region occupies a very small part of the channel.

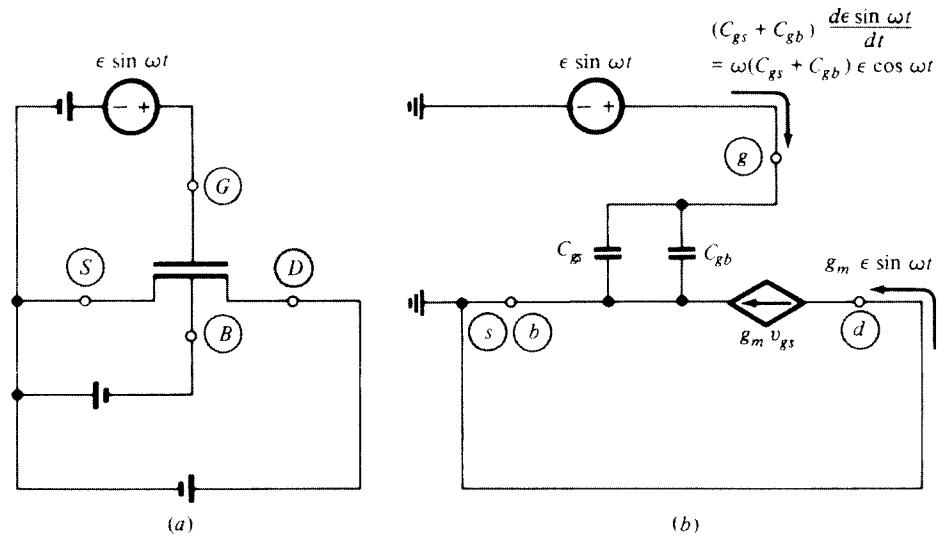


FIGURE 7.22

(a) A transistor operating in the saturation part of the I_{DS} - V_{DS} characteristics with a sinusoidal small-signal excitation; (b) the small-signal equivalent circuit for (a).

the transistor by the circuit of Fig. 7.19, and by substituting all dc voltage sources with short circuits (since for these sources $\Delta V = 0$). Removing now all elements that appear in parallel with short circuits and noting that C_{gd} is zero in saturation, we arrive at the circuit of Fig. 7.22b. The small-signal drain and gate currents can be calculated by using this circuit, and are shown directly on the figure. Defining the “short-circuit current gain” magnitude a_i as the ratio of the *amplitude* of the small-signal drain current to the *amplitude* of the small-signal gate current, we have

$$a_i = \frac{g_m}{\omega(C_{gs} + C_{gb})} \quad (7.3.30)$$

For $\omega \rightarrow 0$ this becomes infinite, which is to be expected since at dc there is no gate current.[†] As ω is increased, a_i drops. The *intrinsic transition frequency*, or *intrinsic cutoff frequency*, denoted by ω_{Ti} , is defined as the value of ω at which a_i drops to the value of 1. From the preceding equation, this value is

$$\boxed{\omega_{Ti} = \frac{g_m}{C_{gs} + C_{gb}}} \quad (7.3.31)$$

Assuming no velocity saturation, using (7.2.29a) and (7.3.25), and neglecting C_{gb} , we get

[†] Recall that we are assuming no gate leakage. This assumption is not valid for very thin oxides (Sec. 5.12).

$$\omega_{Ti} \approx \frac{g_m}{C_{ox}} = \frac{3}{2} \frac{\mu(V_{GS} - V_T)}{\alpha L^2} = \frac{3}{2} \omega_o, \quad \text{strong inversion} \quad (7.3.32)$$

where ω_o is defined in (7.3.6). The value of ω_{Ti} is outside the region of validity of the model used, so it should initially be looked at with suspicion. However, calculation of ω_{Ti} using higher-frequency models (Chap. 8) gives essentially the same value. The reason for this is that the preceding calculation involves only the *magnitudes* of the currents. For these, errors using the simple model are not large at these frequencies. However, significant *phase* errors can result from using the models derived in this chapter at such high frequencies.

As an example, consider a transistor with $\mu = 400 \text{ cm}^2/(\text{V}\cdot\text{s})$, $L = 200 \text{ nm}$, and $\alpha = 1.1$, operating in saturation with $V_{GS} - V_T = 0.5 \text{ V}$. From (7.3.32), we have $\omega_{Ti} = 682 \text{ Grad/s}$. This corresponds to $f_{Ti} = \omega_{Ti}/2\pi = 109 \text{ GHz}$.

In practice, the current gain of a complete transistor becomes unity at some frequency less than ω_{Ti} because of the presence of extrinsic parasitic elements (Sec. 7.4). Velocity saturation can further limit the cutoff frequency; these effects are discussed in Sec. 8.6.

Moderate Inversion In Sec. 6.4.3, it was mentioned that in circuits texts the moderate-inversion region is sometimes neglected. The approach taken in such cases is to consider weak- and strong-inversion expressions as valid in adjacent regions. The resulting error is often acceptable for charge evaluation. However, such charge expressions will result in large error in predicting *capacitances*. This will be seen later in this section when we compare this approach to results derived from complex general models valid in all regions of inversion.

As was true for small-signal conductances, to evaluate the small-signal capacitances in moderate inversion, one resorts to general models, which are valid in all regions of inversion (see below). Interpolation modes are also used; an example will be shown in Table 7.1 of the end of this chapter.

Weak Inversion In weak inversion, things are very simple. The inversion layer charge is negligible throughout the length of the channel, and the gate “sees” the depletion region directly through the oxide. A small increase in V_B will cause some charges to enter through the substrate terminal, and this will be balanced by some gate charge leaving through the gate terminal. The corresponding value of C_{gb} can be found by using the weak inversion Q_G from (6.4.34) and (6.4.33) in (7.3.5):

$$C_{gb} = C_{ox} \frac{\gamma}{2\sqrt{\gamma^2/4 + V_{GB} - V_{FB}}} \quad (7.3.33)$$

Varying V_S or V_D in weak inversion can vary the inversion layer charge drastically in a relative sense. However, this charge remains negligible compared with Q_G and Q_B and can play no significant role in the charge-balancing process. Hence the gate and substrate do not “feel” the variations of V_S and V_D , and the corresponding

ΔQ_G and ΔQ_B in Fig. 7.17b and c are negligible. The values for the corresponding capacitances are then[†]

$$C_{gd} \approx C_{gs} \approx C_{bd} \approx C_{bs} \approx 0 \quad (7.3.34)$$

An intrinsic cutoff frequency can be defined for weak inversion (with $V_{DS} > 5\phi_i$) in a similar manner as in strong inversion; see Fig. 7.22. Using (7.3.34) in (7.3.31), we get

$$\omega_{Ti} \approx \frac{g_m}{C_{gb}} \quad (7.3.35)$$

Using the weak-inversion expressions for g_m and C_{gb} , it can be shown that (Prob. 7.24)

$$\omega_{Ti} \approx \frac{\mu\phi_i}{L^2} \frac{I_{DS}}{I_M}, \quad \text{weak inversion} \quad (7.3.36)$$

where I_{DS} is the current at the particular operating point and I_M is the maximum achievable current in weak inversion (i.e., the current at the upper limit of the weak-inversion region).

As an example, consider a device with $\mu = 400 \text{ cm}^2/(\text{V}\cdot\text{s})$ and $L = 200 \text{ nm}$, operated at room temperature with I_{DS} equal to one-fifth the maximum weak-inversion current. Then (7.3.36) gives $\omega_{Ti} = 5.2 \text{ Grad/s}$. This corresponds to $f_{Ti} = \omega_{Ti}/2\pi = 825 \text{ MHz}$. The same device was seen earlier to have a f_T of 109 GHz in strong inversion, with $V_{GS} - V_T = 0.5 \text{ V}$.

All-Region Models In Sec. 6.4.5, we have shown how the all-region model of Sec. 4.3 can be used to evaluate the various charges. The resulting expressions are valid in all regions of inversion. These charges can be differentiated to produce capacitances (not an easy task). The resulting expressions are too lengthy,⁵² unless an approximation is used for the bulk charge, as in Sec. 4.4.3.⁵¹⁻⁵³ One can also develop capacitance expressions for the body-referenced simplified general model of Sec. 4.4.2. These are given on this book's website.⁹⁹

To show the capacitance variation over all regions of inversion, we can fix V_{DS} and plot vs. V_{GS} . This was the approach taken in the plots of Figs. 6.9 and 7.15. The resulting capacitance plots are shown in Fig. 7.23 and agree very well with experiment in all regions. In the moderate-inversion region, both strong- and weak-inversion expressions fail completely. Yet, in some circuits texts, weak- and strong-inversion expressions are assumed to be valid in adjacent regions.

[†] This discussion only considers the effect of the source and drain through the inversion layer charge. In addition to this, one has the proximity capacitances between the gate *inside* the broken line in Fig. 6.1 and the *inside* side walls of the n^+ regions.⁶⁷ These capacitances are observed in weak inversion and depletion, but reduce to zero in strong inversion, since the inversion layer then acts as a shield between the gate and the *inside* side walls of the n^+ regions. See “Small-Dimension Effects” below.

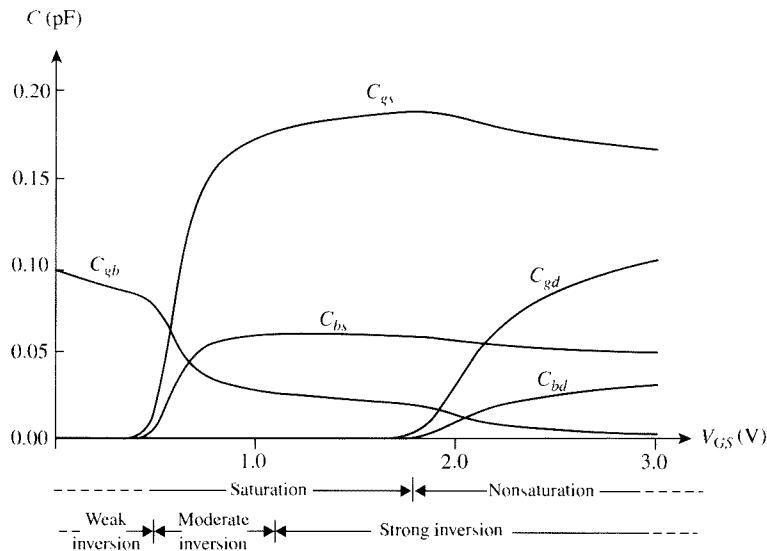


FIGURE 7.23
Small-signal intrinsic capacitances vs. V_{GS} from an accurate all-region model, for a large-gate device with a thick oxide. $t_{ox} = 11$ nm, $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 10 \mu\text{m}$, $V_{DS} = 0.9$ V, $V_{SB} = 0$ V.

In Fig. 7.24, we compare g_{mb}/g_m , C_{bs}/C_{gs} , and C_{bd}/C_{gd} . It is seen that (7.3.19) is approximately verified, if mobility is assumed constant.

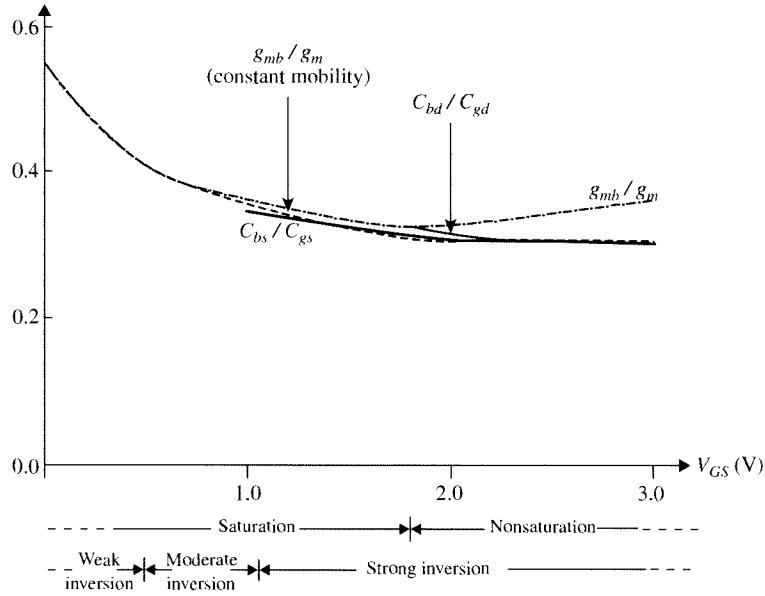
Both simplified all-region models^{51–53} and the EKV model^{54–57} result in capacitance expressions that are valid in all regions of inversion. An example is shown at the end of this chapter.

Depletion and Accumulation In depletion and accumulation, the inversion layer charge is totally negligible, and thus the intrinsic C_{gs} , C_{gd} , C_{bs} , and C_{bd} become zero. The only nonzero capacitance is C_{gb} .[†] This capacitance is important, as it can be varied with bias, and is used in voltage-controlled capacitors (varactors) in RF circuits.^{100–103} Using (6.4.49) in the definition of this capacitance, (7.3.5), we obtain

$$C_{gb} = WLC'_{ox} \left(1 + \frac{\partial \psi_s}{\partial V_B} \right) \quad (7.3.37)$$

The derivative $\partial \psi_s / \partial V_B$ is negative, as can be seen from Fig. 2.12 and the fact that $V_{GB} = V_G - V_B$. Thus C_{gb} is less than the oxide capacitance $C_{ox} = WLC'_{ox}$, which is consistent with Fig. 2.25. Evaluation of this capacitance requires the use of (2.4.25). Simple results are possible in deep depletion and deep accumulation, as is now discussed.

[†] See, however, the discussion of the inner fringing capacitance illustrated in Fig. 6.17.

**FIGURE 7.24**

Comparison of intrinsic capacitance ratios to transconductance ratios for a large-gate device with a thick oxide. Broken line shows result assuming mobility is constant. $t_{\text{ox}} = 11 \text{ nm}$, $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 10 \mu\text{m}$, $V_{DS} = 0.9 \text{ V}$, $V_{SB} = 0 \text{ V}$.

Deep in depletion. Using (6.4.52) with (6.4.51) in the definition (7.3.5), we obtain

$$C_{gb} = WLC'_{\text{ox}} \frac{\gamma}{2\sqrt{\gamma^2/4 + V_{GB} - V_{FB}}} \quad (7.3.38)$$

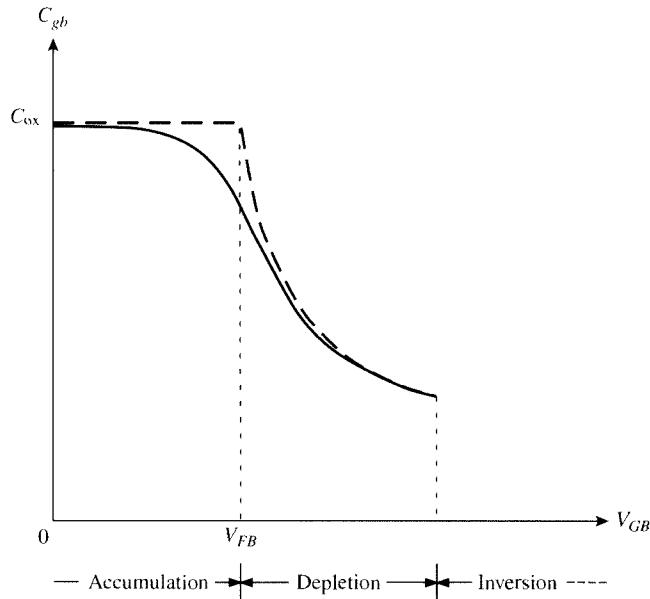
Deep in accumulation. Deep in accumulation, ψ_s varies little with V_{GB} , as can be seen in Fig. 2.12. Neglecting the derivative in (7.3.33), we obtain

$$C_{gb} \approx WLC'_{\text{ox}} \quad (7.3.39)$$

This result is consistent with the picture that, deep in accumulation, the abundance of holes creates a conductive sheet right under the oxide.

Figure 7.25 shows C_{gb} in the regions of accumulation and depletion. The broken line represents (7.3.38) and (7.3.39), and the solid line represents (7.3.37) with (2.4.25).

Small-Dimension Effects The evaluation of intrinsic capacitances of short-channel devices must take into account several effects that we could conveniently neglect in the case of long-channel devices. Such effects include velocity saturation, CLM, two-dimensional charge-sharing and DIBL effects. Some of these effects can approximately be taken into account by using “effective” values for such parameters as the threshold voltage and the gate width in the long-channel expressions, as was done in modeling the current in Chap. 5. To do a better job, the charges must be evaluated

**FIGURE 7.25**

Gate-substrate capacitance in accumulation and depletion plotted vs. V_{GB} . Broken line: (7.3.38) and (7.3.39); solid line: (7.3.37) with (2.4.25).

including such effects, and the resulting expressions differentiated with respect to the terminal voltages, as we have done previously. It should be emphasized that our definitions in (7.3.1) to (7.3.5) are still valid. However, the results are invariably complicated; the interested reader is referred to the literature.^{43,65-77,79,81,82,84,85,89,91,95}

For small-dimension devices, one typically obtains capacitance-voltage plots that are qualitatively similar to those of long-channel devices, but the details of the plots, and the actual values, can be significantly different.⁷²⁻⁷⁷ An example is shown in Fig. 7.26a, which should be compared with the long-channel case in Fig. 7.23. It is seen that for the short-channel device, it is not easy to identify particular regions of operation on the plot, because the transition from region to region is very gradual (this trend is also apparent in the current of short-channel devices; for example, in saturation, V_{DS} continues to affect the current through DIBL and CLM, and makes the transition from nonsaturation to saturation less distinct than for long-channel devices).

A strong inversion plot is shown in Fig. 7.26b, which can be compared with the long-channel case in Fig. 7.20. Again, the transition from nonsaturation to saturation is seen to be less distinct.

The thin oxide effects described in Secs. 5.11 and 5.12 can also affect the device capacitances.⁹⁷ For example, depending on device type, oxide thickness, polysilicon gate doping, and voltage polarity, a depletion region can form in the polysilicon gate immediately above the oxide, and the corresponding capacitance will be in series with the oxide capacitance. This can decrease the gate capacitance by over 10 percent.⁹⁷

The study of small-dimension effects can be greatly aided by two-dimensional computer simulations,⁷²⁻⁷⁵ correlated with measurements. The latter can be very

434 OPERATION AND MODELING OF THE MOS TRANSISTOR

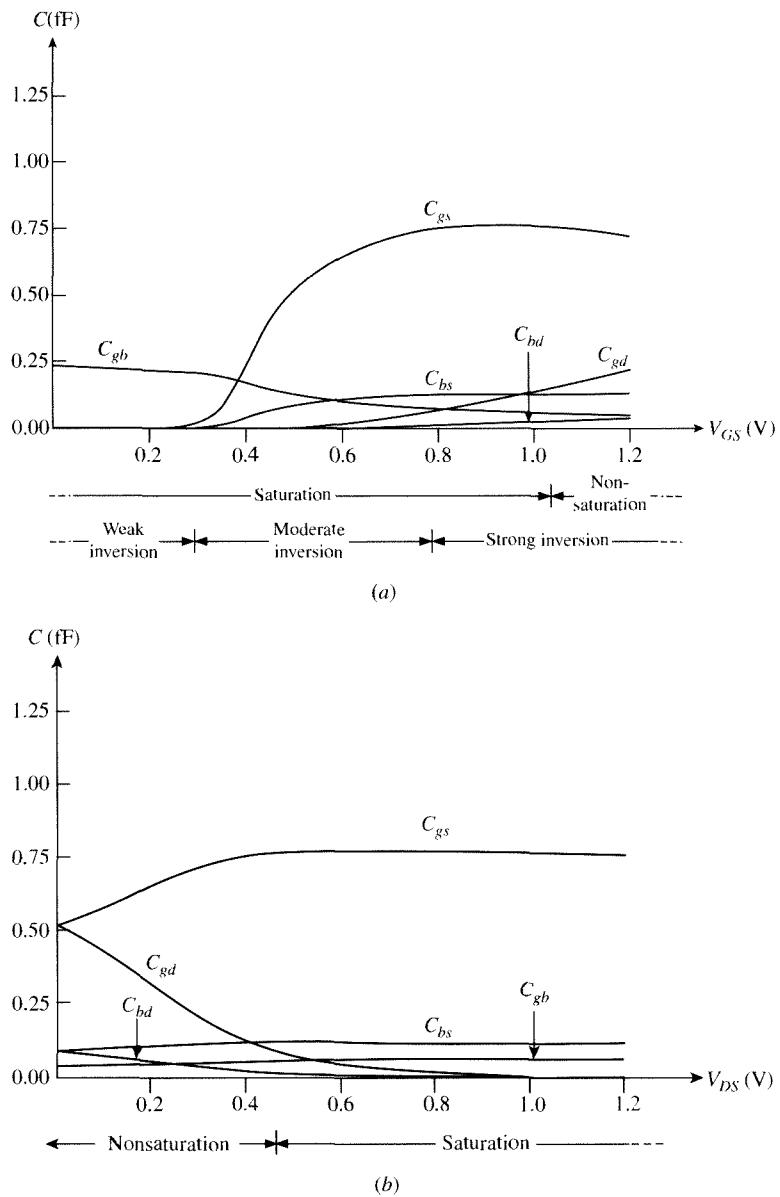


FIGURE 7.26

Small-signal *intrinsic* capacitances from an accurate all-region model, for a short-channel device with thin oxide. (a) vs. V_{GS} ; $V_{DS} = 0.4$ V, $V_{SB} = 0$. (b) vs. V_{DS} in strong inversion. $t_{ox} = 1.2$ nm, $N_A = 1 \times 10^{18}$ cm $^{-3}$, $W = 1$ μ m, $L = 0.065$ μ m, $V_{GS} = 1$ V, $V_{SB} = 0$ V.

difficult since the intrinsic capacitances of small devices are small, and their effect can be masked by the drain current and by parasitic capacitances due to extrinsic effects, packaging, and measuring circuitry. The measurement of MOS transistor capacitances has been the subject of several studies.^{61,66,68,69,72,75,77,93,94}

7.4 Including the Extrinsic Part

Our discussion so far in this chapter has dealt with intrinsic behavior only, i.e., the behavior of the channel and the gate and body immediately above and below it. As discussed in Sec. 6.8, the rest of the transistor exhibits extrinsic parasitic resistances and capacitances; these have been shown in Fig. 6.16. The discussion of capacitances given in Sec. 6.8 is directly applicable to small-signal modeling, as these were incremental (i.e., small-signal) capacitances. Assuming the resistances discussed there are approximately linear, the values given for them can also be used for small-signal purposes.[†]

Adopting the simplification discussed in relation to Fig. 6.20, we can add six extrinsic capacitances and four series resistances to an intrinsic small-signal model, to obtain the augmented model shown in Fig. 7.27. A seventh capacitance, $C_{bb'}$ (shown as broken line), needs to be included if the device body is formed on a well of the opposite doping type than the substrate, to model the junction capacitance between

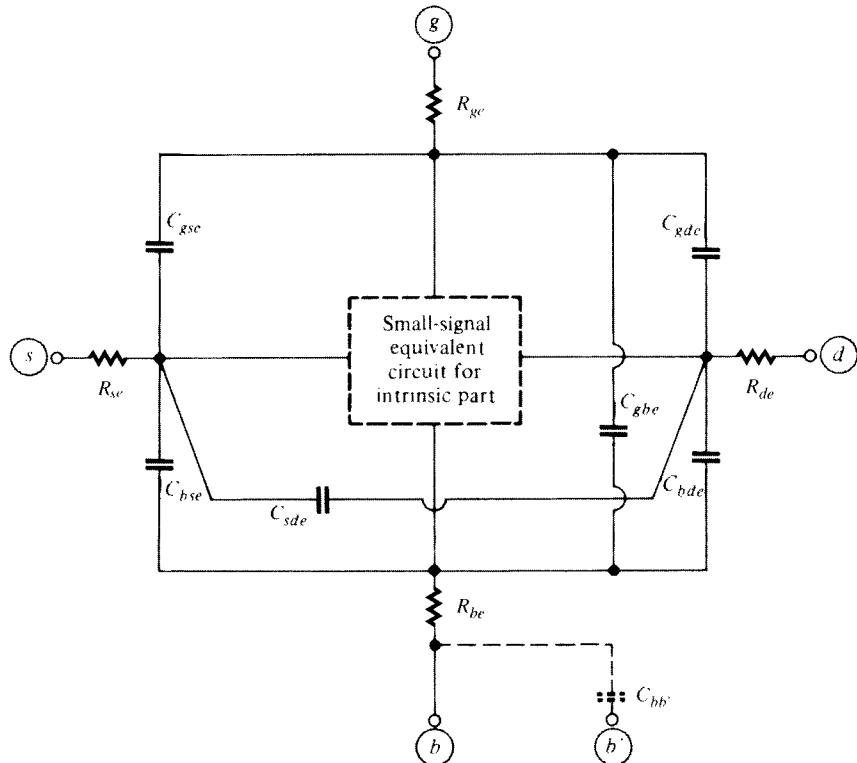


FIGURE 7.27
Extrinsic transistor capacitances added to an intrinsic small-signal model.

[†] A more accurate development would involve differentiating the total voltage–total current relations for the resistances, to develop the corresponding small-signal quantities. This is actually done numerically with models in circuit simulators.

the two. The resistances R_{se} , R_{ge} , R_{de} , and R_{be} are assumed to be approximately equal to the corresponding quantities in Fig. 6.20. The model for the intrinsic part, shown as broken lines, can be the model of Fig. 7.19, possibly augmented with the small-signal conductances associated with gate and body currents, as in Fig. 7.6.

Assuming that the model inside the broken line in Fig. 7.27 is the model in Fig. 7.19, it can be seen that the extrinsic capacitances (with the exception of C_{dse}) appear in parallel with the corresponding intrinsic capacitances; Thus we have, for example, a total gate-source capacitance of $C_{gs} + C_{gse}$, where C_{gs} is the intrinsic component and C_{gse} the extrinsic component. Total (intrinsic plus extrinsic) capacitance plots are shown in Fig. 7.28, for the same device parameters as those used for Fig. 7.26. Because this is a short-channel device, the effect of extrinsic capacitances is strong, as can be seen by comparing these plots to those in Fig. 7.26.

The performance of the transistor is in general decreased by the presence of the extrinsic parasitics. We will return to this topic in Chap. 8.

7.5 Noise

7.5.1 Introduction

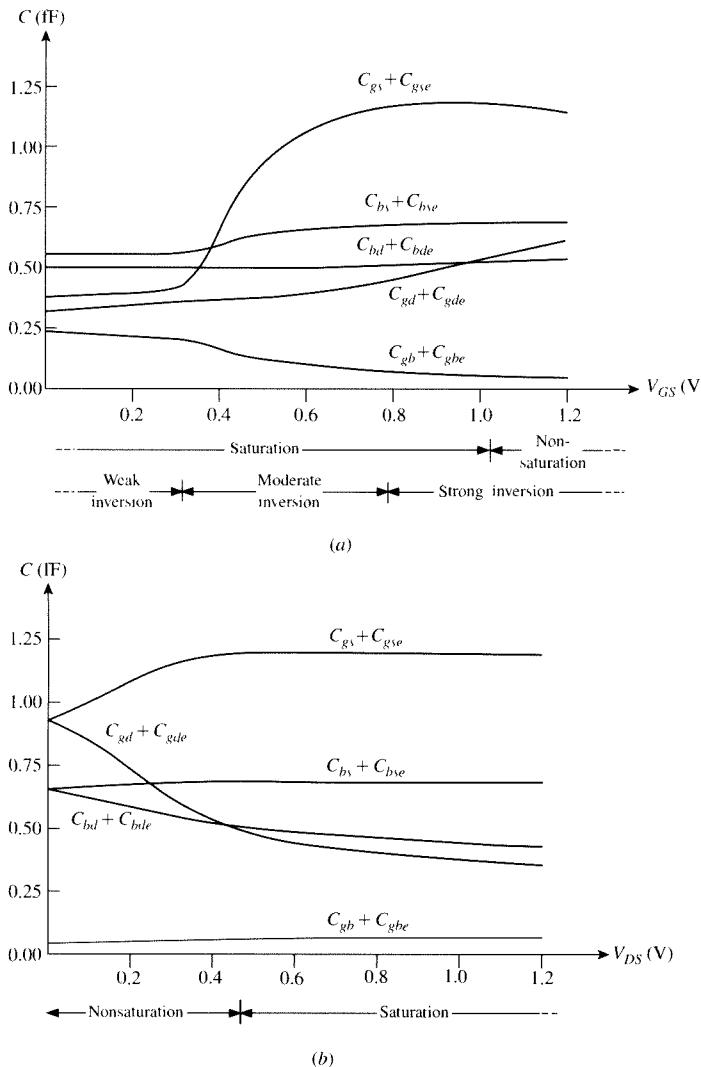
It has so far been assumed that the drain current of a MOS transistor varies with time only if one or more of the terminal voltages vary with time. This is not exactly true. A careful examination of the current reveals minute fluctuations, referred to as *noise*, which are present whether externally applied signals are present or not. Such fluctuations can occur due to several mechanisms (see the following). Noise can interfere with weak signals when the transistor is part of an analog circuit,^{104–106} so ways to predict and possibly reduce noise are very important. For this reason, the subject of noise in MOS transistors has received extensive treatment in the literature. This section is devoted to this subject. We have chosen to treat noise as part of our discussion of small-signal modeling, because noise is, in a sense, an internally generated small signal in the device, and it can be modeled with appropriate additions to the small-signal equivalent circuits we have already developed in this chapter.

Consider a transistor with dc bias voltages, as shown in Fig. 7.29a. The total drain current, shown in Fig. 7.29b, can be expressed as

$$i_{DS}(t) = I_{DS} + i_n(t) \quad (7.5.1)$$

where I_{DS} is the ideal (bias) current and $i_n(t)$ is the noise component, which has zero average value. The instantaneous value of i_n at a given t is unpredictable. Instead, one talks about certain measures characterizing the behavior of $i_n(t)$. In noise work, such measures are the *mean square* value, denoted by $\bar{i_n^2}$, and the *root mean square* (rms) value, $\sqrt{\bar{i_n^2}}$.

In measuring noise quantities, the amount of noise depends on the bandwidth of the measuring instrument. A common measurement involves a very narrow bandwidth, centered on a frequency f . The current noise spectral components within this bandwidth have a certain mean square value. The *ratio* of this value to the bandwidth, as the latter is allowed to approach zero, tends to what is called the *power*

**FIGURE 7.28**

Small-signal *total* (intrinsic plus extrinsic) capacitances from an accurate all-region model, for a short-channel device with a thin oxide. (a) vs. V_{GS} $V_{DS} = 0.4$ V, $V_{SB} = 0$ V; (b) vs. V_{DS} in strong inversion. $t_{ox} = 1.2$ nm, $N_A = 1 \times 10^{18}$ cm $^{-3}$, $W = 1$ μm , $L = 0.065$ μm , $V_{GS} = 1$ V, $V_{SB} = 0$ V. These parameters are the same as for Fig. 7.26.

spectral density[†] of the current noise, denoted by $S_i(f)$. This quantity has units of square amperes per hertz. Often the square root of the power spectral density is used instead, given in A/ $\sqrt{\text{Hz}}$. For a noise voltage v_n one can similarly define a power spectral density $S_v(f)$ (in square volts per hertz) or its square root (in V/ $\sqrt{\text{Hz}}$).

[†] More rigorous definitions can be found in specialized texts.^{104,105}

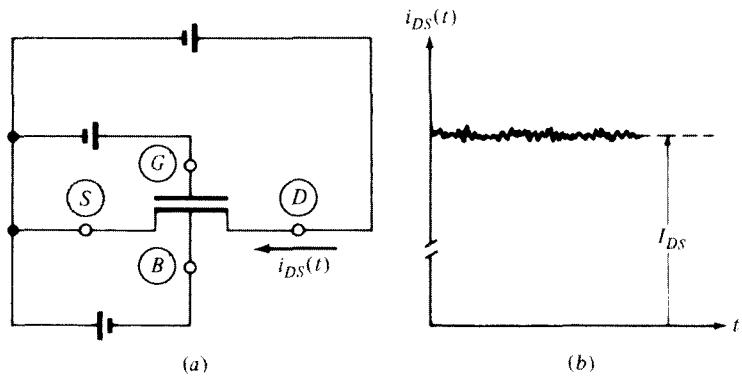


FIGURE 7.29 (a) A MOS transistor biased with fixed noiseless terminal voltages; (b) the drain-to-source current for the connection in (a), including noise.

The total mean square noise current within an arbitrary bandwidth extending from $f = f_1$ to $f = f_2$ can be found by summing the mean square values of the individual components within each subband width Δf . More precisely, using the power spectral density concept, we have

$$\overline{i_n^2} = \int_{f_1}^{f_2} S_i(f) df \quad (7.5.2)$$

A similar result can be obtained for voltage noise.

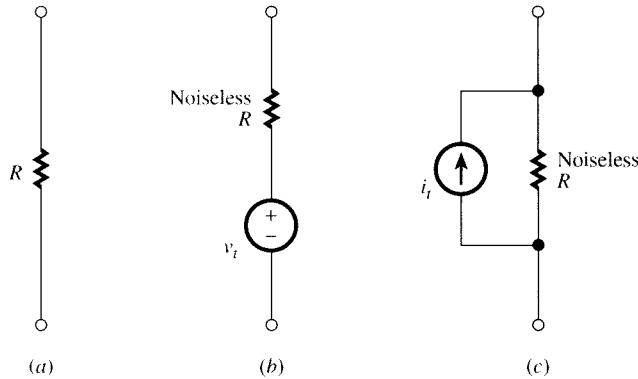
A well-known example of device noise is the *thermal noise* in a resistor (also called *Johnson noise* or *Nyquist noise*). It is due to the random motion of carriers, resulting from scattering events with crystal atoms, impurity atoms, and other carriers in thermal motion.^{104,105} A real (noisy) resistor (Fig. 7.30a) can be represented as an ideal (noiseless) resistor R in a series with a noise voltage source as shown in Fig. 7.30b. Using concepts from statistical physics, it can be shown that the corresponding power spectral density is^{104,105}

$$S_{\text{int}} = 4kT R \quad (7.5.3a)$$

where k is Boltzmann's constant (1.38×10^{-23} V·C/K) and T is the absolute temperature. Thermal noise is said to be *white* noise, because its power spectral density, as given by (7.5.3a), is "flat" up to extremely high frequencies (over 10^{12} Hz).

The circuit in Fig. 7.30b is a “Thévenin equivalent circuit.”¹⁰⁷ This can be converted to a “Norton equivalent circuit,”¹⁰⁷ shown in Fig. 7.30c, with a noise current $i_t = v_t/R$, or equivalently with $i_t^2 = v_t^2/R^2$; thus, the power spectral density of this noise current is

$$S_{ii} = 4kT \frac{1}{R} \quad (7.5.3b)$$

**FIGURE 7.30**

(a) A real (noisy) resistor; (b) Thévenin noise equivalent circuit for (a); (c) Norton noise equivalent circuit for (a).

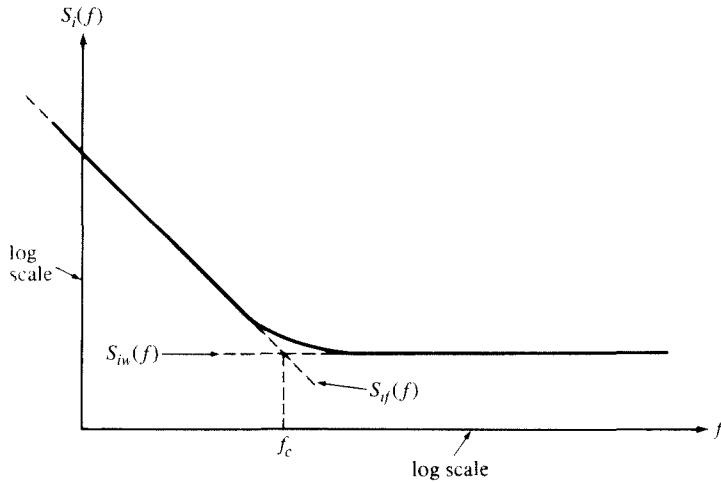
A second type of noise is *shot noise*. This noise is associated with dc flow produced by carriers crossing a potential barrier (such as the one from source to channel) and is due to the discreteness of arriving charges. As discussed in Sec. 5.10, carriers can cross the barrier provided they have sufficient energy to do so; some carriers cross, and some do not. The large numbers of carriers that do cross, each carrying a charge of magnitude q , constitute a current flow whose average is I , the dc value of the current. On top of this is superimposed a minute fluctuation due to the randomness in the flow of individual charge carriers over the barrier, constituting shot noise. This can be shown to have a power spectral density of^{105,108}

$$S_{is} = 2qI \quad (7.5.4)$$

Shot noise can be related to thermal noise in certain cases.^{104,109}

A typical plot of power spectral density for the drain current noise of a MOS device is shown in Fig. 7.31 on log-log axes. Two distinct frequency regions, with different noise behavior in each, can be identified. These regions can be thought of as separated by a *corner frequency* f_c . Values from several hertz to several megahertz are common for this quantity depending on device construction, geometry, and bias. The type of noise dominating at high frequencies in Fig. 7.31 is *white noise* (its power spectral density is constant up to extremely high frequencies) which is, discussed in Sec. 7.5.2. The corresponding noise current component will be denoted by i_w . The noise dominating at low frequencies is called *flicker noise* or $1/f$ noise; the power spectral density for the current caused by this type of noise is nearly proportional to $1/f$ (Sec. 7.5.3). The current component due to flicker noise will be denoted by i_f .

The noise currents i_w and i_f are independent. In calculating the total noise mean square value due to both, one can consider the effect of each separately and then *add the individual mean square values*. To see why, consider a noise current $i_n(t)$ that consists of the two independent contributions $i_w(t)$ and $i_f(t)$. We have $i_n(t) = i_w(t) + i_f(t)$, which gives $i_n^2(t) = i_w^2(t) + i_f^2(t) + 2i_w(t)i_f(t)$. The average of $i_n^2(t)$ is

**FIGURE 7.31**

A typical plot of the drain-noise current power spectral density vs. frequency in log-log axes. Subscript i refers to total noise, iw to white noise, and if to flicker noise.

$\bar{i_n^2}(t) = \bar{i_w^2(t)} + \bar{i_f^2(t)} + 2\bar{i_w(t)i_f(t)}$, where bars denote averages. However, if two contributions are independent, and at least one of them has zero mean, their product averages to zero^{104,105} and thus

$$\bar{i_n^2} = \bar{i_w^2} + \bar{i_f^2} \quad (7.5.5)$$

Now let $\bar{i_n^2}$, $\bar{i_w^2}$, and $\bar{i_f^2}$ be the mean square values of the corresponding noise spectral components within a very small bandwidth, and divide both sides of (7.5.5) by that bandwidth; allowing the bandwidth to approach zero, and recalling the concept of power spectral density discussed previously, we see that $S_i(f)$, the power spectral density of the total noise, will be given by

$$S_i(f) = S_{iw}(f) + S_{if}(f) \quad (7.5.6)$$

where $S_{iw}(f)$ and $S_{if}(f)$ are the power spectral densities of the white and flicker noise components, respectively, indicated in Fig. 7.31.

We now consider MOSFET white noise and flicker noise separately.^{104,105,108–271} Unless stated otherwise, we will assume that the transistors considered are long-channel ones. We will later consider what happens if the channel is short.

7.5.2 White Noise

In this section, we consider the white noise in the intrinsic transistor channel. Although this can be done in a general way, we will initially assume operation in strong inversion, in order to be able to draw some concrete pictures and convey the

fundamentals involved in as simple a manner as possible. We will then give a generalization of the results obtained. In strong inversion, the white noise is actually thermal noise, as discussed for resistors (most of Refs. 104, 105, and 110–162 deal with thermal noise in MOS transistors). Before embarking on an evaluation of this noise, we give here two conventional relations derived for the strongly inverted MOS transistor in Sec. 4.7.1. The drain current (assumed noiseless) was shown there to be, in the absence of velocity saturation,

$$I_{DS} = -\mu W Q'_l(V_{CB}(x)) \frac{dV_{CB}(x)}{dx} \quad (7.5.7)$$

where x is the position along the channel, $V_{CB}(x)$ is the “effective reverse bias” of the strongly inverted channel with respect to the substrate at position x (see Fig. 7.32), Q'_l is the inversion layer charge per unit area, μ is the mobility, and W is the width of the channel. Integrating this equation was seen to give, assuming a constant mobility,

$$I_{DS} = -\frac{W}{L} \mu \int_{V_{SB}}^{V_{DB}} Q'_l(V_{CB}(x)) dV_{CB}(x) \quad (7.5.8)$$

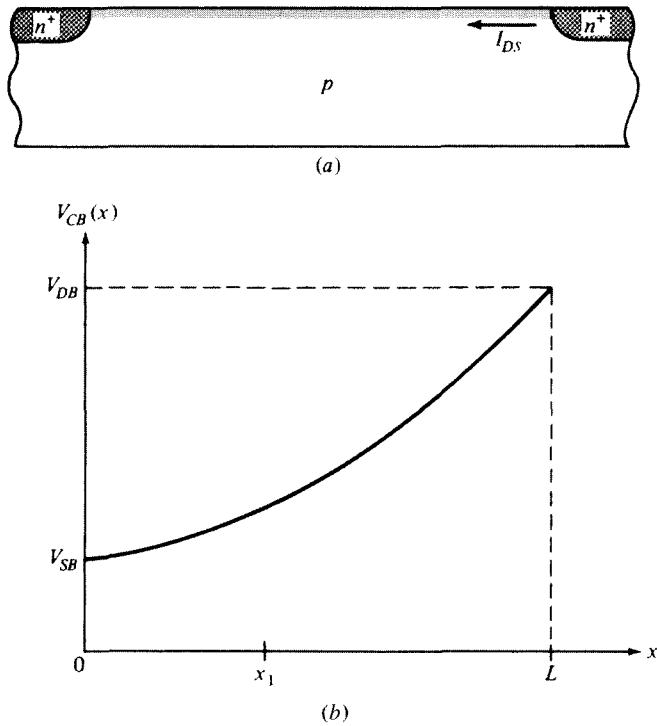


FIGURE 7.32
(a) Part of a MOS transistor in strong inversion; (b) effective reverse bias V_{CB} vs. distance along the channel.

with L being the channel length. Depending on the expression for $Q'_I(V_{CB})$, we have seen that models of differing complexity and accuracy are obtained.

In the following development, we also need an expression for the resistance ΔR of a small element of the channel of length Δx centered around a point $x = x_1$. Since $\Delta V_{CB} = I_{DS} \Delta R$ we have, from (7.5.7),

$$\Delta R = \frac{\Delta x}{-\mu W Q'_I(V_{CB}(x_1))} \quad (7.5.9)$$

where the right-hand side is positive, since $Q'_I < 0$.

Assuming that the small element of the channel acts as a resistor of resistance ΔR , we will observe across it a small noise voltage Δv_t with a power spectral density $4kT \Delta R$ [see (7.5.3)]. Thus, the mean square value of Δv_t over a bandwidth B will be[†]

$$\overline{(\Delta v_t)^2} = 4kT \Delta R \cdot B \quad (7.5.10)$$

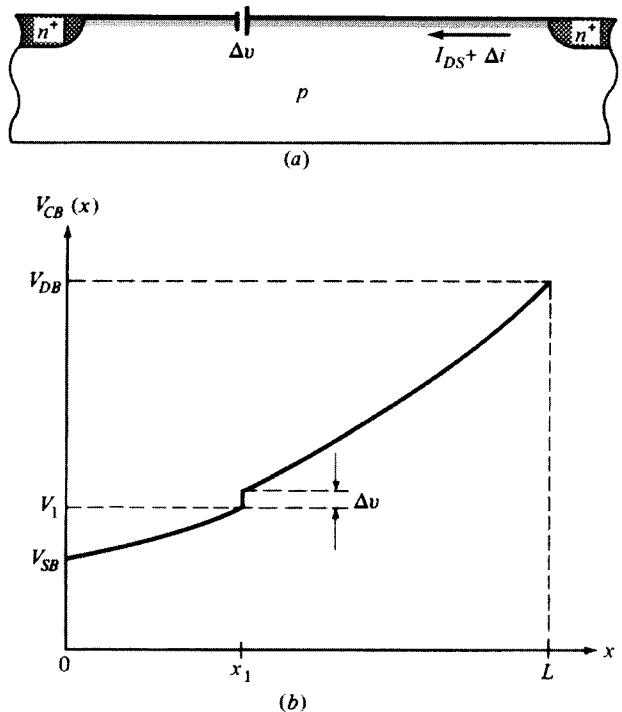
from which, using (7.5.9), we get

$$\overline{(\Delta v_t)^2} = \frac{4kT \Delta x}{-\mu W Q'_I(V_{CB}(x))} B \quad (7.5.11)$$

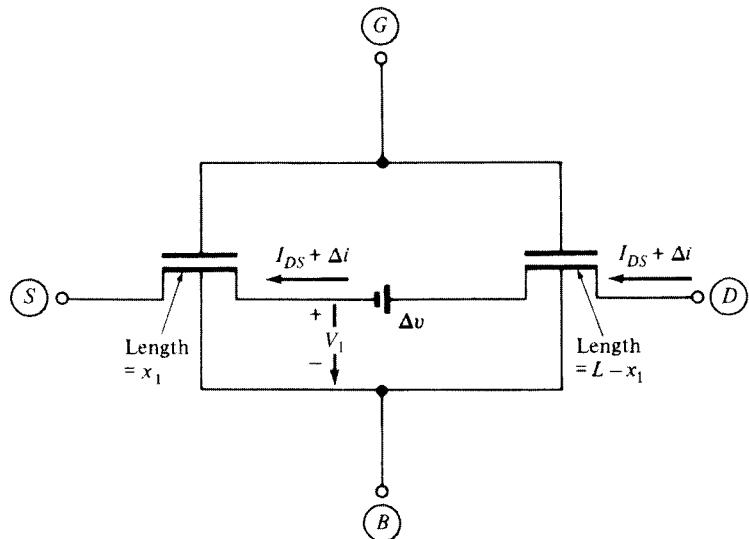
This noise will cause noise in the drain current. To study the mechanism by which this happens, we consider the following “thought” experiment. A fictitious dc voltage source of negligible length and of very small magnitude Δv is inserted at point x_1 in the channel, as shown in Fig. 7.33a. This will create a jump Δv in the potential $V_{CB}(x)$, as shown in Fig. 7.33b,¹⁰⁵ and will cause a change in the drain current, in comparison to that in Fig. 7.32. To get a feeling for this effect, one can consider Fig. 7.34.¹³ Here we have two transistors of lengths x_1 and $L - x_1$. The dimensions of the source-drain regions connected to Δv are assumed to have shrunk to zero. If $\Delta v = 0$, the connection of the two transistors is equivalent to the single transistor in Fig. 7.32, and the current is I_{DS} . If $\Delta v \neq 0$, the drain bias of the transistor on the left and the source bias of the transistor on the right will be disturbed. A new value for the current will then be established, along with a new potential distribution. These will correspond to the situation shown in Fig. 7.34. Let the new drain current value be $I_{DS} + \Delta i$, as noted in Figs. 7.33 and 7.34. We can write equations similar to (7.5.8) for the left and the right transistors. These will be, correspondingly,

$$I_{DS} + \Delta i = -\frac{W}{x_1} \mu \int_{V_{SB}}^{V_1} Q'_I(V_{CB}(x)) dV_{CB}(x) \quad (7.5.12)$$

[†] Here B does not have to be small in order to write (7.5.10), since the power spectral density is constant (the noise is white). If the power spectral density were a function of frequency, B would have to be chosen very small.

**FIGURE 7.33**

(a) The transistor of Fig. 7.32a, with a fictitious dc source placed at point $x = x_1$ in the channel; (b) resulting effective reverse bias V_{CB} vs. distance.

**FIGURE 7.34**

Schematic representation of the situation in Fig. 7.33a.

$$I_{DS} + \Delta i = -\frac{W}{L - x_1} \mu \int_{V_1 + \Delta v}^{V_{DB}} Q'_I(V_{CB}(x)) dV_{CB}(x) \quad (7.5.13)$$

where V_1 is defined in Fig. 7.33b. Eliminating x_1 among these equations and using the assumption that Δv is very small easily gives (Prob. 7.26)

$$I_{DS} + \Delta i = -\frac{W}{L} \mu \int_{V_{SB}}^{V_{DB}} Q'_I(V_{CB}(x)) dV_{CB}(x) + \frac{W}{L} \mu Q'_I(V_{CB}(x_1)) \Delta v \quad (7.5.14)$$

As Δv goes to zero, $V_{CB}(x_1)$ has a well-defined value, in fact the same as in Fig. 7.32b. Recognizing in (7.5.14) the first term on the right-hand side as I_{DS} from (7.5.8), we obtain for the change Δi :

$$\Delta i = \frac{W}{L} \mu Q'_I(V_{CB}(x_1)) \Delta v \quad (7.5.15)$$

In the preceding development, we have assumed that Δv is a dc voltage. However, the result will be valid even if Δv is varying with time as long as the variation is slow enough so that quasi-static behavior is maintained (Sec. 6.2). This implies frequencies several times less than ω_s , just as was the case for the small-signal model we have already developed in Sec. 7.3. Let us now remove the battery and instead consider the thermal noise voltage generated across the small element of the channel centered at x_1 . Let Δv_t represent that part of the total thermal noise voltage that has frequency components in a bandwidth B within the preceding frequency range.[†] If Δi_t represents the corresponding drain current variation, we will have, in analogy with (7.5.15),

$$\Delta i_t(t) = \frac{W}{L} \mu Q'_I(V_{CB}(x_1)) \Delta v_t(t) \quad (7.5.16)$$

The mean square value of Δi_t can be found by noting that, for negligibly small Δv_t , $V_{CB}(x_1)$ has a practically constant, well-defined value, in fact the same value as in Fig. 7.32b, as already mentioned. (We do not need the actual value, as will be seen. We only need to know that it is practically independent of Δv_t for negligibly small Δv_t .) Thus, the mean square value of Δi_t will be

$$\overline{(\Delta i_t)^2} = \left[\frac{W}{L} \mu Q'_I(V_{CB}(x_1)) \right]^2 \overline{(\Delta v_t)^2} \quad (7.5.17)$$

Using now (7.5.11) in the previous relation, we obtain

$$\overline{(\Delta i_t)^2} = -4kT \frac{\mu}{L^2} W Q'_I(V_{CB}(x_1)) \Delta x \cdot B \quad (7.5.18)$$

[†] This assumption will result in models sufficient for many applications. Noise at higher frequencies is discussed in Chap. 8.

This gives the contribution of the element at x_1 to the drain current noise. The contributions of all similar elements in the channel are assumed uncorrelated and with zero average value, and one can thus find the mean square value of their combined effect by adding the individual mean square values. In the limit, letting Δx become a differential and integrating over the channel length (assuming constant mobility), we obtain

$$\bar{i_t^2} = -4kT \frac{\mu}{L^2} \left(\int_0^L Q'_I W dx \right) B \quad (7.5.19)$$

where we have used $\bar{i_t^2}$ to denote the total noise current mean square value in a bandwidth B . Recognizing the integral in the previous equation as the total inversion layer charge Q_I , and dividing both sides by B , we obtain the power spectral density of the white noise

$$S_{iw} = 4kT \frac{\mu}{L^2} (-Q_I)$$

(7.5.20)

Although this formula has been derived assuming strong inversion, it can also be derived without making this assumption. This is because, although V_{CB} in (7.5.7) and (7.5.8) was assumed to be the effective reverse bias between the channel and the body, in the general case the same equations can be written if one considers V_{CB} to be the potential corresponding to the quasi-Fermi potential split (Sec. 4.5). Due to this, it can be shown that (7.5.20) is valid in all regions of inversion.^{109,236a,236b} We now present some simple results in specific regions of inversion.

Strong Inversion Equation (7.5.20) is valid for any model, provided the appropriate expression is used for Q_I . In particular, for the simplified source-referenced strong-inversion model, we can use Q_I from (6.4.14). This gives

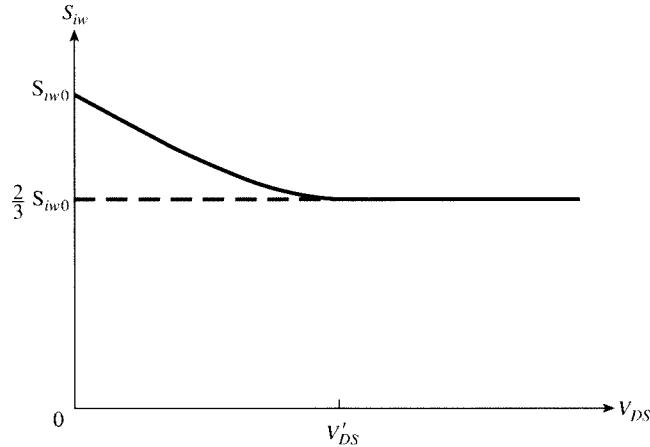
$$S_{iw} = 4kT \left[\frac{W}{L} \mu C'_{ox} (V_{GS} - V_T) \frac{2}{3} \frac{1 + \eta + \eta^2}{1 + \eta} \right] \quad (7.5.21)$$

where η has been defined in (4.7.25). S_{iw} is plotted vs. V_{DS} in Fig. 7.35. For a given bias, the power spectral density is *independent* of frequency, at least in the range of frequencies where the assumption of quasi-static operation is valid.

The quantity $(W/L)\mu C'_{ox}(V_{GS} - V_T)$ used in (7.5.21) can be easily measured or computed from simulation. This is because we have, from (7.2.37)

$$\frac{W}{L} \mu C'_{ox} (V_{GS} - V_T) = g_{sd} \Big|_{V_{DS}=0} \quad (7.5.22)$$

We now discuss the value of S_{iw} at two particular points of interest, nonsaturation with $V_{DS} = 0$ ($\eta = 1$), and saturation ($\eta = 0$). Using these values of η in (7.5.21), and employing (7.5.22) and (7.2.29a), we see that

**FIGURE 7.35**

Normalized power spectral density of MOS transistor thermal noise current versus V_{DS} . Operation is in strong inversion. S_{iw0} is the value of the power spectral density at $V_{DS} = 0$.

$$S_{iw} = 4kTg_{sd}, \quad V_{DS} = 0 \quad (7.5.23a)$$

$$= 4kT \frac{2}{3}\alpha g_m, \quad V_{DS} > V'_{DS} \quad (7.5.23b)$$

We thus see that in nonsaturation with $V_{DS} = 0$, the channel behaves as a resistor of value $R = 1/g_{sd}$, consistent with (7.5.3b). In saturation, the noise is equivalent to that of a resistor with conductance $\frac{2}{3}(\alpha g_m)$. Notice that S_{iw} is proportional to g_m only in saturation.[†]

Weak Inversion As mentioned below (7.5.20), this equation is valid in all regions of inversion. Q_l can be obtained from (6.4.36), repeated here:

$$Q_l = WL \frac{Q'_{l0} + Q'_{lL}}{2} \quad (7.5.24)$$

Expressions for the inversion layer charge per unit area at the source and drain ends of the channel were developed in Sec. 4.8. Using these in (7.5.24), and using the current expression (4.8.10) in the result, we easily get (Prob. 7.29)

[†] In some circuits literature and some circuit simulator models, (7.5.23b) has sometimes been used indiscriminately in both nonsaturation and saturation, which can lead to very wrong results. For example, at $V_{DS} = 0$, we have $g_m = 0$, which would predict zero noise if this formula were used! This is physically impossible and clearly false, as follows from (7.5.23a).

$$Q_I = \frac{L^2}{2\mu\phi_i} I'_{DS} (1 + e^{-V_{DS}/\phi_i}) \quad (7.5.25)$$

where I'_{DS} is the current in the saturation part of the I_{DS} - V_{DS} curve ($V_{DS} > 5\phi_i$). Using this in (7.5.20) and recalling that $\phi_i = kT/q$, where q is the electron charge, gives

$$S_{iw} = [2qI'_{DS} (1 + e^{-V_{DS}/\phi_i})] \quad (7.5.26)$$

For large V_{DS} ($> 5\phi_i$), this reduces to

$$S_{iw} = 2qI_{DS}, \quad V_{DS} > 5\phi_i \quad (7.5.27a)$$

and, using (7.2.50), we have

$$S_{iw} = 2kTng_m, \quad V_{DS} > 5\phi_i \quad (7.5.27b)$$

The physical origin of white noise in weak inversion¹¹⁴ is sometimes disputed. In some treatments, it is taken to be thermal noise,^{142,144} whereas in others it is assumed to be shot noise.¹³⁹ Both assumptions lead to the same result.[†] This is not surprising; as already mentioned in Sec. 7.5.1, shot noise and thermal noise are related.^{104,109}

Moderate Inversion and All-Region Models No simple moderate-inversion model has been derived analytically. However, as already mentioned, (7.5.20) is valid in all regions of inversion, including moderate inversion.¹⁵⁹ We can use in it the general expression for the inversion layer charge, (6.4.47), to obtain

$$S_{iw} = 4kT\mu \frac{W}{L} \left[-\frac{\frac{2}{3}(Q'_{IL}^2 + Q'_{IL}Q'_{I0} + Q'_{I0}^2) - \alpha\phi_i C'_{ox} (Q'_{IL} + Q'_{I0})}{Q'_{IL} + Q'_{I0} - 2\alpha\phi_i C'_{ox}} \right] \quad (7.5.28)$$

[†] It would appear that this equivalence does not hold as V_{DS} is reduced, since the right-hand side of (7.5.25) increases (in agreement with experiment^{142,144}), whereas the expression $2qI_{DS}$ predicts a decreasing value with decreasing V_{DS} (in fact, that value would be zero at $V_{DS} = 0$, a physically impossible result). This would tend to favor the assumption that the noise in weak inversion is of thermal origin.^{142,144} This discrepancy between the two theories is, however, removed if one views I_{DS} as the superposition of two components, one associated with the drain and one with the source^{151,158} (a similar superposition view is encountered in bipolar transistor theory). Thus, from (4.8.3) we can write $I_{DS} = I_1 - I_2$, where $I_1 = -(W/L)\mu\phi_i Q'_{I0}$ and $I_2 = -(W/L)\mu\phi_i Q'_{IL}$. If each of I_1 and I_2 is assumed to have shot noise (which has zero average value) and the two noise components are assumed uncorrelated, their mean square values will add. Using then equations from Sec. 4.8 it is easy to show (Prob. 7.29) that the combined power spectral density of the shot noise in the two currents is given *identically* by the right-hand side of (7.5.26) for *all* values of V_{DS} . In this way, both the shot noise assumption and the thermal noise assumption are seen to produce identical results for all bias points in weak inversion. A similar observation for the noise in a zero-biased *pn* junction is well known.¹⁰⁵

The charge densities in this expression can be evaluated as discussed in Chap. 4. Interpolation models have also been used for this purpose.^{55,57} An example will be shown in Table 7.1 at the end of this chapter.

Induced Gate Noise The preceding results have been based on random fluctuations of the potential in the channel. These fluctuations are coupled to the gate terminal through the oxide capacitance, and they “induce” a minute noise current in the gate terminal even if all external voltages are fixed.^{104,105,112,115} This phenomenon is usually neglected in models intended for operation at low or medium frequencies (which are the subject of this chapter). Induced gate noise is considered in Chap. 8.

Short-Channel Effects The results presented so far are for long-channel devices. Short-channel devices can be noisier than the previous formulas would predict.^{137,141,226–236} One reason for this is CLM, which can be accounted for by replacing L by $L - l_p$, where l_p is the pinchoff region, as discussed in Sec. 5.3. Velocity saturation cannot be accounted for that simply, as the entire derivations given previously were for long-channel devices, implicitly assuming that this effect is absent. This phenomenon affects both carrier velocity and the inversion charge density distribution along the channel. This changes the resistance of each incremental element along the channel and therefore changes the thermal noise. To include velocity saturation, rather than starting from (7.5.7), one can start from appropriate relations in Sec. 5.2, such as (5.2.6) with (5.2.14).^{236a,236d–239} The result can be put in the form of the ones we have presented previously, but multiplied by a factor, representing the “excess noise,” that accounts for velocity saturation.

Another short-channel phenomenon is that high electric fields present near the drain produce “hot carriers” in that vicinity, for sufficiently large V_{DS} values (Sec. 5.8). The effect of these can be modeled by an equivalent “carrier temperature” that is higher than the lattice temperature,¹⁰⁴ so it is not surprising that the thermal noise of short-channel devices increases.^{150,154,236d,240,241} An *excess noise factor* can be defined again, as the factor by which the noise current power spectral density is higher than that predicted by long-channel theory. Some measurements^{150,227} and models²³⁴ have claimed a significant excess noise factor due to hot electrons; however, this has been disputed. Irrespective of the carrier temperature, which is greater than the lattice temperature at the drain end of the channel, the inversion charge density decreases at the drain end, to a negligible value in saturation, so the contribution of the inversion charge near the drain end of a device to the overall thermal noise should be small. Also, as noise models have been improved, by including short-channel effects and by adding shot noise sources for substrate and gate currents and thermal noise sources for parasitic resistance (see the following), the discrepancies between models and experimental data have decreased over a wide range of technologies,^{240–243} and some of the noise increase observed in short-channel devices can be explained without invoking hot electrons. Nevertheless, the issue of excess noise is not considered resolved and is still an active area of investigation.

We have seen in Sec. 5.8 that impact ionization is responsible for the drain-to-substrate current, I_{BD} . This current shows noise of its own,²²⁷ at low substrate current values, this has been found to be shot noise (Sec. 7.5.1), with power spectral density of $2qI_{DB}$.^{237,245–247} Because the drain current includes I_{DB} as a component [see (5.8.5)], the noise of this current is added to the drain-source current noise. Also, because I_{DB} produces a voltage drop across the substrate resistance, its noise modulates the effective substrate potential; the fluctuations in the latter are coupled to the drain current through the substrate transconductance g_{mb} and can contribute to an increase in the observed drain current noise. Figure 7.36 shows noise power spectral density for the drain current vs. V_{DS} , with V_{GS} as a parameter. While at low V_{DS} , the general behavior is qualitatively similar to that in Fig. 7.35, increased noise due to impact ionization can be seen at larger values of V_{DS} .

The currents in the source-bulk and drain-bulk reverse-biased junctions, even with enhanced leakage induced by the high fields in the gate-source and gate-drain overlap regions (Sec. 5.13), are generally small, and so is the shot noise contributed by them.

As discussed in Sec. 5.12, modern devices with thin gate dielectrics exhibit a dc gate tunneling current, I_G . This current exhibits shot noise, with power spectral density $2qI_G$.^{237,248} Flicker noise has also been reported for this current and is variously attributed to the influence of oxide traps on tunneling and on trap-assisted tunneling.^{237,247}

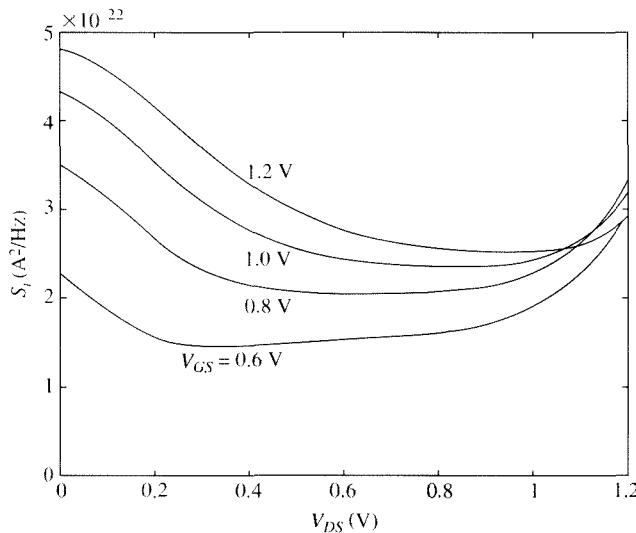


FIGURE 7.36

Power spectral density of thermal noise current vs. V_{DS} , with V_{GS} as a parameter, for an n MOS device with $W = 10 \mu\text{m}$ and $L = 0.09 \mu\text{m}$. $t_{\text{ox}} = 1.5 \text{ nm}$, $N_A = 3 \times 10^{18} \text{ cm}^{-3}$.

7.5.3 Flicker Noise

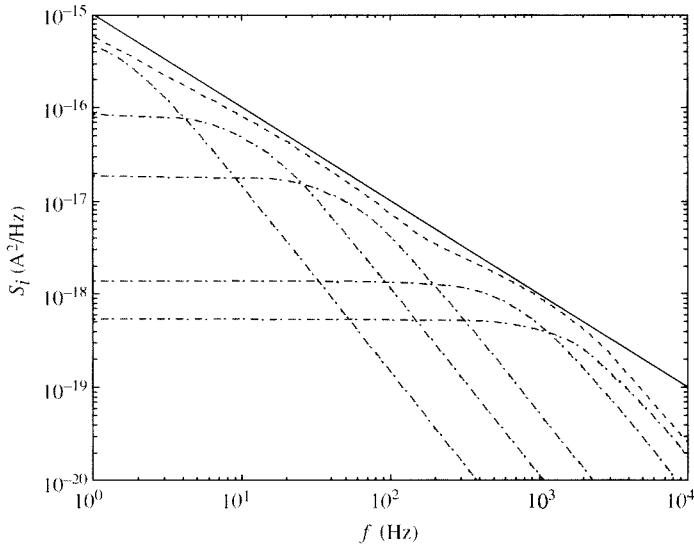
As shown in Fig. 7.31, at low frequencies flicker noise is dominant.[†] This noise is also called “ $1/f$ ” noise, because the power spectral density is nearly proportional to the inverse of the frequency. Flicker noise in MOS transistors has been the subject of intensive studies for several decades. There have been several theories for the origin of this noise, with involved physics and sometimes conflicting conclusions; the reader is referred to the literature for more information.^{110,163–232,249,251–258} Here we will only summarize the main conclusions of the two dominant classical theories, and then we will present a more recent, unifying approach.

Random Carrier Number Fluctuations The first theory attributes the origin of flicker noise to the random fluctuation of the number of carriers in the channel, due to fluctuations in the surface potential; the fluctuations are in turn caused by trapping and releasing of carriers by traps located near the Si-SiO₂ interface.^{163,166,173,174,176,178,197,216,251,252} Traps are spatially localized available energy levels in what is normally the forbidden energy gap, caused by imperfections in the lattice structure. When traps capture or release a carrier from the channel (even ones within the oxide can do this, through tunneling), the oxide charge changes. We have seen in Sec. 2.2 that the oxide charge can be represented by an effective interface charge per unit area, Q'_o . Thus, when a trap captures or releases a carrier, Q'_o changes by a minute amount, and so does the flatband voltage, V_{FB} in (2.2.6). This, for a given gate-source voltage, changes the channel current, I_{DS} . Let us assume initially that only one trap is present. Then, the captures and release of carriers will vary I_{DS} by a small amount ΔI_{DS} . The quantity ΔI_{DS} , plotted vs. time, will be a square wave with random transition times, corresponding to the random times of capture and release. The power spectral density of such a waveform can be shown to be of the form¹⁰⁴

$$S_{i,\text{single trap}}(f) = \frac{c_t \tau_t}{1 + (2\pi f)^2 \tau_t^2} \quad (7.5.29)$$

where c_t is a constant, and τ_t is a characteristic time, related to the average time between captures and releases by the trap in question. The deeper the trap is into the oxide, the less likely it is that it will capture a carrier, and the larger τ_t will be. In fact, τ_t is an exponential function of the distance between the trap site and the oxide-semiconductor interface. This type of spectrum is called “Lorentzian” and does not have a $1/f$ characteristic; rather it is constant at low frequencies and decreases as $1/f^2$ at high frequencies. However, if a large number of traps, with uniformly distributed τ_t values, operate simultaneously and independently of each other, the superposition

[†] Note that the fact that flicker noise is dominant at low frequencies does not imply that it affects circuits only at such frequencies. This is because if two signals of frequencies f_1 and f_2 are passed through a nonlinear circuit, the output of the circuit can contain not just components at these two frequencies, but also components at frequencies of the form $nf_1 \pm mf_2$, where n and m are integers. Thus, low-frequency noise can “mix” with high-frequency signals and appear as high-frequency noise. Through this mechanism, flicker noise can affect the performance of high-frequency circuits such as oscillators, even in the gigahertz range.

**FIGURE 7.37**

Drain current noise spectral power density vs. frequency. Solid line is ideal $1/f$ characteristic. Dash-dot lines are Lorentzian spectra from five traps with random time constants. Dashed line is the sum of the five individual trap spectra; even with only five traps it rather well approximates $1/f$ behavior.

of the individual spectra approximates $1/f$ behavior. This is illustrated in Fig. 7.37; although only five traps are used in this example, the combined spectrum (upper broken line) approaches a $1/f$ characteristic (solid line, shown for comparison). For larger-area devices, containing many traps, a constant slope characteristic is approached much more closely. Experimental data show that this slope sometimes deviates from exact $1/f$ behavior, following rather a dependence of the form $1/f^c$, where c is between 0.7 and 1.2. This has been quantitatively explained by the trap distribution in the gate dielectric not being uniform, but varying with distance from the interface.^{166,188}

A quantitative analysis of flicker noise in MOS devices is possible, but it is involved. Here we will give some intuitive arguments, which will allow us to expect certain key features of such results. As suggested previously, the capturing and releasing of carriers by traps can be thought of as randomly varying the effective interface charge density Q'_o by a small amount $\Delta Q'_o$, thus modulating the flat-band voltage V_{FB} through the term Q'_o/C'_{ox} in (2.2.6), by the amount $\Delta Q'_o/C'_{ox}$.²¹⁰ The mean square value of this voltage will thus be proportional to $(1/C'_{ox})^2$, and its effect on the mean-square value of the drain-source current noise can be found by multiplying by g_m^2 , where g_m is the transconductance.[†] Finally, because flicker noise

[†] In various expressions, we have seen for the drain-source current, V_{FB} is encountered in differences with the applied gate-source bias voltage, V_{GS} ; thus, variations in V_{FB} affect the drain current in the same way as variations in V_{GS} (other than an algebraic sign difference, which plays no role because there is a squaring operation involved).

results from the superposition of such variations due to many traps, the larger the gate area WL , the more the effects of these variations will tend to average out, and the smaller the resulting flicker noise. All of these observations can be confirmed in the final form for the power spectral density of the current noise, which can be shown to be^{197,210}

$$S_{if}(f) = g_m^2 S_{vf}(f) \quad (7.5.30)$$

where[†]

$$S_{vf}(f) = \frac{K_1}{C'_{ox}^2} \frac{1}{WL} \frac{1}{f^c} \quad (7.5.31)$$

where, as already mentioned, the exponent c is between 0.7 and 1.2, and K_1 is a quantity dependent on fabrication details, in general being lower for cleaner processes and for processes that avoid surface damage. Typical values for this quantity are in the 10^{-31} to $10^{-29} \text{ C}^2 \cdot \text{cm}^{-2}$ range.

Mobility Fluctuations A second theory attributes flicker noise to mobility fluctuations, due to lattice scattering.^{184,192–194,216,222,253} Results based on this theory suggest that the power spectral density for the equivalent input noise current is given by (7.5.30), but with $S_{vf}(f)$ given by

$$S_{vf}(f) = \frac{K(V_{GS})}{C'_{ox}} \frac{1}{WL} \frac{1}{f^c} \quad (7.5.32)$$

where $K(V_{GS})$ is a bias-dependent quantity. The inverse proportionality to C'_{ox} is not universally accepted. The noise of *p*MOS devices has in the past been attributed to mobility fluctuations, with $K(V_{GS})$ of the order of 6×10^{-26} to $2 \times 10^{-23} \text{ V}^2 \cdot \text{F}$ at $|V_{GS} - V_T|$ of about 1 V. For *p*MOS devices, it has been found^{192,216} that, in strong inversion, $K(V_{GS})$ increases approximately linearly with $|V_{GS} - V_T|$ up to several fold; it has also been reported that, in weak inversion, $K(V_{GS})$ increases as $|V_{GS}|$ is decreased.²¹⁹

Unified Theory The unified theory holds that flicker noise is due to both carrier number fluctuations and mobility fluctuations.^{155,196,211,254,255} After all, the carrier number fluctuation theory talks about a randomly varying charge at traps near the interface, and we have seen that charge can affect mobility through “Coulomb

[†] The quantity $S_{if}(f)$ is sometimes considered to be the power spectral density of an “equivalent input noise voltage,” i.e., modeled as a voltage source in series with the gate. However, such a model would yield the same current noise as (7.5.30) only if the external terminal of this source is grounded, or at least is connected to a low-impedance point at the frequencies of interest. This issue is considered in Prob. 7.28. The correct representation of noise in terms of equivalent input sources requires two such elements: a voltage source and a current source.¹⁰⁶

scattering" (Sec. 4.10); so it is not surprising that both effects may be present and correlated in a given device.²⁰⁷ Note that the unified approach does not merge the separate theories discussed previously; rather, it adds a correlated mobility fluctuation term to the carrier number fluctuation theory (and it attributes mobility fluctuations to Coulomb scattering, which is not the case with the classical mobility fluctuations theory).

To analytically determine flicker noise using this theory, it is typically assumed that the traps are distributed uniformly over the gate area. However, the *occupied* trap density depends on the inversion charge density, which is a function of position.^{123,155,188,197} In turn, the charge in the semiconductor is linked to the trapped charge through the charge balance equation, (2.3.4). The drain current at position x in the channel is as in (4.3.4), and in that equation *both* the inversion layer charge *and* the mobility are assumed to be affected by the random capture and release of charge by the traps.^{155,255} In this way, the local disturbance at position x can be determined, and then the combined effect of all such disturbances on the drain-source current can be determined through integration, in a manner similar to that used in Sec. 7.5.2. Several assumptions and approximations are made on the way. The resulting expressions are complex, not suitable for hand analysis, and are given on the book's website,⁹⁹ along with a detailed derivation. Such results can be approximated by using (7.5.30), while augmenting (7.5.31) by making K_1 a function of V_{GS} :

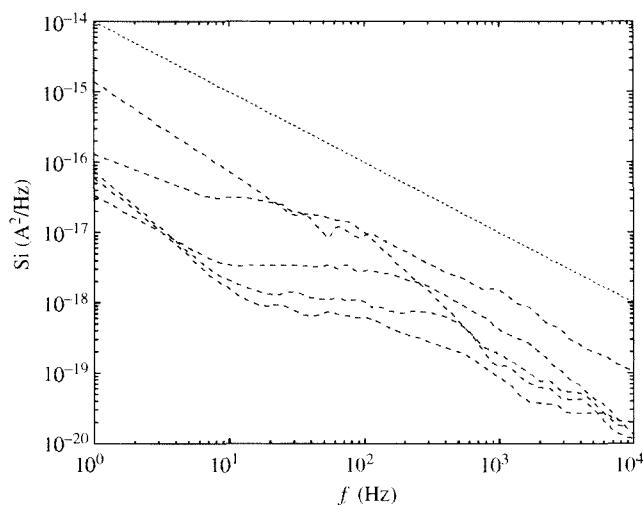
$$S_{vf}(f) = \frac{K_1(V_{GS})}{C_{ox}'^2} \frac{1}{WL} \frac{1}{f^c} \quad (7.5.33)$$

where in weak inversion $K_1(V_{GS})$ is constant, close to the values given previously, and in strong inversion is roughly proportional to $(V_{GS} - V_T)^2$.^{256c} Deep in strong inversion, increases by a factor of 2 are not uncommon for *n*MOS devices; for *p*MOS devices, increases by a factor as high as 50 have been reported. Results from several fabrication processes can be found in Ref. 257. Due to the large variation of behavior in this respect from process to process, often empirical parameters are introduced into CAD models of flicker noise.

The unified flicker noise model, or variations of it, are widely used and have proven to fit experimental data reasonably well. However, the model is not universally accepted, one criticism being that Coulomb scattering due to the varying trap charge is too small to produce significant mobility fluctuations.²⁴⁹

It has been reported that for *p*MOS devices, buried-channel operation (Sec. 9.3.4) results in much lower flicker noise (by up to two orders of magnitude).^{201,222} This is attributed to the fact that the channel is farther away from the Si-SiO₂ interface and is thus less affected by interface traps. However, in modern processes both *p*MOS and *n*MOS devices are surface-channel.

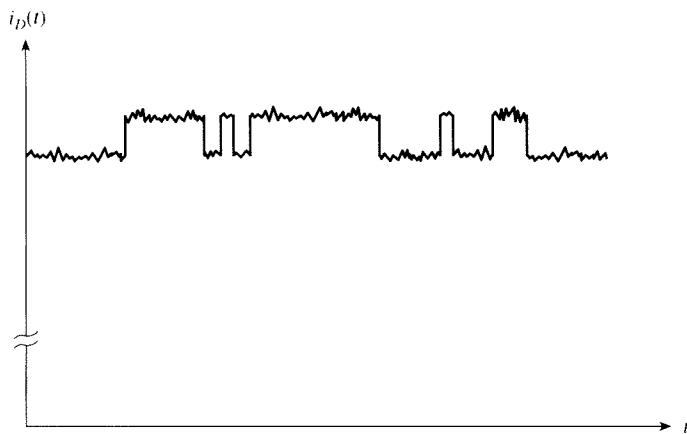
Short-Channel Effects We have seen that the 1/f characteristic of flicker noise is the result of the superposition of many spectra due to many traps.^{237,257,258} However, if the gate area is very small, there will be only a few traps that can

**FIGURE 7.38**

Measured noise characteristics of identical *n*MOS devices from five different sites on a wafer. $W = 0.7 \mu\text{m}$, $L = 0.28 \mu\text{m}$, $V_{GS} = 1.1 \text{ V}$, and $V_{DS} = 1.0 \text{ V}$. Dotted line is $1/f$ reference slope.

exchange charge with the channel, and their individual effects will be noticed, rather than tending to average out as in the case of large gate area. In that case, the overall spectrum cannot be expected to follow $1/f$ behavior. Figure 7.38 shows measured noise spectra of identical *n*MOS devices measured at five different sites on a wafer. An approximate Lorentzian characteristic [see discussion in association with (7.5.29)] is discernible for some of the devices, and the variation between devices is significant. This is consistent with the number of traps, and their characteristic time constants, varying statistically between devices. It is, in fact, possible that only a single trap of this type exists in a very small device. Then, as it captures and releases charge, abrupt changes in the drain current can be noticed, as shown in Fig. 7.39; these changes will be on top of the more common noise variation.^{250,259–266} This is reminiscent of the so-called random telegraph signals (RTS) in communication theory, and for this reason is often referred to as “*random telegraph noise* (RTN)”. The variation of the drain current due to this effect can be significant (e.g., up to a few tenths of a percent of the dc current value for a device with gate area of about $0.5 \mu\text{m}^2$ in strong inversion, but it can increase significantly if the gate area and or $V_{GS} - V_T$ are decreased; measurements in the tens of percent have been reported in weak inversion for very small devices).²⁵⁰

Because, with only very few traps present, we can no longer assume that they are uniformly distributed over the channel area, it is possible that more traps are present near the drain than near the source. When a device is operated in the forward saturation region, the extra traps near the drain may not influence the current much, as they may be located in the pinchoff region, which does not affect significantly the value of the current. If, however, the device is operated in nonsaturation, that region

**FIGURE 7.39**

Typical variation of the drain current in a device with extremely small gate area, due to random telegraph noise.

of the channel affects the channel charge and the current, and the extra traps cause an increase in flicker noise. The same happens if the preceding device is operated in reverse saturation (with the source and drain terminals interchanged), in which case the extra traps will lie next to the new “source.” Such effects can be exacerbated if hot carrier effects have compromised the quality of the oxide near the interface and have introduced additional traps near the interface in the area close to the drain.^{226,228–232,235} This can happen, for example, if a device has been operated at or near punchthrough for several hours. When the drain voltage is reduced and the device is operated normally, the preceding effects can be strong, and excess flicker noise of over an order of magnitude can be observed.

Flicker noise is essentially the superposition of many RTS waveforms.²⁶⁰ It is reasonable to ask at this point whether it is possible to have *no* traps, and thus *no* flicker noise! There is no fundamental reason as to why this would not be possible.[†]

Other effects not discussed here involve quantum effects in the inversion layer, which have been claimed to be responsible for what we have described previously as an increase of $K_l(V_{GS})$ with V_{GS} in strong inversion.^{164,236c}

Flicker noise also appears to be larger in modern short-channel devices formed with halo implants.²⁷¹

[†] In fact, a well-known researcher in the field has mentioned to one of the authors that, a long time ago, he made such devices. The reason he has decided not to announced this startling result (and does not want to be quoted) is that he has not been able to reproduce it since then! He still has the devices on his shelf, and they still show no flicker noise, but whenever he has tried to make more devices using the same process, the flicker noise was back. Something went right that day. Although this information is far from specific, perhaps it can serve as motivation for device designers!

7.5.4 Noise in Extrinsic Resistances

The complete MOS transistor includes parasitic resistances associated with each terminal (Secs. 6.8 and 7.4). Each of these contributes thermal noise (Sec. 7.5.1) and can be modeled as in Fig. 7.30. Gate and body resistance noise couples to the drain current through the gate and body transconductance, respectively.^{148,267} The series source and drain resistances also contribute thermal noise to the drain current, although the amount is relatively small compared with other noise sources,²³⁷ provided these resistances are kept sufficiently low through multiple contacts.

7.5.5 Including Noise in Small-Signal Circuits

Noise voltages and currents are small signals, and noise analysis in circuits is done through small-signal analysis. Noise sources are added to small-signal equivalent circuits in the appropriate places. For example, in the model of Fig. 7.2, thermal and flicker noise in the drain current can be included by adding a noise current source as shown in Fig. 7.40. The spectral density of this current source will be given by (7.5.6). Additional noise sources can be included to model the four resistances associated with the four terminals and the impact ionization drain-body current and gate leakage current (not shown).

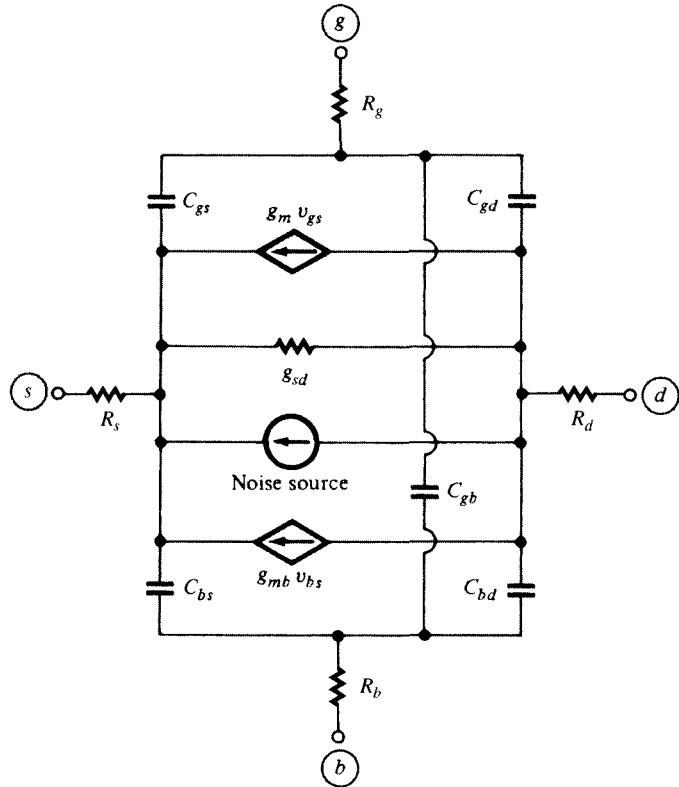
In our discussion of small-signal models, we have come up with a number of small-signal resistances, all of which are shown in Fig. 7.6. We caution that it would be a mistake to assume that these resistances contribute thermal noise. This is because these are *not* real, physical resistances, but rather fictitious elements representing derivatives of voltages with respect to currents (Sec. 7.2). Thus, for example, no noise should be assigned to g_{sd} in Fig. 7.40.

When using such models to calculate noise in circuits, one has to work with mean square or root mean square values of noise currents and voltages, rather than with the noise currents and voltages themselves, and this requires some care. Convenient techniques for such calculations in hand analysis are given in several texts.^{104–106} The efficient calculation of noise in circuit simulation programs is discussed in Ref. 268.

7.6 All-Region Models

General expressions for small-signal parameters can be developed for the all-region model^{51–53} discussed in Sec. 4.3, as well as for the EKV model^{54–58,236b} discussed in Sec. 4.4.4. An example of a set of expressions valid in weak-, moderate-, and strong-inversion *saturation* regions is shown in Table 7.1. Except for CLM, short-channel effects are not included. All symbols not defined in the table are as used in the corresponding sections of this chapter.

The value of n shown is a compromise for medium accuracy in all regions, and it is adequate for simple estimates. In strong inversion, if the expected voltage ranges are significant, better accuracy for I_{DS} is obtained if n is replaced by α as used in Sec. 4.7.3. However, unless the transition between n and α is made gradually as the moderate inversion is crossed, discontinuities will occur.

**FIGURE 7.40**

A small-signal equivalent circuit for the intrinsic part of a transistor, with a noise current source added; the power spectral density of this source is given by (7.5.6).

TABLE 7.1
A saturation region long-channel nMOS model valid in all regions of inversion[†]

$$\text{Drain current: } I_{DS} = I_Z \ln^2 \left[1 + \exp \left(\frac{V_{GS} - V_T}{2n\phi_f} \right) \right]$$

Small-signal conductances:

$$g_m = \frac{I_{DS}}{n\phi_f} \frac{1}{f(u)}$$

$$g_{mb} = (n-1)g_m$$

$$g_{sd} = \frac{I_{DS}}{V_A}$$

(Continued)

[†] The current equation²⁶⁹ has been adopted from Refs. 54–58 for using the source as a reference and for making possible the inclusion of an accurately evaluated V_T . The small-signal and noise expressions are from Refs. 55 and 57. The expression for $f(u)$ is from Refs. 52 and 53.

TABLE 7.1 *Continued*

Small-signal *intrinsic* capacitances:

$$C_{gs} = WLC'_{ox} \left[\frac{3}{2} + \frac{f(u)}{u} \right]^{-1}$$

$$C_{bs} = (n-1)C_{gs}$$

$$C_{gb} = WLC'_{ox} \frac{n-1}{n} \left[1 - \left(\frac{3}{2} + \frac{f(u)}{u} \right)^{-1} \right]$$

$$C_{gd} = C_{bd} = 0$$

Noise power spectral densities:

$$S_{bv}(f) = 4kT \left\{ \frac{1}{2} + \frac{1}{6} \frac{u}{[f(u)]^2} \right\} n g_m$$

$$S_{vf}(f) = \frac{K_1}{C'^2_{ox}} \cdot \frac{1}{WL} \cdot \frac{1}{f}$$

$$I_Z = \frac{W}{L} \mu C'_{ox} (2n\phi_t^2)$$

$$V_T = V_{T0} + \gamma \left(\sqrt{V_{SB} + \phi_0} - \sqrt{\phi_0} \right)$$

$$n = 1 + \frac{\gamma}{2\sqrt{V_{SB} + \phi_0}}$$

$$u = \frac{I_{DS}}{I_Z}$$

$$f(u) = \frac{1}{2} \left(\sqrt{1+4u} + 1 \right)$$

References

1. H. J. K. Ihantola and J. L. Moll, "Design theory of a surface field-effect transistor," *Solid-State Electronics*, vol. 7, pp. 423–430, 1964.
2. C. T. Sah, "Characteristics of the metal-oxide semiconductor transistors," *IEEE Transactions on Electron Devices*, vol. ED-11, pp. 324–345, 1964.
3. S. R. Hofstein and G. Warfield, "Carrier mobility and current saturation in the MOS transistor," *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 129–138, 1965.
4. V. K. G. Reddi and C. T. Sah, "Source to drain resistance beyond pinch-off in metal-oxide semiconductor transistors (MOST)," *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 139–141, 1965.
5. M. H. White and R. C. Gallagher, "Metal oxide semiconductor (MOS) small-signal equivalent circuits," *Proceedings of the IEEE*, vol. 53, pp. 314–315, 1965.

6. C. T. Sah and H. C. Pao, "The effects of fixed bulk charge on the characteristics of metal-oxide semiconductor transistors," *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 393–409, 1966.
7. M. B. Das, "Charge-control analysis of M.O.S. and junction-gate field-effect transistors," *IEE Proceedings*, vol. 113, pp. 1565–1570, 1966.
8. F. A. Lindholm, R. J. Balda, and J. L. Clements, "Characterization of the four-terminal MOS transistor for digital and linear applications," *Proceedings of the IEEE International Electronics Conference*, Toronto, 1967, pp. 116–117.
9. M. B. Das, "Dependence of the characteristics of MOS transistors on the substrate resistivity," *Solid-State Electronics*, vol. 11, pp. 305–322, 1968.
10. J. E. Schroeder and R. S. Muller, "IGFET analysis through numerical solution of Poisson's equation," *IEEE Transactions on Electron Devices*, vol. ED-15, no. 12, pp. 954–961, 1968.
11. T. L. Chiu and C. T. Sah, "Correlation experiments with a two-section model theory of the saturation drain conductance of MOS transistors," *Solid-State Electronics*, vol. 11, pp. 1149–1163, 1968.
12. D. Frohman-Bentchkowsky and A. S. Grove, "Conductance of MOS transistors in saturation," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 108–113, 1969.
13. R. S. C. Cobbold, *Theory and Applications of Field-Effect Transistors*, Wiley-Interscience, New York, 1970.
14. G. Baum and H. Beneking, "Drift velocity saturation in MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-17, pp. 481–482, 1970.
15. D. J. Hamilton, F. A. Lindholm, and A. H. Marshak, *Principles and Applications of Semiconductor Device Modeling*, Holt, Rinehart, and Winston, New York, 1971.
16. J. E. Meyer, "MOS models and circuit simulation," *RCA Review*, vol. 32, pp. 42–63, 1971.
17. G. Merckel, J. Borel, and N. Z. Cupeea, "An accurate large-signal MOS transistor model for use in computer-aided design," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 681–690, 1972.
18. A. Popa, "An injection level dependent theory of the MOS transistor in saturation," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 774–781, 1972.
19. E. Vittoz and J. Fellrath, "MOS analog integrated circuits based on weak inversion operation," *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 224–231, 1977.
20. J. Fellrath and E. Vittoz, "Small signal model of MOS transistors in weak inversion," *Procédures des Journées d'Electronique 1977*, Session C4, École Polytechnique Fédérale de Lausanne, Switzerland, pp. 315–324, 1977.
21. P. Rossel, H. Martinot, and G. Vassilieff, "Accurate two-sections model for MOS transistors in saturation," *Solid-State Electronics*, vol. 19, pp. 51–56, 1976.
22. F. M. Klaassen, "A MOS model for computer-aided design," *Philips Research Reports*, vol. 31, pp. 71–83, 1976.
23. Y. A. El-Mansy and A. R. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Transactions on Electron Devices*, vol. ED-24, pp. 254–262, 1977.
24. G. W. Taylor, "The effects of two-dimensional charge sharing on the above-threshold characteristics of short-channel devices," *Solid-State Electronics*, vol. 22, pp. 701–717, 1979.
25. H. C. Poon, " V_{th} and beyond," presented at the Workshop on Device Modelling for VLSI, Burlingame, Calif., March 29, 1979; also L. Cong, Bell Laboratories, private communication.
26. T. Poorter and J. H. Satter, "A D.C. model for an MOS-transistor in the saturation region," *Solid-State Electronics*, vol. 23, pp. 765–772, 1980.
27. G. Cardinali, S. Graffi, M. Impronta, and G. Masetti, "DC MOSFET model for analogue circuit simulation employing process-empirical parameters," *IEE Proceedings*, vol. 129, part 1, pp. 61–66, 1982.
28. S. Liu and L. W. Nagel, "Small-signal MOSFET models for analog circuit design," *IEEE Journal of Solid-State Circuits*, vol. SC-17, pp. 983–998, 1982.
29. C. Turchetti, G. Masetti, and Y. Tsividis, "On the small-signal behavior of the MOS transistor in quasi-static operation," *Solid-State Electronics*, vol. 26, pp. 941–949, 1983.
30. M. El-Nokali and H. Miranda, "A simple model for the MOS transistors in saturation," *Solid-State Electronics*, vol. 29, pp. 591–596, 1986.
31. C. Nguyen-Duc, S. Cristoloveanu, and G. Ghibaudo, "A three-piece model of channel length modulation in submicrometer MOSFETs," *Solid-State Electronics*, vol. 31, pp. 1057–1063, 1988.

32. F. S. Shoucair, "A semi-empirical model of the MOSFET's small-signal drain conductance in saturation for analog circuit design," *IEEE Transactions on Electron Devices*, vol. 39, pp. 1246–1248, 1992.
33. J. H. Huang, Z. H. Liu, M. C. Jeng, P. K. Ko, and C. Hu, "A physical model for MOSFET output resistance," *Procedures of the IEEE International Electron Devices Meeting*, pp. 569–579, 1992.
34. M. Fujishima and K. Asada, "A nonpinchoff gradual channel model for deep-submicron MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, pp. 1883–1885, 1993.
35. R. M. D. A. Velghe, D. B. M. Klaassen, and F. M. Klaassen, "Compact MOS modeling for analog circuit simulation," *Procedures of the IEEE International Electron Devices Meeting*, pp. 485–488, December 1993.
36. M. Miura-Mattausch, "Analytical MOSFET model for quarter-micron technologies," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 13, pp. 610–615, 1994.
37. J. K. Seon and K. S. Yoon, "A precision output conductance model for analog CMOS circuit simulations," *Procedures of the IEEE International Electron Devices Meeting*, pp. 1584–1587, December 1995.
38. W. R. Bandy and R. S. Winton, "A new approach for modeling the MOSFET using a simple, continuous analytical expression for drain conductance which includes velocity-saturation in a fundamental way," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, pp. 475–843, 1996.
39. Y. Cheng, M.-C. Jeng, Z. Liu, K. Chen, M. Chan, C. Hu, and P. K. Ko, "An investigation on the robustness, accuracy and simulation performance of a physics-based deep-submicrometer BSIM model for analog/digital circuit simulation," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 321–324, 1996.
40. Y. G. Chen and J. B. Kuo, "A unified triode/saturation model with an improved continuity in the output conductance suitable for CAD of VLSI circuits using deep sub-0.1 μm NMOS devices," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, pp. 256–258, 1996.
41. Y. Cheng, M.-C. Jeng, Z. Liu, J. Huang, M. Chan, K. Chen, P. K. Ko, and C. Hu, "A physical and scalable I-V model in BSIM3v3 for analog/digital circuit simulation," *IEEE Transactions on Electron Devices*, vol. 44, pp. 277–287, 1997.
42. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
43. N. Arora, *MOSFET Models for VLSI Circuit Simulation—Theory and Practice*, Springer-Verlag, Vienna, 1993.
44. K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, N.J., 1993.
45. Y. Tsividis and G. Masetti, "Problems in precision modeling of the MOS transistor for analog applications," *IEEE Transactions on Computer-Aided Design*, vol. CAD-3, pp. 72–79, 1983.
46. C. C. McAndrew, B. K. Bhattacharyya, and O. Wing, "A single-piece C₋continuous MOSFET model including subthreshold conduction," *IEEE Electron Device Letters*, vol. 12, pp. 565–567, 1991.
47. S. Csereny, "MOS small-signal conductances in the weak avalanche multiplication region," *IEEE Transactions on Electron Devices*, vol. 39, pp. 1233–1235, 1992.
48. Y. Tsividis, "Moderate inversion in MOS devices," *Solid-State Electronics*, vol. 25, pp. 1099–1104, 1982; see also Erratum, *Solid-State Electronics*, vol. 26, p. 823, 1983.
49. H.-S. Wong, M. H. White, T. J. Krutsick, and R. V. Booth, "Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFETs," *Solid-State Electronics*, vol. 30, pp. 953–968, 1987.
50. J. R. Brews, "A charge sheet model of the MOSFET," *Solid-State Electronics*, vol. 21, pp. 345–355, 1978.
51. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An explicit physical model for the long-channel MOS transistor including small-signal parameters," *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
52. A. I. A. Cunha, *Um Modelo do Transistor MOS para Projeto de Circuitos Integrados*, Ph.D. thesis, Universidade Federal de Santa Catarina, December 1996 (in Portuguese).
53. A. I. A. Cunha, O. C. Gouveia-Filho, M. C. Schneider, and C. Galup-Montoro, "A current-based model for the MOS transistor," *Proceedings of the International Symposium on Circuits and Systems*, Hong Kong, 1997, pp. 1608–1611.

54. H. J. Oguey and S. Cserveny, "MOS Modeling at Low Current Density," Summer Course on Process and Device Modeling, Catholic University of Leuven, Belgium, 1983.
55. C. Enz, *High Precision CMOS Micropower Amplifiers*, Ph.D. thesis no. 802, École Polytechnique Fédéral de Lausanne, Switzerland, 1989.
56. E. A. Vittoz, "Micropower Techniques," J. E. Franca and Y. Tsividis (editors), *Design of Analog-Digital VLSI Circuits for Telecommunications and Signal Processing*, Prentice Hall, Englewood Cliffs, N.J., 1994.
57. C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, 1995.
58. M. Bucher, C. Lallement, C. Enz, F. Theodoloz, and F. Krummenacher, "Scalable GM/I based MOSFET model," *Proceedings of the International Semiconductor Device Research Symposium*, Charlottesville, Va., December 1997, pp. 615–618.
59. Y. P. Tsividis, "Reaction between incremental intrinsic capacitances and transconductances in MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 946–948, 1980.
60. D. E. Ward, "Charge-based modeling of capacitance in MOS transistors," *Technical Report G201-11*, Integrated Circuits Laboratory, Stanford University, Stanford, Calif., 1981.
61. J. J. Paulos, D. A. Antoniadis, and Y. P. Tsividis, "Measurement of intrinsic capacitances of MOS transistors," *Proceedings of the IEEE International Solid-State Circuits Conference*, San Francisco, 1982, pp. 238–239.
62. J. J. Paulos and D. A. Antoniadis, "Limitations of quasi-static capacitance models for the MOS transistor," *IEEE Electron Device Letters*, vol. EDL-4, pp. 221–224, 1983.
63. C. Turchetti, G. Masetti, and Y. Tsividis, "On the small-signal behavior of the MOS transistor in quasi-static operation," *Solid-State Electronics*, vol. 26, pp. 941–949, 1983.
64. G. I. Serhan and S.-Y. Yu, "A simple charge-based model for MOS transistor capacitances: A new production tool," *IEEE Transactions on Computer-Aided Design*, vol. CAD-2, p. 48, 1983.
65. E. W. Greeneich, "An analytical model for the gate capacitance of small-geometry MOS structures," *IEEE Transactions on Electron Devices*, ED-30, pp. 1838–1839, 1983.
66. J. Oristian, H. Iwai, J. Walker, and R. Dutton, "Small geometry MOS transistor capacitance measurement method using simple on-chip circuits," *IEEE Electron Device Letters*, vol. EDL-5, pp. 395–397, 1984.
67. B. J. Sheu and P. K. Ko, "An analytical model for intrinsic capacitances of short-channel MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, San Francisco, 1984, pp. 300–303.
68. H. Iwai, J. E. Oristian, J. T. Walker, and R. W. Dutton, "A scalable technique for the measurement of intrinsic MOS capacitance with attofarad resolution," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 344–356, 1985.
69. J. J. Paulos and D. A. Antoniadis, "Measurement of minimum-geometry MOS transistor capacitances," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 357–363, 1985.
70. H. Iwai, M. R. Pinto, C. S. Rafferty, J. E. Oristian, and R. W. Dutton, "Velocity saturation effect on short-channel MOS transistor capacitance," *IEEE Electron Device Letters*, vol. EDL-6, pp. 120–122, 1985.
71. T. Shima, "Table lookup MOSFET capacitance model for short-channel devices," *IEEE Transactions on Computer-Aided Design*, vol. CAD-5, pp. 624–632, 1986.
72. H. Iwai, M. R. Pinto, C. S. Rafferty, J. E. Oristian, and R. W. Dutton, "Analysis of velocity saturation and other effects on short-channel MOS transistor capacitances," *IEEE Transactions on Computer-Aided Design*, vol. CAD-6, pp. 173–184, 1987.
73. Y. Ohkura, T. Toyabe, and H. Masuda, "Analysis of MOSFET capacitances and their behavior at short-channel lengths using an AC device simulator," *IEEE Transactions on Computer-Aided Design*, vol. CAD-6, pp. 423–430, 1987.
74. F. Riedel, "Kleinsignalmodell für kurz- und schmalkanal-MOS-transistoren," *AEÜ*, vol. 41, pp. 13–20, 1987.
75. Y.-T. Yeow, "Measurement and numerical modeling of short-channel MOSFET gate capacitances," *IEEE Transactions on Electron Devices*, vol. ED-35, pp. 2510–2520, 1987.
76. H. Masuda, Y. Aoki, J. Mano, and O. Yamashiro, "MOSTSM: A physically based charge conservative MOSFET model," *IEEE Transactions on Computer-Aided Design*, vol. 7, pp. 1229–1236, 1988.
77. B.J. Sheu and P.-K. Ko, "Measurement and modeling of short-channel MOS transistor gate capacitances," *IEEE Journal of Solid-State Circuits*, vol. SC-22, pp. 464–472, 1987.

78. B.J. Sheu, W.-J. Hsu, and P. K. Ko, "An MOS transistor charge model for VLSI design," *IEEE Transactions on Computer-Aided Design*, vol. 7, pp. 520–527, 1988.
79. R. Gharabagi and M. El-Nokali, "A model for the intrinsic gate capacitances of short channel MOSFETs," *Solid-State Electronics*, vol. 32, pp. 57–63, 1989.
80. K. A. Sakallah, Y.-T. Yen, and S. S. Greenberg, "A first-order charge conserving MOS capacitance model," *IEEE Transactions on Computer-Aided Design*, vol. 9, pp. 99–108, 1990.
81. W. Budde and W. H. Lamfried, "A charge-sheet capacitance model based on drain current modeling," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1678–1687, 1990.
82. R. Gharabagi and M. El-Nokali, "An analytical model for the capacitances in short-channel MOSFETs," *Solid-State Electronics*, vol. 33, pp. 235–241, 1990.
83. C. S. Oh, W. H. Chang, B. Davari, and Y. Taur, "Voltage dependence of the MOSFET gate-to-source/drain overlap," *Solid-State Electronics*, vol. 33, pp. 1650–1652, 1990.
84. R. Gharabagi and M. A. El-Nokali, "A charge-based model for short-channel MOS transistor capacitances," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1064–1073, 1990.
85. H.-J. Park, P. K. Ko, and C. Hu, "A charge sheet capacitance model of short channel MOSFETs for SPICE," *IEEE Transactions on Computer-Aided Design*, vol. 10, pp. 376–389, 1991.
86. A. Afzali-Kushaa and M. El-Nokali, "Modeling subthreshold capacitances of MOS transistors," *Solid-State Electronics*, vol. 35, pp. 45–49, 1992.
87. P. Klein, K. Hoffmann, and B. Lemaitre, "Description of the bias dependent overlap capacitance at LDD MOSFETs for circuit applications," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 493–496, December 1993.
88. K.-M. Rho, K. Lee, M. Shur, and T. A. Fjeldly, "Unified quasi-static MOSFET capacitance model," *IEEE Transactions on Electron Devices*, vol. 40, pp. 131–135, 1993.
89. T. Smedes and F. M. Klaassen, "Influence of channel series resistances on dynamic MOSFET behaviour," *Solid-State Electronics*, vol. 37, pp. 251–254, 1994.
90. J. Katzenelson and A. Unikovski, "A network charge-oriented MOS transistor model," *International Journal of High Speed Electronics and Systems*, vol. 6, pp. 285–316, 1995.
91. D.-H. Cho, S.-M. Kang, K.-H. Kim, and S.-H. Lee, "An accurate intrinsic capacitance modeling for deep submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. 42, pp. 540–548, 1995.
92. V. I. Kol'dyaev, A. Clerix, R. M. Arteaga, and L. Deferm, "Characterisation of the overlap capacitance of submicron LDD MOSFETs," *Proceedings of the 25th European Solid State Device Research Conference*, The Netherlands, September 1995, pp. 757–760.
93. B. Riccò, R. Versari, and D. Esseni, "A novel method to characterize parasitic capacitances in MOSFETs," *IEEE Electron Device Letters*, vol. 16, pp. 485–487, 1995.
94. T. Manku, "A methodology for measuring the gate-drain capacitance of CMOS devices," *IEEE Electron Device Letters*, vol. 17, pp. 312–314, 1996.
95. C. H. Wang, "Identification and measurement of scaling-dependent parasitic capacitance of small-geometry MOSFETs," *IEEE Transactions on Electron Devices*, vol. 43, p. 965, 1996.
96. R. Shrivastava and K. Fitzpatrick, "A simple model for the overlap capacitance of a VLSI MOS device," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 1870–1875, 1982.
97. K. S. Krisch, J. D. Bude, and L. Manchanda, "Gate capacitance attenuation in MOS devices with thin gate dielectrics," *IEEE Electron Device Letters*, vol. 17, pp. 521–524, 1996.
98. P. Klein, "A compact charge LDD MOSFET model," *IEEE Transactions on Electron Devices*, vol. 44, pp. 1483–1490, 1997.
99. www.oup.com/us/tsividis_mcandrew
100. F. Svelto, P. Erratico, S. Manzini, and R. Castello, "A metal-oxide-semiconductor varactor," *IEEE Electron Device Letters*, vol. 20, no. 3, pp. 164–166, 1999.
101. S. Pavan, Y. Tsividis, and K. Nagaraj, "Modeling of accumulation MOS capacitors for analog design in digital VLSI processes," *Proceedings of the IEEE International Conference on Circuits and Systems*, pp. 202–205, May 1999.
102. A. Porret, T. Melly, C. C. Enz, and E. A. Vittoz, "Design of high-Q varactors for low-power wireless applications using a standard CMOS process," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 3, pp. 337–345, 2000.

103. P. Andreani and S. Mattisson, "On the use of MOS varactors in RF VCOs," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 6, pp. 905–910, 2000.
104. A. van der Ziel, *Noise in Solid State Devices and Circuits*, Wiley-Interscience, New York, 1986.
105. A. Ambrozy, *Electronic Noise*, McGraw-Hill, New York, 1982.
106. P. R. Gray P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 4th ed., Wiley, New York, 2001.
107. C. A. Desoer and E. S. Kuh, *Basic Circuit Theory*, McGraw-Hill, New York, 1969.
108. A. van der Ziel, *Noise: Sources, Characterization, and Measurement*, Prentice Hall, Englewood Cliffs, N.J., 1970.
109. C. Galup-Montoro and M. C. Schneider, *MOSFET Modeling for Circuit Analysis and Design*, World Scientific, Singapore, 2007.
110. C. T. Sah, "Theory and experiments on the $1/f$ surface noise of MOS insulated-gate field-effect transistors," *IEEE Transactions on Electron Devices* (Abstract), vol. ED-11, p. 534, 1964.
111. A. G. Jordan and N. A. Jordan, "Theory of noise in metal oxide semiconductor devices," *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 148–156, 1965.
112. H. Johnson, "Noise in field-effect transistors," chap. 6, *Field-Effect Transistors*, T. Wallmark and H. Johnson (editors), Prentice Hall, Englewood Cliffs, N.J., 1966.
113. S. M. Bozic, "Noise in the metal oxide semiconductor transistor," *Electronic Engineering*, vol. 38, pp. 40–41, 1966.
114. C. T. Sah, S. Y. Wu, and F. H. Hielscher, "The effects of fixed bulk charge on the thermal noise in metal-oxide-semiconductor transistors," *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 410–414, 1966.
115. M. Shoji, "Analysis of high-frequency thermal noise of enhancement mode M.O.S. field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 520–524, 1966.
116. I. Flinn, G. Bew, and F. Berz, "Low frequency noise in M.O.S. field effect transistors," *Solid-State Electronics*, vol. 10, pp. 833–845, 1967.
117. J. Mavor, "Noise parameters for metal-oxide-semiconductor transistors," *IEE Proceedings*, vol. 113, pp. 1463–1467, 1967.
118. F. M. Klaassen and J. Prins, "Thermal noise of M.O.S. transistors," *Philips Research Reports*, vol. 22, pp. 505–514, 1967.
119. R. Paul, "Thermisches Rauschen von MOS-Transistoroen," *Nachrichtentechnik*, vol. 17, pp. 458–466, 1967.
120. S. Y. Wu, "Theory of the generation-recombination noise in MOS transistors," *Solid-State Electronics*, vol. 11, pp. 25–32, 1968.
121. A. Leupp and M. J. O. Strutt, "Noise behavior of the MOSFET at VHF and UHF," *Electronics Letters*, vol. 4, pp. 313–314, 1968.
122. H. E. Halladay and A. van der Ziel, "Test of the thermal noise hypothesis in MOSFETS," *Electronics Letters*, vol. 4, pp. 366–367, 1968.
123. L. D. Yau and C. T. Sah, "Theory and experiments of low-frequency generation recombination noise in MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 170–177, 1969.
124. F. M. Klaassen and J. Prins, "Noise of field-effect transistors at very high frequencies," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 952–957, 1969.
125. M. Nakahara, "Anomalous low-frequency noise enhancement in silicon MOS transistors," *Proceedings of the IEEE*, vol. 57, pp. 2177–2178, 1969.
126. L. D. Yau and C. T. Sah, "On the excess 'white noise' in MOS transistors," *Solid-State Electronics*, vol. 12, pp. 927–936, 1969.
127. J. W. Haslett and F. N. Trofimenkoff, "Gate noise in MOSFET's at moderately high frequencies," *Solid-State Electronics*, vol. 14, pp. 239–245, 1971.
128. A. van der Ziel, "Noise resistance of FETs in the hot electron regime," *Solid-State Electronics*, vol. 14, pp. 347–350, 1971.
129. P. S. Rao and A. van der Ziel, "Noise and y parameters in MOSFETs," *Solid-State Electronics*, vol. 14, pp. 939–944, 1971.
130. E. W. Kirk, "Induced gate noise in MOSFETs," *Solid-State Electronics*, vol. 14, pp. 945–948, 1971.

131. M. B. Das and J. M. Moore, "Measurements and interpretation of low frequency noise in FETS," *IEEE Transactions on Electron Devices*, vol. ED-21, pp. 247–257, 1974.
132. S. T. Hsu and A. van der Ziel, "Thermal noise in ion-implanted MOSFETs," *Solid-State Electronics*, vol. 18, pp. 509–510, 1975.
133. N. Nakamura, O. Kudoh, and M. Kamoshida, "Noise characteristics of ion-implanted MOS transistors," *Journal of Applied Physics*, vol. 46, pp. 3189–3193, 1975.
134. W. Fichtner, E. Hochmair, and D. Kranzer, "Noise measurements on SOSMOS transistors," *Proceedings of the European Solid-State Device Research Conference*, Munich, September 1976.
135. W. Fichtner and E. Hochmair, "Current-kink noise or n-channel enhancement ESFI MOS SOS transistors," *Electronics Letters*, vol. 13, pp. 675–676, 1977.
136. P. Gentil and S. Chausse, "Low-frequency measurement on silicon-on-sapphire (SOS) MOS transistors," *Solid-State Electronics*, vol. 20, pp. 935–940, 1977.
137. K. Takagi and K. Matsumoto, "Noise in silicon and FETs at high electric fields," *Solid-State Electronics*, vol. 20, pp. 1–3, 1977.
138. W. A. Baril, "High-frequency thermal noise in MOSFETs," *Solid-State Electronics*, vol. 21, pp. 589–592, 1978.
139. J. Fellrath, "Shot noise behavior of subthreshold MOS transistors," *Revue de Physique Appliquée*, vol. 13, pp. 719–723, 1978.
140. K. Takagi and A. van der Ziel, "Drain noise in MOSFETs at zero drain bias as a function of temperature," *Solid-State Electronics*, vol. 22, pp. 87–88, 1979.
141. K. Takagi and A. van der Ziel, "Excess high frequency noise and flicker noise in MOSFETs," *Solid-State Electronics*, vol. 22, pp. 289–292, 1979.
142. S. T. Liu and A. van der Ziel, "High-frequency noise in weakly inverted metal-oxide-semiconductor transistors," *Applied Physics Letters*, vol. 37, pp. 950–951, 1980.
143. R. P. Jindal and A. van der Ziel, "Effect of transverse electric field on Nyquist noise," *Solid-State Electronics*, vol. 24, pp. 905–906, 1981.
144. G. Reimbold and P. Gentil, "White noise of MOS transistors operating in weak inversion," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 1722–1725, 1982.
145. H. S. Park and A. van der Ziel, "Noise measurements in ion implanted MOSFETs," *Solid-State Electronics*, vol. 26, pp. 747–751, 1983.
146. A. van der Ziel, R. J. J. Zijstra, H. S. Park, and S. T. Liu, "Alternate explanation of 1/f noise in ion-implanted MOSFETs," *Solid-State Electronics*, vol. 26, pp. 927–928, 1983.
147. S. A. Hayat and B. K. Jones, "Thermal noise in inversion layers," *Solid-State Electronics*, vol. 27, pp. 687–688, 1984.
148. R. P. Jindal, "Noise associated with distributed resistance of MOSFET gate structures in integrated circuits," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1505–1509, 1984.
149. R. P. Jindal, "Distributed substrate resistance noise in fine-line NMOS field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2450–2453, 1985.
150. R. P. Jindal, "High frequency noise in fine line NMOS field effect transistors," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1985, pp. 68–71.
151. E. A. Vittoz, private communication, 1985.
152. S. A. Hayat and B. K. Jones, "Thermal noise in inversion layers," *Solid-State Electronics*, vol. 27, pp. 687–688, 1984.
153. E. N. Wu and A. van der Ziel, "On the influence of substrate doping on the input conductance and the induced gate noise in MOSFETs," *Solid-State Electronics*, vol. 27, pp. 945–946, 1984.
154. A. A. Abidi, "High-frequency noise measurements on FETs with small dimensions," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 1801–1805, 1986.
155. K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A physics-based MOSFET noise model for circuit simulators," *IEEE Transactions on Electron Devices*, vol. 37, pp. 1323–1333, 1990.
156. L.-J. Pu and Y. Tsividis, "Small-signal parameters and thermal noise of the four-terminal MOSFET in non-quasi-static operation," *Solid-State Electronics*, vol. 33, pp. 513–521, 1990.
157. L. Goldminz and Y. Nemirovsky, "Thermal noise in buried-channel MOSFET," *IEEE Transactions on Electron Devices*, vol. 39, pp. 2315–2322, 1992.
158. R. Sarapeshkar, T. Delbrück, and C. A. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits and Devices Magazine*, pp. 23–30, 1993.

159. S. Tedja, J. Van der Spiegel, and H. H. Williams, "Analytical and experimental studies of thermal noise in MOSFETs," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2069–2075, 1994.
160. B. Wang, J. R. Hellums, and C. G. Sodini, "MOSFET thermal noise modeling for analog integrated circuits," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 833–835, 1994.
161. F. Danneville, H. Happy, G. Dambrine, J.-M. Belquin, and A. Cappy, "Microscopic noise modeling and macroscopic noise models: How good a connection?" *IEEE Transactions on Electron Devices*, vol. 41, pp. 779–786, 1994.
162. E. P. Vandamme, L. K. Vandamme, C. Claeys, E. Simoen, and R. J. Schreutelkamp, "Impact of silicidation on the excess noise behaviour of MOS transistors," *Solid-State Electronics*, vol. 38, pp. 1893–1897, 1995.
163. G. Abowitz, E. Arnold, and E. Leventhal, "Surface states and $1/f$ noise in M.O.S. transistors," *IEEE Transactions on Electron Devices*, vol. ED-14, pp. 775–777, 1967.
164. I. R. M. Mansour, R. J. Hawkins, and G. G. Bloodworth, "Measurement of current noise in MOS transistors from $5 \cdot 10^{-5}$ to 1 Hz," *The Radio and Electronic Engineer*, vol. 35, pp. 212–216, 1968.
165. E. A. Leventhal, "Derivation of $1/f$ noise in silicon inversion layers from carrier motion in a surface band," *Solid-State Electronics*, vol. 11, pp. 621–627, 1968.
166. S. Christensson, I. Lundstrom, and C. Svensson, "Low frequency noise in MOS transistors—I. Theory," *Solid-State Electronics*, vol. 11, pp. 796–812, 1968; "Low frequency noise in MOS transistors—II. Experiments," *Solid State Electronics*, vol. II, pp. 813–820, 1968.
167. I. R. M. Mansour, R. J. Hawkins, and G. G. Bloodworth, "Physical model for the current noise spectrum of MOSTS," *British Journal of Applied Physics (Journal of Physics D: Applied Physics)* vol. 2, pp. 1063–1082, 1969.
168. H. E. Halladay and A. van der Ziel, "On the high frequency excess noise and equivalent circuit representation of the MOSFET with *n*-type channel," *Solid-State Electronics*, vol. 12, pp. 161–176, 1969.
169. P. S. Rao, "The effect of the substrate upon the gate and drain noise parameters of MOSFETs," *Solid-State Electronics*, vol. 12, pp. 549–556, 1969.
170. L. D. Yau and A. van der Ziel, "Geometrical dependencies of the low-frequency generation-recombination noise in MOS transistors," *Solid-State Electronics*, vol. 12, pp. 903–905, 1969.
171. A. Takagi and A. van der Ziel, "Non-thermal noise in MOSFETs and MOS tetrodes," *Solid State Electronics*, vol. 12, p. 907, 1969.
172. N. R. Mantena and R. C. Lucas, "Experimental study of flicker noise in MIS field effect transistors," *Electronics Letters*, vol. 5, pp. 607–608, 1969.
173. F. Berz, "Theory of low frequency noise in Si MOSTs," *Solid-State Electronics*, vol. 13, pp. 631–647, 1970.
174. S. T. Hsu, "Surface state related $1/f$ noise in MOS transistors," *Solid-State Electronics*, vol. 13, pp. 1451–1459, 1970.
175. J. W. Hawkins and G. G. Bloodworth, "Two components of $1/f$ noise in MOS transistors," *Solid-State Electronics*, vol. 14, pp. 932–939, 1971.
176. F. M. Klaassen, "Characterization of low $1/f$ noise in MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-18, pp. 887–891, 1971.
177. J. W. Haslett and F. N. Trofimenkoff, "Effects of the substrate on surface state noise in silicon MOSFETs," *Solid-State Electronics*, vol. 15, pp. 117–131, 1972.
178. H. S. Fu and C. T. Sah, "Theory and experiments on surface $1/f$ noise," *IEEE Transactions on Electron Devices*, vol. ED-19, pp. 273–285, 1972.
179. R. S. Ronen, "Low-frequency $1/f$ noise in MOSFETs," *RCA Review*, vol. 34, pp. 280–307, 1973.
180. S. T. Hsu and A. van der Ziel, "A new type of flicker noise in microwave MOSFETs," *Solid-State Electronics*, vol. 18, pp. 885–886, 1975.
181. H. Katto, Y. Kamigaki, and Y. Itoh, "MOSFETs with reduced low frequency $1/f$ noise," *Japanese Journal of Applied Physics*, vol. 44, pp. 243–248, 1975.
182. S. T. Hsu, "Trapping noise in SOSMOSFs," *Proceedings of the Fourth International Conference on Physical Aspects of Noise in Solid State Devices*, Noordwijkerhout, The Netherlands, 1975.
183. A. van der Ziel, "Limiting flicker noise in MOSFETs," *Solid-State Electronics*, vol. 18, p. 1031, 1975.
184. F. N. Hooge, " $1/f$ noise," *Physica*, vol. 83B, pp. 14–23, 1976.

185. A. van der Ziel, "Dependence of flicker noise in MOSFETs on geometry," *Solid-State Electronics*, vol. 20, p. 267, 1977.
186. K. L. Wang, "Measurements of residual defects and $1/f$ noise in ion-implanted p-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 478–484, 1978.
187. A. van der Ziel, "Some general relationships for flicker noise in MOSFETs," *Solid-State Electronics*, vol. 21, pp. 623–624, 1978.
188. R. P. Jindal and A. van der Ziel, "Carrier fluctuations noise in a MOSFET channel due to traps in the oxide," *Solid-State Electronics*, vol. 21, pp. 901–903, 1978.
189. P. Gentil, "Bruit basse fréquence du transistor MOS—1ère partie," *L'onde électrique*, vol. 58, pp. 565–575, 1978; 2nd part, *L'onde électrique*, vol. 58, pp. 645–652, 1978.
190. P. Victorovich and P. Gentil, "Influence of the depth of interface states in the insulator on the noise properties of MOS transistors," *Solid-State Electronics*, vol. 22, pp. 21–23, 1979.
191. W. V. Backenstos and C. R. Viswanathan, "Bias-dependent $1/f$ noise model of an m.o.s. transistor," *IEE Proceedings*, vol. 127, part I, pp. 87–93, 1980.
192. L. K. J. Vandamme and H. M. M. de Werd, " $1/f$ noise model for MOSTs biased in nonohmic region," *Solid-State Electronics*, vol. 23, pp. 325–329, 1980.
193. L. K. J. Vandamme, "Model for $1/f$ noise in MOS transistors biased in the linear region," *Solid-State Electronics*, vol. 23, pp. 317–323, 1980.
194. P. Gentil and A. Mounib, "Equivalent input spectrum and drain current spectrum for $1/f$ noise in short channel MOS transistors," *Solid-State Electronics*, vol. 24, pp. 411–414, 1981.
195. H. S. Park, A. van der Ziel, and S. T. Liu, "Comparison of two $1/f$ noise models in MOSFETs," *Solid-State Electronics*, vol. 23, pp. 213–217, 1982.
196. H. Mikoshiba, " $1/f$ noise in n-channel silicon-gate MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 965–970, 1982.
197. G. Reimbold, "Modified $1/f$ trapping noise theory and experiments in MOS transistors biased from weak to strong inversion—Influence of interface states," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1190–1198, 1984.
198. J. M. Pimbley and G. Gildenblat, "Effect of hot-electron stress on low frequency MOSFET noise," *IEEE Electron Device Letters*, vol. EDL-5, pp. 345–347, 1984.
199. Z. Celik and T. Y. Hsiang, "Study of $1/f$ noise in N-MOSFETs: linear region," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2797–2802, 1985.
200. A. van der Ziel, "Integral expression for $1/f$ noise in MOSFETs at arbitrary drain bias," *Solid-State Electronics*, vol. 29, pp. 29–30, 1986.
201. T. Watanabe, "Low-noise operation in buried-channel MOSFETs," *IEEE Electron Device Letters*, vol. EDL-6, pp. 317–319, 1985.
202. A. van der Ziel, "Reconciliation of Klaassen's and Reimbold's theories of $1/f$ noise in MOSFETs," *Solid-State Electronics*, vol. 29, pp. 967–968, 1986.
203. B. Pellegrini, "On mobility-fluctuation origin on $1/f$ noise," *Solid-State Electronics*, vol. 29, pp. 1279–1287, 1986.
204. Z. Celik-Butler and T. Y. Hsiang, "Spectral dependence of $1/f$ noise on gate bias in N-MOSFETs," *Solid-State Electronics*, vol. 30, pp. 419–423, 1987.
205. G. Ghibaudo, "A simple derivation of Reimbold's drain current spectrum formula for flicker noise in MOSFETs," *Solid-State Electronics*, vol. 30, pp. 1037–1038, 1987.
206. N. Mutoh and N. Teranishi, "New empirical relation for MOSFET $1/f$ noise unified over linear and saturation regions," *Solid-State Electronics*, vol. 31, pp. 1675–1680, 1988.
207. C. Surya and T. Y. Hsiang, "Surface mobility fluctuations in metal-oxide-semiconductor field-effect transistors," *Physical Review B*, vol. 35, p. 6342, 1987.
208. F. Grabowski, "Influence of dynamical interactions between density and mobility of carriers in the channel on $1/f$ noise of MOS transistors below saturation—I. Mechanisms," *Solid-State Electronics*, vol. 32, pp. 909–913, 1989.
209. F. Grabowski, "Influence of dynamical interactions between density and mobility of carriers in the channel of $1/f$ noise of MOS transistors below saturation—II. Implications," *Solid-State Electronics*, vol. 32, pp. 915–918, 1989.
210. G. Ghibaudo, "On the theory of carrier number fluctuations in MOS devices," *Solid-State Electronics*, vol. 32, pp. 563–565, 1989.

211. K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *IEEE Transactions on Electron Devices*, vol. 37, pp. 654–665, 1990.
212. T. G. M. Kleinpenning, "On $1/f$ trapping noise in MOSTs," *IEEE Transactions on Electron Devices*, vol. 37, pp. 2084–2089, 1990.
213. Z.-H. Fang, A. Chovet, Q.-P. Zhu, and J.-N. Zhao, "Theory and applications of $1/f$ trapping noise in MOSFETs for the whole biasing ranges," *Solid-State Electronics*, vol. 34, pp. 327–333, 1991.
214. S.-L. Jang, "Analytical low-frequency $1/f$ noise model for lightly-doped-drain MOSFETs operating in the linear region," *Solid-State Electronics*, vol. 36, pp. 899–903, 1993.
215. S.-L. Jang and P.-C. Chang, "Low-frequency noise characteristics of lightly doped-drain MOSFETs," *Solid-State Electronics*, vol. 36, pp. 1007–1010, 1993.
216. L. K. J. Vandamme, X. Li, and D. Rigaud, " $1/f$ noise in MOS devices, mobility or number fluctuations?" *IEEE Transactions on Electron Devices*, vol. 41, pp. 1936–1945, 1994.
217. J. H. Scofield, N. Borland, and D. M. Fleetwood, "Reconciliation of different gate-voltage dependencies of $1/f$ noise in n-MOS and p-MOS transistors," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1946–1952, 1994.
218. D. M. Fleetwood, T. L. Meisenheimer, and J. H. Scofield, " $1/f$ noise and radiation effects in MOS devices," *IEEE Transactions on Electron Devices*, vol. 41, p. 1953, 1994.
219. J. Chang, A. A. Abidi, and C. R. Viswanathan, "Flicker noise in CMOS transistors from subthreshold to strong inversion at various temperatures," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1965–1971, 1994.
220. D. R. Wolters, A. T. A. Zegers-van Duijnhoven, and R. Augur, "Model for $1/f$ noise in MOSFETs and interconnects," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 177–180, 1994.
221. S.-L. Jang, Comments on "Analytical low-frequency $1/f$ noise model for lightly-doped-drain MOSFETs operating in the linear region," *Solid-State Electronics*, vol. 37, pp. 1903–1904, 1994.
222. X. Li, C. Barros, E. P. Vandamme, and K. L. Vandamme, "Parameter extraction and $1/f$ noise in a surface and a bulk-type, p-channel LDD MOSFET," *Solid-State Electronics*, vol. 37, pp. 1853–1862, 1994.
223. C. Hu, J. Zhao, G. P. Li, P. Liu, E. Worley, J. White, and R. Kjar, "The effects of plasma etching induced gate oxide degradation on MOSFET's $1/f$ noise," *IEEE Electron Device Letters*, vol. 16, pp. 61–63, 1995.
224. C. Hu, G. P. Li, E. Worley, and J. White, "Consideration of low-frequency noise in MOSFETs for analog performance," *IEEE Electron Device Letters*, vol. 17, pp. 552–554, 1996.
225. P. Morfouli, G. Ghibaudo, T. Ouisse, E. Vogel, W. Hill, V. Misra, P. McLarty, and J. J. Wortman, "Low-frequency noise characterization of n- and p-MOSFET's with ultrathin oxynitride gate films," *IEEE Electron Device Letters*, vol. 17, pp. 395–397, 1996.
226. M. Stegherr, "Flicker noise in hot electron degraded short channel MOSFETs," *Solid-State Electronics*, vol. 27, pp. 1055–1056, 1984.
227. R. P. Jindal, "Noise phenomena in submicron channel length nMOS transistors," *Noise in Physical Systems and 1/f Noise*, Elsevier Science, Amsterdam, The Netherlands, 1986.
228. Z. H. Fang, S. Cristoloveanu, and A. Chovet, "Analysis of hot-carrier-induced aging from $1/f$ noise in short-channel MOSFETs," *IEEE Electron Device Letters*, vol. EDL-7, pp. 371–373, 1986.
229. B. Boukriss, H. Haddara, S. Cristoloveanu, and A. Chovet, "Modeling of the $1/f$ noise overshoot in short-channel MOSFETs locally degraded by hot-carrier injection," *IEEE Electron Device Letters*, vol. 10, pp. 433–436, 1989.
230. A. Comeau, Comments on "Modeling of the $1/f$ noise overshoot in short-channel MOSFET's locally degraded by hot-carrier injection," *IEEE Electron Device Letters*, vol. 11, p. 129, 1990.
231. M.-H. Tsai and T.-P. Ma, " $1/f$ noise in hot-carrier damaged MOSFET's: Effects of oxide charge and interface traps," *IEEE Electron Device Letters*, vol. 14, pp. 256–258, 1993.
232. C.-H. Cheng and C. Surya, "The effect of hot-electron injection on the properties of flicker noise in N-channel MOSFETs," *Solid-State Electronics*, vol. 36, pp. 475–479, 1993.
233. D. Rigaud and M. Valenza, "Correlation measurement of carrier multiplication noise sources in MOS transistors at low frequencies," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2076–2081, 1994.

234. D. P. Triantis, A. N. Birbas, and D. Kondis, "Thermal noise modeling for short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 43, pp. 1950–1955, 1996.
235. M. Aoki and M. Kato, "Hole-induced 1/f noise increase in MOS transistors," *IEEE Electron Device Letters*, vol. 17, pp. 118–120, 1996.
236. D. P. Triantis, A. N. Birbas, and S. E. Plevridis, "Induced gate noise in MOSFET revisited: The submicron case," *Solid-State Electronics*, vol. 41, pp. 1937–1942, 1997.
- 236a. M. Miura-Mattausch, H. J. Mattausch, and T. Ezaki, *The Physics and Modeling of MOSFETs*, World Scientific, Singapore, 2008.
- 236b. C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling*, Wiley, Hoboken, N.J., 2006.
- 236c. D. M. Binkley, *Tradeoffs and Optimization in Analog Circuit Design*, Wiley, Hoboken, N.J., 2008.
- 236d. G. Knoblinger, P. Klein, and M. Tiebout, "A new model for thermal channel noise of deep-submicron MOSFETs and its application in RF-CMOS design," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 5, pp. 831–837, 2001.
237. A. J. Scholten, L. F. Tiemeijer, R. van Langevelde, R. J. Havens, A. T. A. Zegers-van Duijnhoven, and V. C. Venezia, "Noise modeling for RF CMOS circuit simulation," *IEEE Transactions on Electron Devices*, vol. 50, no. 3, pp. 618–632, 2003.
238. T.-L. Chen and G. Gildenblat, "Symmetric bulk charge linearisation in charge-sheet MOSFET model," *Electronics Letters*, vol. 37, no. 12, pp. 791–793, 2001.
239. A. J. Scholten, G. D. J. Smit, B. A. De Vries, L. F. Tiemeijer, J. A. Croon, D. B. M. Klaassen, R. van Langevelde, X. Li, W. Wu, and G. Gildenblat, "The new CMC standard compact MOS model PSP: Advantages for RF applications," *Proceedings of the IEEE Radio Frequency Integrated Circuits Symposium*, pp. 247–250, April 2008.
240. M. J. Deen, C.-H. Chen, and Y. Cheng, "MOSFET modeling for low noise, RF circuit design," *Proceedings of the Custom Integrated Circuits Conference*, pp. 201–208, May 2002.
241. A. S. Roy and C. C. Enz, "Compact modeling of thermal noise in the MOS transistor," *IEEE Transactions on Electron Devices*, vol. 52, no. 4, pp. 611–614, 2005.
242. R. Brederlow, G. Wenig, and R. Thewes, "Investigation of the thermal noise of MOS transistors under analog and RF operating conditions," *Proceedings of the European Solid-State Device Research Conference*, pp. 87–90, September 2002.
243. A. J. Scholten, L. F. Tiemeijer, R. van Langevelde, R. J. Havens, A. T. A. Zegers-van Duijnhoven, R. de Kort, and D. B. M. Klaassen, "Compact modelling of noise for RF CMOS circuit design," *IEE Proceedings—Circuits, Devices and Systems*, vol. 151, no. 2, pp. 167–174, 2004.
244. R. P. Jindal, "Compact noise models for MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-53, pp. 2051–2061, 2006.
245. A. van der Ziel and E. R. Chenette, "Noise in solid state devices," *Advances in Electronics and Electron Physics*, vol. 44, pp. 313–383, 1978.
246. R. P. Jindal, "Noise associated with substrate current and fine-line NMOS field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 6, pp. 1047–1052, 1985.
247. F. M. Klaassen, "On the substrate current noise in MOS transistors beyond pinchoff," *Proceedings of the IEEE*, vol. 59, no. 2, pp. 331–332, 1971.
248. M. Valenza, A. Laigle, F. Martinez, A. Hoffman, and D. Rigaud, "Impact of gate current noise on drain current noise in 90-nm CMOS technology," *Proceedings of the European Solid-State Device Research Conference*, pp. 287–290, September 2003.
- 248a. G. B. Alers, K. S. Krish, D. Monroe, E. E. Weir, and A. M. Chang, "Tunneling current noise in thin gate oxides," *Applied Physics Letters*, vol. 69, pp. 2885–2887, 1996.
249. E. P. Vandamme and L. K. J. Vandamme, "Critical discussion on unified 1/f noise models for MOSFETs," *IEEE Transactions on Electron Devices*, vol. 47, no. 11, pp. 2146–2152, 2000.
250. J. P. Campbell, J. Qin, K. P. Cheung, L. Yu, J. S. Suehle, A. Oates, and K. Sheng, "The origins of random telegraph noise in highly scaled SiON nMOSFETs," *IEEE 2008 International Integrated Reliability Workshop Final Report*, South Lake Tahoe, 2008, pp. 1–16.
251. A. L. McWorther, "1/f noise and related surface effects in germanium," *Massachusetts Institute of Technology Lincoln Laboratory Report 80*, May 1955.
252. A. L. McWorther, "1/f noise and germanium surface properties," *Semiconductor Surface Physics*, R. H. Kingston (editor), pp. 207–228, University of Pennsylvania Press, Philadelphia, 1957.

253. F. N. Hooge and L. K. Vandamme, "Lattice scattering causes $1/f$ noise," *Physics Letters*, vol. 66A, pp. 315–316, 1978.
254. H. Mikoshiba, " $1/f$ noise in n-channel silicon-gate MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 6, pp. 965–970, June 1982.
255. R. Jayaram and C. Sodini, "A $1/f$ noise technique to extract the oxide trap density near the conduction band edge of silicon," *IEEE Transactions on Electron Devices*, vol. 36, no. 9, pp. 1773–1782, 1989.
256. R. Brederlow, W. Weber, D. Schmitt-Landsiedel, and R. Thewes, "Fluctuations of the low frequency noise of MOS transistors and their modeling in analog and RF-circuits," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 159–162, December 1999.
257. G. I. Wirth, J. Koh, R. da Silva, R. Thewes, and R. Brederlow, "Modeling of statistical low-frequency noise of deep-submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. 52, no. 7, pp. 1576–1588, 2005.
258. M. Erturk, T. Xia, R. Anna, K. M. Newton, and E. Adler, "Statistical BSIM model for MOSFET $1/f$ noise," *Electronics Letters*, vol. 41, no. 22, pp. 1208–1209, 2005.
259. K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, "Discrete resistance switching in submicrometer silicon inversion layers: Individual interface traps and low-frequency ($1/f$) noise," *Physical Review Letters*, vol. 52, p. 228, 1984.
260. M. J. Uren, D. J. Day, and M. J. Kirton, " $1/f$ and random telegraph noise in silicon metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, vol. 47, p. 1195, 1985.
261. K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "Random telegraph noise of deep-submicrometer MOSFETs," *IEEE Electron Device Letters*, vol. 11, pp. 90–92, 1990.
262. P. Fang, K. K. Hung, P. K. Ko, and C. Hu, "Hot-electron-induced traps studied through the random telegraph noise," *IEEE Electron Device Letters*, vol. 12, pp. 273–275, 1991.
263. O. R. D. Buisson, G. Ghibaudo, and J. Brini, "Model for drain current RTS amplitude in small-area MOS transistors," *Solid-State Electronics*, vol. 35, pp. 1273–1276, 1992.
264. E. Simoen, B. Dierickx, C. L. Claeys, and G. J. Declerck, "Explaining the amplitude of RTS noise in submicrometer MOSFETs," *IEEE Transactions on Electron Devices*, vol. 39, pp. 422–429, 1992.
265. M.-H. Tsai and T.-P. Ma, "The impact of device scaling on the current fluctuations in MOSFETs," *IEEE Transaction on Electron Devices*, vol. 41, pp. 2061–2068, 1994.
266. Z. Shi, J.-P. Miéville, and M. Dutoit, "Random telegraph signals in deep submicron n-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1161–1168, 1994.
267. K. K. Thornber, "Resistive gate induced thermal noise in IGFETs," *IEEE Journal of Solid-State Circuits*, vol. SC-16, no. 4, pp. 414–415, 1981.
268. R. Rohrer, L. Nagel, R. Meyer, and L. Weber, "Computationally efficient electronic-circuit noise calculations," *IEEE Journal of Solid-State Circuits*, vol. SC-6, no. 4, pp. 204–213, 1971.
269. Y. Tsividis, K. Suyama, and K. Vavelidis, "A simple 'Reconciliation' MOSFET model valid in all regions," *Electronics Letters*, vol. 31, pp. 506–508, 1995.
270. C.C. McAndrew, G. Coram, A. Blaum, and O. Pillard, "Correlated noise modeling and simulation," *Proceedings of the NSTI Workshop on Compact Modeling*, pp. 40–45, May 2005.
271. J. W. Wu, C. C. Ceng, K. L. Chiu, J. C. Guo, W. Y. Lien, C. S. Chang, G. W. Huang, and T. Wang, "Pocket implantation effect on drain current flicker wise in analog nMOSFET devices," *IEEE Transactions on Electron Devices*, vol. 51, no. 8, pp. 1262–1266, 2004.

Problems

- 7.1.** Discuss what types of I - V plots you would need in order to determine g_m , g_{mb} , and g_{sd} graphically, and describe the procedure for determining these parameters. Discuss qualitatively the accuracy you would expect from such a procedure.
- 7.2.** Prove (7.2.12).
- 7.3.** For an n -channel device in strong inversion with $V_{T0} = 0.4$ V, $V_{SB} = 0$ V, $(W/L)\mu C'_ox = 30$ $\mu\text{A}/\text{V}^2$, and $\alpha = 1.1$, plot: (a) g_m vs. V_{GS} with V_{DS} as a parameter, (b) g_m vs. V_{DS} with V_{GS} as a parameter.

- 7.4.** (a) Prove (7.2.25)–(7.2.27) starting from the definitions given in Secs. 7.2.2 and 7.2.3. Do this in two ways: (1) mathematically and (2) using the small-signal equivalent circuit of Fig. 7.6. (b) Show that all nine parameters in the model of Fig. 7.6 can be determined from external measurements by exciting the gate, drain, and body terminals one at a time and observing the currents at those three terminals. For example, assume $\Delta V_D = \Delta V_B = \Delta V_S = 0$; apply a nonzero voltage change ΔV_G at the gate, find expressions for the currents ΔI_G , ΔI_D , and ΔI_B from the model, and set them equal to the corresponding externally measured current changes. Repeat this, always for the same currents, by exciting instead only the drain and finally by exciting only the body. This will produce nine equations that can be solved for the nine small-signal parameters in Fig. 7.6. (There is no need to also excite the source; this would result in equations that would be linearly dependent on the above nine and would thus provide no more information.) (c) Determine the simplifications possible in the results in (b) if the quantities related to the gate and body currents are negligible.
- 7.5.** Prove (7.2.28) and (7.2.29).
- 7.6.** Assume that the effective mobility varies with V_{GS} according to (4.11.31), with $\theta_B = 0$. Show how (7.2.28) should be modified to take this effect into account.
- 7.7.** Prove (7.2.31), (7.2.32), and (7.2.34).
- 7.8.** For a given I_{DS} and V_{SB} , state qualitatively what will happen to the ratios g_{mb}/g_m and g_{sd}/g_m if: (a) the substrate doping is increased; (b) the channel length is increased.
- 7.9.** Prove (7.2.38) to (7.2.40).
- 7.10.** Assume operation in saturation, and consider only CLM, modeled by one-dimensional analysis and leading to (5.3.3). Show that the corresponding small-signal source-drain conductance, g_{sd} , is given by $\sqrt{2\epsilon_s/(qN_A)}[I'_{DS}/(2L)](\phi_D + V_{DS} - V'_DS)^{-1/2}$.
- 7.11.** Prove (7.2.42) and (7.2.43).
- 7.12.** Show that an appropriate “charge sharing” model (Sec. 5.4) can predict an increase of g_{sd} in saturation with increasing drain junction depth, other things being equal.
- 7.13.** Prove (7.2.48) and (7.2.49).
- 7.14.** Prove (7.2.50) to (7.2.53).
- 7.15.** Prove (7.2.59) and (7.2.60).
- 7.16.** Using the equivalent small-signal circuit of Fig. 7.19, perform all “measurements” illustrated in Figs. 7.1 and Fig. 7.17. Show that in each case the elements do not interfere with each other; i.e., the results given next to each figure are still obtained.
- 7.17.** Prove (7.3.15).
- 7.18.** Prove (7.3.8), (7.3.9), (7.3.14), (7.3.15) and (7.3.20) using the simplifying assumptions stated before them.
- 7.19.** Show that, for the device of Fig. 7.21, assuming V_D and V_S are almost equal, a very small change ΔV of V_S only will decrease the gate charge by $\frac{1}{2}C_{ox}\Delta V$.
- 7.20.** Derive the results in (7.3.25) to (7.3.29) by differentiating the saturation charge expressions (6.4.26) and (6.4.30), using the simplifying assumptions made in Sec. 7.3 (i.e., that $\alpha = \alpha_t$ and that the variation of α_t with V_S and V_B is negligible).
- 7.21.** Show that (6.4.21) and (6.4.25) cannot be used in the capacitance definitions to derive (7.3.21) and (7.3.22). Explain the reasons for this.
- 7.22.** For the device of Prob. 7.3 operating in strong inversion, plot all capacitances (a) vs. V_{GS} , with V_{DS} as a parameter, and (b) vs. V_{DS} , with V_{GS} as a parameter.

- 7.23.** Prove (7.3.33).
- 7.24.** Prove (7.3.36). [*Hint:* Use (4.8.2), (4.8.6), (4.8.7), and the fact that, at the upper limit of weak inversion, $\psi_s \approx 2\phi_F + V_{SB}$ (Sec. 3.4.1).]
- 7.25.** Show that the models of Fig. 7.30b and c are equivalent.
- 7.26.** Derive (7.5.14), assuming Δv is very small.
- 7.27.** (a) For a transistor in saturation, with $\mu = 450 \text{ cm}^2/(\text{V}\cdot\text{s})$, $t_{ox} = 2 \text{ nm}$, $\alpha = 1.2$, $W = 1 \mu\text{m}$, $L = 0.2 \mu\text{m}$, $V_T = 0.4 \text{ V}$, and $V_{GS} = 0.8 \text{ V}$, evaluate S_{iw} and S_{vw} , at $T = 200$, 300, and 400 K.
(b) Find the mean square value of the thermal noise current for the above device at 300 K, in the band from 100 to 200 kHz.
- 7.28.** The MOS transistor is sometimes modeled by a noiseless transistor with a noise voltage source in series with its gate, representing the equivalent input noise voltage. If capacitances are to be added to this model, on which side of the preceding source should C_g , C_{gd} , and C_{gb} be connected? Show that, for one of these choices, a problem will result at high frequencies if the device is driven by a signal source (between gate and source) with a significant internal resistance.
- 7.29.** Prove (7.5.25) and (7.5.27) as well as the claims in the footnote following (7.5.27).
- 7.30.** A transistor is operating in weak inversion at $T = 300 \text{ K}$, with $I_{DS} = 10 \text{ nA}$, and $n = 1.5$. Plot S_{iw} vs. V_{DS} .
- 7.31.** A device with the process parameters of Prob. 7.27 and with $W = L = 5 \mu\text{m}$, operating at $T = 300 \text{ K}$, exhibits a “corner frequency” f_c (see Fig. 7.31) of 10 kHz. Assuming $c = 1$ in (7.5.31) and $k_1 = 7 \times 10^{-31} \text{ C}^2 \cdot \text{cm}^{-2}$, plot the total power spectral density of the noise current (flicker plus thermal) from 1 Hz to 10 MHz, for $V_{GS} = 0.6$, 0.8, and 1.0 V. Calculate the mean square value of the total current noise (flicker plus thermal) in the band of 20 Hz to 200 kHz.
- 7.32.** Develop an expression for an equivalent input noise resistance representing both thermal and flicker noise in strong inversion. The equivalent noise resistance is a fictitious resistance of such value that its noise power spectral density is equal to that of the equivalent input noise voltage discussed in Prob. 7.28.
- 7.33.** Assume that the power spectral density of noise current S_i , in A^2/Hz , is proportional to $X^a W^b L^c$, where X is some electrical property of a device, such as I_{DS} or g_m , and a , b , and c are constants. Ignoring short-length and narrow-width effects, the noise from a single W/L device should be the same as from two $(W/2)/L$ devices connected in parallel and as from two $W/(L/2)$ devices connected in series. Consider operation at low V_{DS} so that Q'_f can be assumed constant along the channel, so the channel acts like a resistor that is modulated by the gate voltage. Use the equivalent circuit of Fig. 7.30c for each device or section of a device, and consider that a voltage source V_{DS} biases the drain, so it is a short circuit for small signals (noise circuit analysis is ac small-signal analysis). Calculate the noise power spectral density in the drain current for each of the three cases. You need to use the fact that X will halve when the width halves, but will not halve when the length halves (this is clear for I_{DS} ; explain why this is true if X is g_m). Note that you need to treat individual noise sources as being statistically independent. Do the circuit analysis in terms of root mean square quantities (e.g., the square root of current power spectral densities, in $\text{A}/\sqrt{\text{Hz}}$), and then sum the squares of the circuit analysis results (this is trickiest for the series connected devices). Compare the results from the single W/L device and the two $(W/2)/L$ devices connected in parallel to determine a relation between a and b . Compare the results from the single W/L device and

the two $W(L/2)$ devices connected in series to determine a value for c . Express the general result in terms of X/W , the exponent a , and length and width. This expression is quite general and holds for any device where the current flows longitudinally, such as resistors and junction FETs, and for both white noise and flicker noise.²⁷⁰

- 7.34. Compare the white noise model (7.5.20) and the flicker noise model (7.5.30) and (7.5.31) with the relation you derived in Prob. 7.33. Do they satisfy the requirement stated in that problem? [Hint: Write the inversion charge term in (7.5.20) in terms of Q'_f . For thermal noise, in the resulting relation there should be no separate direct dependence on I_{DS} or g_m , therefore $a = 0$].

8

High-Frequency Small-Signal Models

8.1 Introduction

In this chapter we study models that are valid in a wider frequency range than the five-capacitance quasi-static model of Chap. 7.¹⁻⁵⁵ The first model considered is what will be called the *complete quasi-static* model. This model represents an attempt to take every possible advantage of the quasi-static assumption and gives an improved upper frequency limit of validity. Beyond that limit, though, the model becomes very inaccurate. It is then necessary to consider *non-quasi-static* models. Such models will be considered after a general discussion of *y*-parameter models. A comparison of all models presented will be performed. Noise at high frequencies will then be considered. The chapter will conclude with a discussion of modeling for radio-frequency (RF) applications, including the effects of extrinsic elements.

All effects considered in this chapter will be understood to be for the *intrinsic* part of the transistor (see Fig. 6.1) except in Secs. 8.3 and 8.6. No small-channel effects will be assumed, unless stated otherwise. Also, gate and body leakage currents will be assumed to be zero, and thus only charging currents will be present in those terminals. If desired, the dc small-signal conductances due to gate and body leakage currents can be added as discussed in Sec. 7.2.3.

8.2 A Complete Quasi-Static Model for the Intrinsic Part

8.2.1 Complete Description of Intrinsic Capacitance Effects

In Sec. 7.3, we assumed quasi-static operation (defined in Sec. 6.2), and we modeled the capacitance effect of the drain, source, and substrate on the gate, and the effects of the drain and source on the substrate. Clearly, we did not consider all possible combinations in choosing these five effects. Nevertheless, the model was claimed to be satisfactory for many applications up to a certain frequency. We will now undertake the rigorous development of a *complete* quasi-static model. The model will be complete in the sense that the capacitance effect of *every* terminal on *every* other will be modeled.^{18,28–37,39,47,53}

Consider an intrinsic transistor with time-varying voltages, as shown in Fig. 8.1a. Lowercase letters with capital subscripts denote *total* quantities (as opposed to bias or small-signal quantities). In Sec. 6.3, we showed how the transport and charging components of the drain and source currents in quasi-static operation can be evaluated. The meaning of the drain and source charging components is rather subtle and was discussed at length. The gate and substrate currents consist only of charging components. The four charging currents satisfy (6.3.15) and are given by (6.3.16a) to (6.3.16d). We now assume that the total voltages consist of a dc bias part and a small-signal part, as shown in Fig. 8.1b. Bias quantities will be represented by

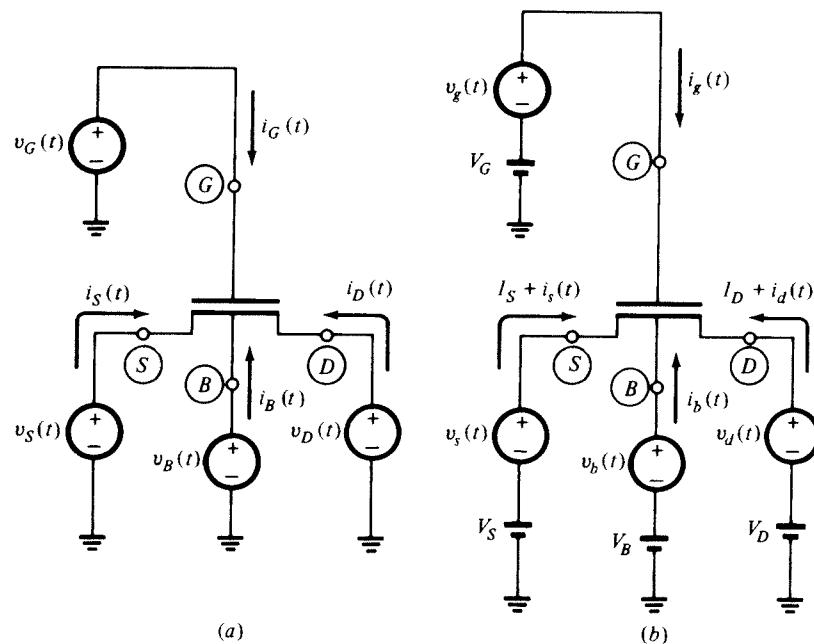


FIGURE 8.1
(a) A transistor with four time-varying terminal voltages; (b) a transistor with terminal voltages consisting of a dc bias plus a time-varying small signal.

capital symbols with capital subscripts; small-signal quantities will be represented by lowercase symbols with lowercase subscripts. If the small signals $v_d(t) = v_g(t) = v_b(t) = v_s(t) = 0$, for all t , then $dv_D/dt = dv_G/dt = dv_B/dt = dv_S/dt = 0$, and thus all currents in (6.3.16) are zero. If the small-signal voltages are nonzero but vary sufficiently slowly, then all charging currents in (6.3.16) will be very small. They will be represented by $i_{da}(t)$, $i_g(t)$, $i_b(t)$, and $i_{sd}(t)$. Over the small range of variation of the small-signal voltages, the slopes of the form $\partial q_K/\partial v_L$, where q_K is any of the four charges and v_L any of the voltages in (6.3.16), will be assumed constant and equal to their values at the bias (or “operating”) point ($v_D = V_D$, $v_G = V_G$, $v_B = V_B$, $v_s = V_S$). Denoting evaluation at this point by o , we define

$$C_{kk} \equiv + \left. \frac{\partial q_K}{\partial v_K} \right|_o \quad (8.2.1a)$$

$$C_{kl} \equiv - \left. \frac{\partial q_K}{\partial v_L} \right|_o, \quad l \neq k \quad (8.2.1b)$$

The choice of the algebraic signs in the preceding definitions is common. It will prove convenient in developing small-signal equivalent circuits later in this section and in relating these circuits to the ones in Sec. 7.3. The sign choice in (8.2.1b) is consistent with (7.3.1) to (7.3.5), the negative sign of which was discussed before (7.3.1) (see also Prob. 8.1).

Of the capacitance parameters defined previously, five (C_{gs} , C_{bs} , C_{gd} , C_{bd} , and C_{gb}) have precisely the meaning discussed in Sec. 7.3. Using the above definitions, we have, from (6.3.16) and the preceding discussion, the following expressions for the small-signal charging currents:

$$i_{da}(t) = +C_{dd} \frac{dv_d}{dt} - C_{dg} \frac{dv_g}{dt} - C_{db} \frac{dv_b}{dt} - C_{ds} \frac{dv_s}{dt} \quad (8.2.2a)$$

$$i_g(t) = -C_{gd} \frac{dv_d}{dt} + C_{gg} \frac{dv_g}{dt} - C_{gb} \frac{dv_b}{dt} - C_{gs} \frac{dv_s}{dt} \quad (8.2.2b)$$

$$i_b(t) = -C_{bd} \frac{dv_d}{dt} - C_{bg} \frac{dv_g}{dt} + C_{bb} \frac{dv_b}{dt} - C_{bs} \frac{dv_s}{dt} \quad (8.2.2c)$$

$$i_{sd}(t) = -C_{sd} \frac{dv_d}{dt} - C_{sg} \frac{dv_g}{dt} - C_{sb} \frac{dv_b}{dt} + C_{ss} \frac{dv_s}{dt} \quad (8.2.2d)$$

It will be helpful in this discussion not to associate the various capacitance parameters above with any physical capacitor-like structures in the MOS transistor. It is better to consider them for the present as simply quantities defined in the precise manner of (8.2.1). We note here that in general $C_{kl} \neq C_{lk}$. For example, consider a long-channel device in saturation. Varying the voltage at the drain will not affect the rest of the device because of pinchoff (assuming no channel length modulation).

Hence, the gate charge will not change [see (6.4.30)], and, from (8.2.1b), C_{gd} will be zero. However, varying the *gate* voltage will change the inversion layer charge. As explained in Sec. 6.3, this change will be accomplished in part by the drain current temporarily becoming different from the transport value. The difference under small-signal conditions is $i_{da}(t)$ and (assuming all other voltages are kept constant) is equal to $-C_{dg}(dv_g/dt)$ from (8.2.2a), which can only be nonzero if $C_{dg} \neq 0$. Another way to see this is to observe that the charge “associated with the drain”[†] does depend on the gate voltage even in saturation [see (6.4.28)]. Thus, from (8.2.1b), $C_{dg} \neq 0$. Therefore, it is seen that $C_{gd} \neq C_{dg}$, and *this can be verified by measurements*, as will be seen. This fact may seem strange at first, because we may have a tendency to think of C_{gd} and C_{dg} as the capacitances of two-terminal capacitors between gate and drain. However, such an interpretation is *not* correct. C_{gd} represents the effect of the drain on the gate, and C_{dg} represents the effect of the gate on the drain, in terms of charging currents. There is no reason to expect that the two effects are the same in general, just as there is no reason to expect that, at dc in saturation, the effect of the drain on the gate current (which is zero assuming no leakage) is the same as the effect of the gate on the drain current (which can be large). We will expand on this point later on.

We now make some important observations about the capacitance parameters and (8.2.2). First, assume that $v_d(t) = v_g(t) = v_b(t) = v_s(t) = v(t)$ in Fig. 8.1b. This is equivalent to the situation shown in Fig. 8.2. From (8.2.2a), we will have

$$i_{da}(t) = (C_{dd} - C_{dg} - C_{db} - C_{ds}) \frac{dv}{dt} \quad (8.2.3)$$

However, since there is no small-signal voltage across any two of the terminals in Fig. 8.2, all terminal small-signal currents must be zero. Since this must be true even for nonzero dv/dt , (8.2.3) implies that

$$C_{dd} - C_{dg} - C_{db} - C_{ds} = 0 \quad (8.2.4)$$

Let us now make another observation. As follows from (6.3.15), the small-signal charging currents must add up to zero:

$$i_{da}(t) + i_g(t) + i_b(t) + i_{sa}(t) = 0 \quad (8.2.5)$$

Consider a case in which $dv_g/dt = dv_b/dt = dv_s/dt = 0$. Using this in (8.2.2a) to (8.2.2d) and the results in (8.2.5), we have

$$(C_{dd} - C_{gd} - C_{bd} - C_{sd}) \frac{dv_d}{dt} = 0 \quad (8.2.6)$$

which, since it must be valid even for nonzero dv_d/dt , implies that

$$C_{dd} - C_{gd} - C_{bd} - C_{sd} = 0 \quad (8.2.7)$$

[†] To appreciate the points made here requires a careful interpretation of the term “charge associated with the drain.” This term can easily be misinterpreted. A related extensive discussion, which is *very* relevant to our present topic, has been given in Sec. 6.3.

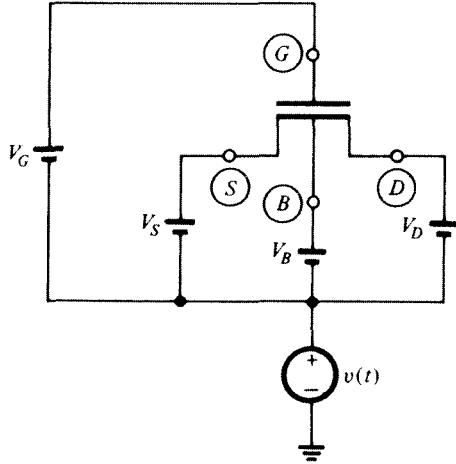


FIGURE 8.2
A transistor with all four terminal-to-ground small-signal voltages equal.

Equations (8.2.4) and (8.2.7) provide two expressions for C_{dd} in terms of other capacitance parameters. Similar expressions can be derived in the same manner for C_{gg} , C_{bb} , and C_{ss} . Thus, we have

$$C_{dd} = C_{dg} + C_{db} + C_{ds} = C_{gd} + C_{bd} + C_{sd} \quad (8.2.8a)$$

$$C_{gg} = C_{gd} + C_{gb} + C_{gs} = C_{dg} + C_{bg} + C_{sg} \quad (8.2.8b)$$

$$C_{bb} = C_{bd} + C_{bg} + C_{bs} = C_{db} + C_{gb} + C_{sb} \quad (8.2.8c)$$

$$C_{ss} = C_{sd} + C_{sg} + C_{sb} = C_{ds} + C_{gs} + C_{bs} \quad (8.2.8d)$$

An interesting result that can be derived from these equations is considered in Prob. 8.2.

Our next observation is simply that, if three of the small-signal charging currents are known, the fourth can be determined from (8.2.5). Thus, any one among the four equations (8.2.2) can be omitted without losing any information. We will omit the last equation from now on.

Without loss of generality, we can write (see Fig. 8.3)

$$v_D = v_{DS} + v_S \quad (8.2.9a)$$

$$v_G = v_{GS} + v_S \quad (8.2.9b)$$

$$v_B = v_{BS} + v_S \quad (8.2.9c)$$

Using analogous expressions for the small-signal voltages in (8.2.2a) gives

$$i_{dd}(t) = C_{dd} \frac{dv_{ds}}{dt} - C_{dg} \frac{dv_{gs}}{dt} - C_{db} \frac{dv_{bs}}{dt} + (C_{dd} - C_{dg} - C_{db} - C_{ds}) \frac{dv_s}{dt} \quad (8.2.10)$$

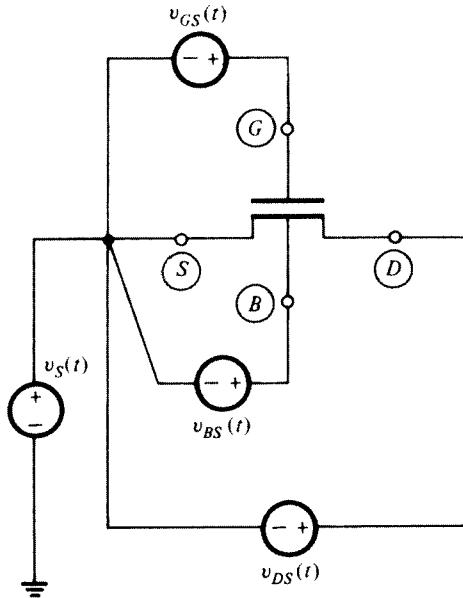


FIGURE 8.3
A transistor with terminal voltages referenced to the source.

The quantity in parentheses is equal to zero, as can be seen from (8.2.8a). Thus,

$$i_{da}(t) = C_{dd} \frac{dv_{ds}}{dt} - C_{dg} \frac{dv_{gs}}{dt} - C_{db} \frac{dv_{bs}}{dt} \quad (8.2.11)$$

Similar relations can be obtained from (8.2.2b) and (8.2.2c). Equation (8.2.2d) will not be considered anymore, as explained earlier. Thus, we finally obtain

$$i_{da}(t) = +C_{dd} \frac{dv_{ds}}{dt} - C_{dg} \frac{dv_{gs}}{dt} - C_{db} \frac{dv_{bs}}{dt} \quad (8.2.12a)$$

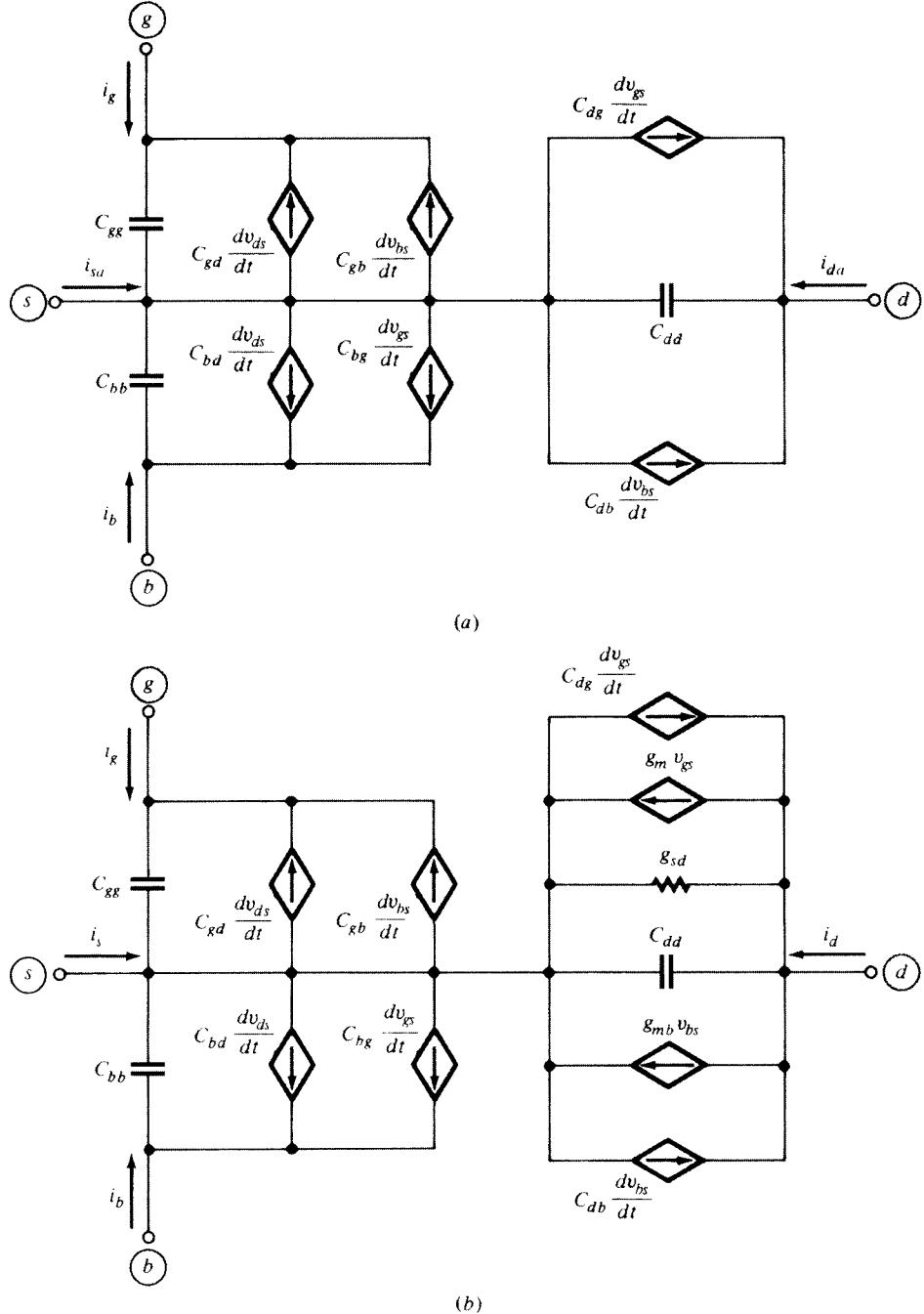
$$i_g(t) = -C_{gd} \frac{dv_{ds}}{dt} + C_{gg} \frac{dv_{gs}}{dt} - C_{gb} \frac{dv_{bs}}{dt} \quad (8.2.12b)$$

$$i_b(t) = -C_{bd} \frac{dv_{ds}}{dt} - C_{bg} \frac{dv_{gs}}{dt} + C_{bb} \frac{dv_{bs}}{dt} \quad (8.2.12c)$$

It is thus clear from the above discussion that *a complete small-signal description of the charging mechanisms requires no less (and no more) than nine independent capacitance parameters*.

8.2.2 Small-Signal Equivalent Circuit Topologies

Small-signal equivalent circuits can be derived to represent (8.2.12). Many such circuits can be constructed; the most straightforward one is shown in Fig. 8.4a. This circuit can be verified by writing Kirchhoff's current law for terminals *g*, *d*, and *b*, which results in (8.2.12). We do not yet have a complete model for the transistor

**FIGURE 8.4**

(a) A small-signal equivalent circuit for the intrinsic charging current action; (b) a complete quasi-static model resulting from (a) by adding the transport current modeling elements of Fig. 7.2.

though, because the transport component of the drain and source currents, originally in (6.3.4), was omitted from the subsequent development. Now, if the small-signal voltages in Fig. 8.1b are zero, all terminal voltages and currents are fixed. The charging components are thus zero, and (6.3.4) gives (using capital I to represent dc current)

$$I_D = I_T \quad (8.2.13a)$$

$$I_S = -I_T \quad (8.2.13b)$$

In the general case, where the small-signal voltages are not zero, (6.3.4) leads to

$$I_D + i_d(t) = I_T + i_l(t) + i_{da}(t) \quad (8.2.14a)$$

$$I_S + i_s(t) = -I_T - i_l(t) + i_{sa}(t) \quad (8.2.14b)$$

where now both small-signal transport components $i_l(t)$ and small-signal charging components appear in the drain and source currents. Subtracting (8.2.13) from (8.2.14), we obtain equations relating only the small-signal currents:

$$i_d(t) = i_l(t) + i_{da}(t) \quad (8.2.15a)$$

$$i_s(t) = -i_l(t) + i_{sa}(t) \quad (8.2.15b)$$

The small-signal transport current has been modeled by the three-element combination of Fig. 7.2. Since this current *adds* to the charging current in (8.2.15a), the corresponding part of the model of Fig. 7.2 should be added in *parallel* with the rest of the elements between drain and source in Fig. 8.4a. This yields the complete model of Fig. 8.4b.

A different small-signal equivalent circuit will now be derived. Let us use

$$v_{ds} = v_{dg} + v_{gs} = -v_{gd} + v_{gs} \quad (8.2.16a)$$

$$v_{bs} = v_{bg} + v_{gs} = -v_{gb} + v_{gs} \quad (8.2.16b)$$

Substituting these equations in (8.2.12b), we obtain

$$\begin{aligned} i_g(t) &= -C_{gd} \left(-\frac{dv_{gd}}{dt} + \frac{dv_{gs}}{dt} \right) + C_{gg} \frac{dv_{gs}}{dt} - C_{gb} \left(-\frac{dv_{gb}}{dt} + \frac{dv_{gs}}{dt} \right) \\ &= C_{gd} \frac{dv_{gd}}{dt} + C_{gb} \frac{dv_{gb}}{dt} + (C_{gg} - C_{gd} - C_{gb}) \frac{dv_{gs}}{dt} \end{aligned} \quad (8.2.17)$$

Using (8.2.8b), we find that this becomes

$$i_g(t) = C_{gd} \frac{dv_{gd}}{dt} + C_{gb} \frac{dv_{gb}}{dt} + C_{gs} \frac{dv_{gs}}{dt} \quad (8.2.18)$$

Using similar manipulations in (8.2.12a) and (8.2.12c), we can write (8.2.12) in the form

$$i_{da}(t) = C_{gd} \frac{dv_{dg}}{dt} + C_{sd} \frac{dv_{ds}}{dt} + C_{bd} \frac{dv_{db}}{dt} - C_m \frac{dv_{gs}}{dt} - C_{mb} \frac{dv_{bs}}{dt} \quad (8.2.19a)$$

$$i_g(t) = C_{gd} \frac{dv_{gd}}{dt} + C_{gb} \frac{dv_{gb}}{dt} + C_{gs} \frac{dv_{gs}}{dt} \quad (8.2.19b)$$

$$i_b(t) = C_{bd} \frac{dv_{bd}}{dt} + C_{gb} \frac{dv_{bg}}{dt} - C_{mx} \frac{dv_{gb}}{dt} + C_{bs} \frac{dv_{bs}}{dt} \quad (8.2.19c)$$

where

$$C_m = C_{dg} - C_{gd} \quad (8.2.20a)$$

$$C_{mb} = C_{db} - C_{bd} \quad (8.2.20b)$$

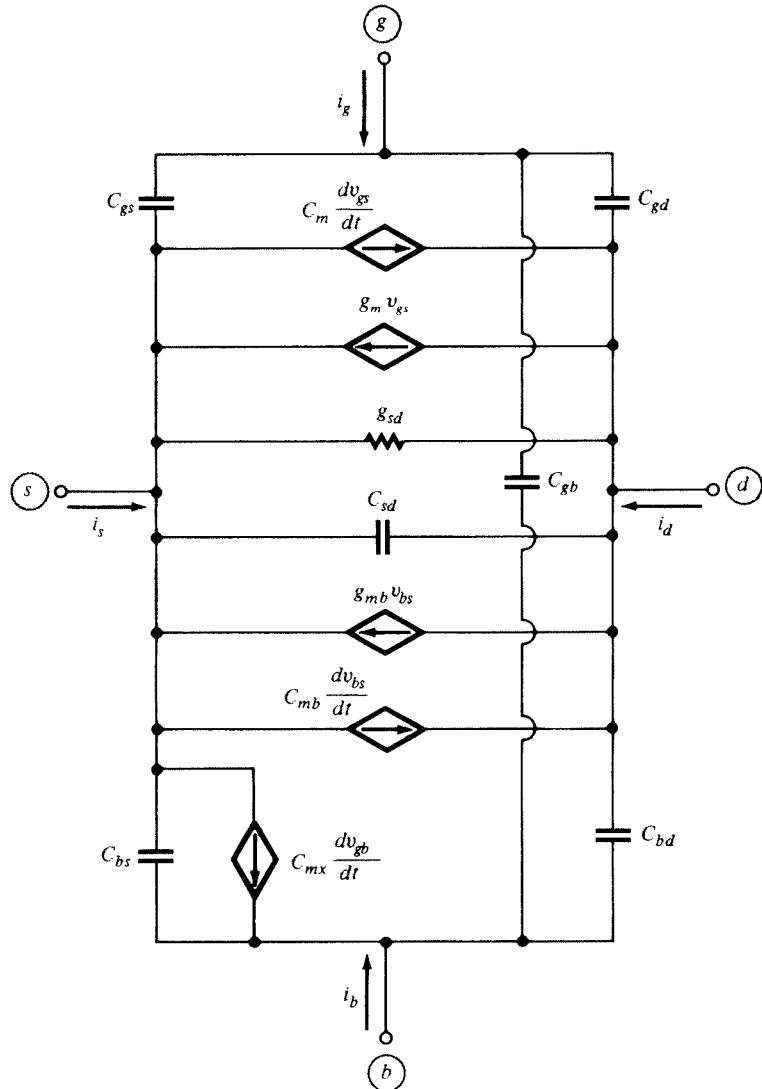
$$C_{mx} = C_{bg} - C_{gb} \quad (8.2.20c)$$

Equation (8.2.19b) has already been proved. To prove (8.2.19a) and (8.2.19c), simply express all voltages in these equations in terms of v_{ds} , v_{gs} , and v_{bs} and use (8.2.20) and (8.2.8); this will give (8.2.12a) and (8.2.12c).

Equations (8.2.19) are easily represented by an equivalent circuit. When the elements of Fig. 7.2 are added to it, we obtain the result shown in Fig. 8.5.³⁶ Five capacitors shown in this figure (C_{gs} , C_{gd} , C_{bs} , C_{bd} , and C_{gb}) are exactly the same as those in the model of Fig. 7.19. In other words, the model of Fig. 8.5 can be viewed as resulting from simply *augmenting* the model of Fig. 7.19 by four more elements, but *without* having to modify the elements already in that popular model, neither in meaning nor in value. This is an important property of the model in Fig. 8.5, in contrast to other models proposed in the literature. This point is considered further in Probs. 8.5 and 8.6.

Note that, in Fig. 8.5, at sufficiently low frequencies dv_{gs}/dt and dv_{bs}/dt will be small, and the currents proportional to these quantities can be neglected in comparison to the currents $g_m v_{gs}$ and $g_{mb} v_{bs}$, respectively. Similarly, the current through C_{sd} can then be neglected in comparison to the current through g_{sd} . Finally, we will find that for the approximate strong-inversion model $C_{mx} = 0$. Thus, the model of Fig. 8.5 reduces to that of Fig. 7.19 at sufficiently low frequencies. More on model comparison will be found later in this chapter.

To provide more feeling for the model in Fig. 8.5, we will consider two experiments, as illustrated in Fig. 8.6. In Fig. 8.6a, a small-signal voltage is applied only at the drain, and the resulting small-signal current entering the gate is observed; in Fig. 8.6b, the opposite is done. The small-signal equivalent circuit used in each case on the right is that of Fig. 8.5, but only elements with nonzero current through them are shown for simplicity (short-circuited capacitances or resistances, and current sources proportional to zero voltages are omitted). As seen in Fig. 8.6a, the small-signal current entering the gate is $-C_{gd}(dv_d/dt)$. Thus, C_{gd} actually represents

**FIGURE 8.5**

A complete quasi-static small-signal model for the intrinsic part of the MOS transistor. Although independently derived, the model can be viewed as resulting from the simpler model of Fig. 7.19 by adding four elements to it.

the effect of terminal d on terminal g . However, although the capacitance C_{gd} in Fig. 8.5 is connected between g and d in the general case, it *does not* represent the total effect of g on d . This can be seen in Fig. 8.6b. Indeed, the current entering the drain is $g_m v_g - (C_{gd} + C_m)(dv_g/dt)$. Note that not only is there a conductive current here but also the capacitive current is *different* from that in Fig. 8.6a. Thus, $C_{gd} + C_m$ represents the capacitive effect of g on d . From (8.2.20a), $C_{gd} + C_m = C_{dg}$, which, in

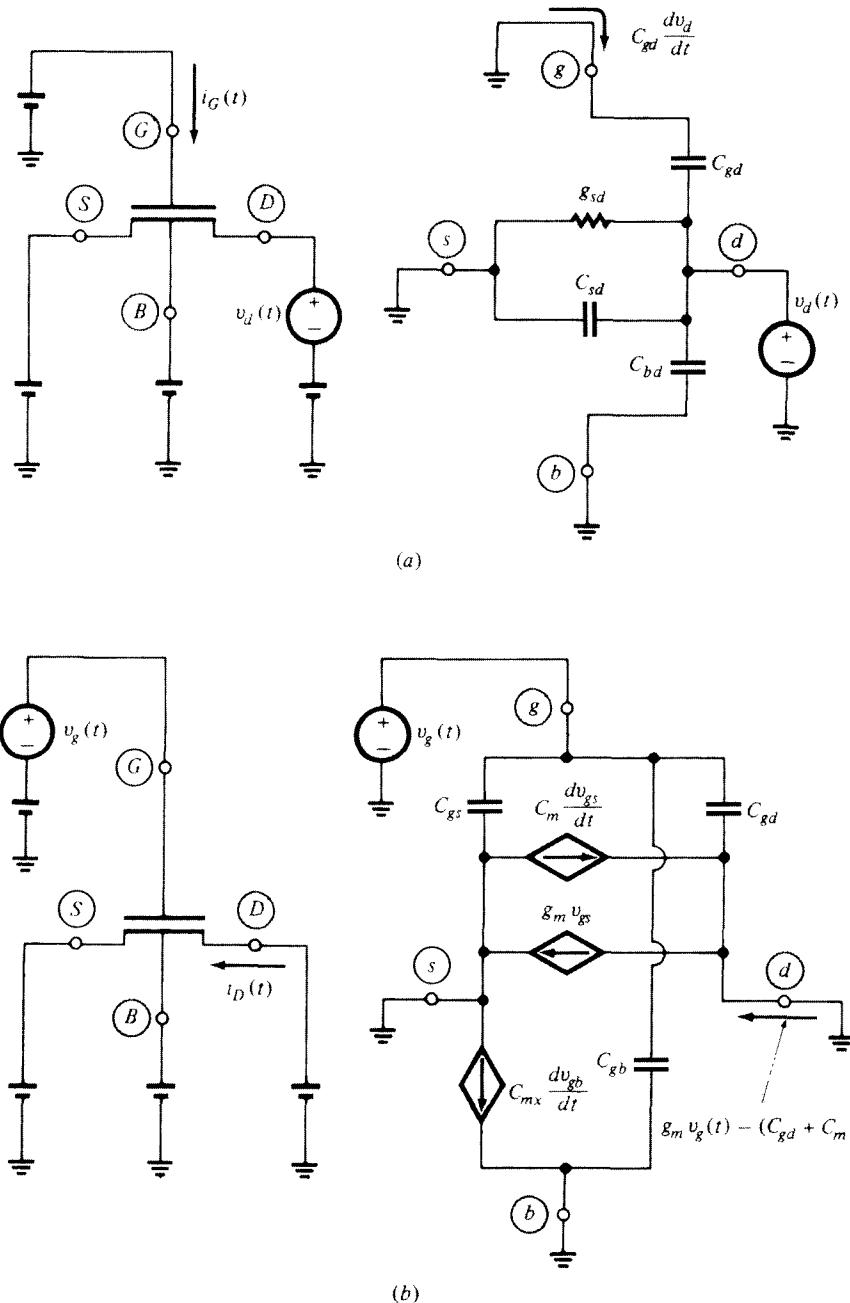
**FIGURE 8.6**

Illustration of the way in which the different effect of the gate and drain on each other is taken into account in the model of Fig. 8.5: (a) effect of drain voltage on gate current; (b) effect of gate voltage on drain current. In each part, the left half shows the complete circuit; the right half shows the small-signal equivalent circuit, in which elements with zero current through them are omitted.

general, is *different* from C_{gd} , as we have already argued, and as we will see quantitatively in the following subsection. C_m is then a *transcapacitance*, taking care of the different effect of the gate and drain on each other in terms of charging currents, just as g_m is a transconductance taking care of the different effect of these two terminals on each other in terms of transport currents. Similar comments hold for C_{mb} and C_{mv} .

8.2.3 Evaluation of Capacitances

Strong Inversion In Sec. 7.3, we used the charges corresponding to the approximate model to evaluate five capacitances: C_{gs} , C_{bs} , C_{gd} , C_{bd} , C_{gb} , and we gave several plots of these capacitances vs. V_{DS} and V_{GS} . The evaluation of the new capacitances that have been defined in this section is done in the same manner. Again, in deriving expressions for these capacitances, we will use the simplifying assumptions of Sec. 7.3, that is, $\alpha = \alpha_1$ as given by (7.3.7), and $d\alpha_1/dV_s$ and $d\alpha_1/dV_B$ are negligible. C_{dg} is found by using $q_D = Q_D$ from (6.4.19) in the definition (8.2.1b).[†] This gives, after much algebra,

$$C_{dg} = C_{ox} \frac{4 + 28\eta + 22\eta^2 + 6\eta^3}{15(1+\eta)^3} \quad (8.2.21)$$

with η as defined in (4.7.25) and plotted in Fig. 4.25, and $C_{ox} = WLC'_{ox}$ the total intrinsic oxide capacitance. For C_{db} , we find, using the same approach as for C_{bd} in Sec. 7.3:

$$C_{db} = b_c C_{dg} \quad (8.2.22)$$

where, from (7.3.18), $b_c = C_{bc}/C_{ox}$, with C_{bc} the body-channel capacitance given by (7.3.12).

Using (6.4.15), we get

$$C_{bg} = \frac{C_{ox} \times C_{bc}}{C_{ox} + C_{bc}} \frac{1}{3} \left(\frac{1-\eta}{1+\eta} \right)^2 \quad (8.2.23)$$

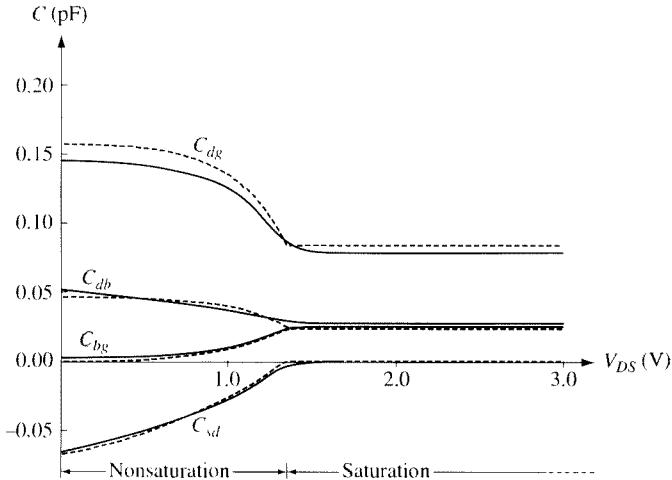
Comparing the last equation to (7.3.20), we see that, *for the approximate model with the simplifying assumptions we have made*, $C_{bg} = C_{gb}$.

To evaluate C_{sd} , we use Q_S from (6.4.20) in the definition (8.2.1b) and obtain

$$C_{sd} = -\frac{4}{15} C_{ox} \alpha_1 \frac{\eta + 3\eta^2 + \eta^3}{(1+\eta)^3} \quad (8.2.24)$$

Notice that this quantity is *negative* in nonsaturation (in saturation, C_{sd} becomes zero). This is in agreement with measurements.³² The negative value can be viewed

[†] The charges and voltages in Sec. 6.4 were assumed to be dc quantities. As before, we will use the expressions developed in that section, with Q replaced by q and V replaced by v , under the assumption of quasi-static operation (Sec. 6.2).

**FIGURE 8.7**

Drain-gate, drain-body, body-gate, and source-drain small-signal intrinsic capacitances vs. V_{DS} in strong inversion for a large-gate device with a thick oxide. Dashed lines are for the approximate strong-inversion model discussed in the text; solid lines are for an accurate surface potential-based model. $t_{ox} = 11 \text{ nm}$, $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 10 \mu\text{m}$, $V_{GS} = 2.5 \text{ V}$, $V_{SB} = 0 \text{ V}$.

intuitively as follows. Raising the drain voltage by an amount ΔV_D will increase the effective reverse bias at the drain end and will cause the magnitude of the inversion layer charge to decrease. Since Q_I is negative, this means a change $\Delta Q_I > 0$. This positive change in Q_I is shared by positive ΔQ_S and positive ΔQ_D . Hence, $C_{sd} = -\partial Q_S / \partial V_D$ will be negative.

Plots of C_{sd} , C_{dg} , C_{db} , and C_{bg} are given in Fig. 8.7. Comparisons to accurate results using surface potential-based models show that good accuracy is obtained except for C_{db} and C_{bg} when V_{SB} is small and V_{GS} and V_{DS} are large.

The preceding relations, along with (7.3.8), (7.3.9), (7.3.16), (7.3.17), and (7.3.20) give values for nine capacitances: C_{gs} , C_{bs} , C_{gd} , C_{dg} , C_{bd} , C_{db} , C_{gb} , C_{bg} , and C_{sd} . Any other capacitance parameter defined in this section can be found from these and (8.2.8) or (8.2.20). For example, using the latter we find

$$C_m = \frac{4}{15} C_{ox} \frac{1 + 2\eta - 2\eta^2 - \eta^3}{(1 + \eta)^3} \quad (8.2.25)$$

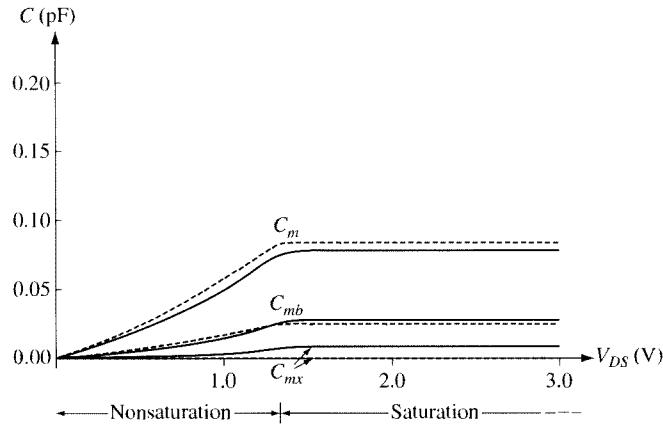
$$C_{mb} = b_c C_m \quad (8.2.26)$$

$$C_{mx} = 0 \quad (8.2.27)$$

where, again, b_c is given by (7.3.18).

Thus, we have expressions for all nine capacitance parameters in the model of Fig. 8.5. Plots of C_m , C_{mb} , and C_{mx} vs. V_{DS} are given in Fig. 8.8. Note that our simple model predicts $C_{bg} = C_{gb}$ or $C_{mx} = 0$. Accurate calculations³⁶ using surface potential-based models give, however,

$$C_{bg} > C_{gb} \quad (8.2.28)$$

**FIGURE 8.8**

Small-signal transcapacitances, C_m , C_{mb} , and C_{mx} vs. V_{DS} in strong inversion for a large-gate device with a thick oxide. Dashed lines are for the approximate strong-inversion model discussed in the text; solid lines are for an accurate surface potential-based model. $t_{ox} = 11 \text{ nm}$, $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 10 \mu\text{m}$, $V_{GS} = 2.5 \text{ V}$, $V_{SB} = 0 \text{ V}$.

and, consequently,

$$C_{mx} > 0 \quad (8.2.29)$$

Nevertheless, C_{mx} is very small and will be unimportant in most practical cases.

Using results from above and (7.3.19), one obtains (if V_{DS} or V_{GS} is small and/or V_{SB} is large)

$$\frac{C_{db}}{C_{dg}} \approx \frac{C_{sb}}{C_{sg}} \approx \frac{C_{bd}}{C_{gd}} \approx \frac{C_{bs}}{C_{gs}} \approx \frac{C_{bb}}{C_{gg}} \approx \frac{C_{mb}}{C_m} \approx \frac{C_{bc}}{C_{ox}} \approx \frac{g_{mb}}{g_m} \approx \frac{dV_T}{dV_{SB}} = \alpha_1 - 1, \quad \text{small } V_{DS} \text{ or small } V_{GS} \quad (8.2.30)$$

We consider now the values of the capacitance parameters in two special cases of interest.

Nonsaturation with $V_{DS} = 0$. Using $V_{DS} = 0$ ($\eta = 1$) in the preceding results, we easily obtain the following values (we also repeat the results obtained in Sec. 7.3 for completeness):

$$C_{gg} = C_{ox} \quad (8.2.31a)$$

$$C_{dg} = C_{gd} = C_{gs} = C_{sg} = \frac{1}{2} C_{ox} \quad (8.2.31b)$$

$$C_{bb} = C_{bc,0} \quad (8.2.31c)$$

$$C_{db} = C_{bd} = C_{bs} = C_{sb} = \frac{1}{2} C_{bc,0} \quad (8.2.31d)$$

$$C_{gb} = C_{bg} = 0 \quad (8.2.31e)$$

$$C_{dd} = C_{ss} = \alpha_1 \frac{C_{\text{ox}}}{3} \quad (8.2.31f)$$

$$C_{ds} = C_{sd} = -\alpha_1 \frac{C_{\text{ox}}}{6} \quad (8.2.31g)$$

$$C_m = C_{mb} = C_{mx} = 0 \quad (8.2.31h)$$

where is $C_{bc,0}$ the value of C_{bc} at $V_{DS} = 0$, found from (7.2.12) and (7.2.13) with $\eta = 1$.

We note that, for the case we are considering ($V_{DS} = 0$), we have $C_{kl} = C_{lk}$, where k and l represent any two among s , d , g , and b . This is actually the only case where such an equality is observed exactly by all C_{kl} ; with $V_{DS} \neq 0$, one finds $C_{kl} \neq C_{lk}$ (for $k \neq l$) in general, for reasons already discussed.

Some of the preceding results have already been obtained and discussed in Sec. 7.3. Here we will attempt to make plausible the rest of them. The value of the inversion layer charge per unit area, Q'_l at the source and drain ends of the channel is, from Sec. 3.4.2,

$$Q'_{l0} = -C'_{\text{ox}}(V_{GS} - V_{TS}) \quad (8.2.32a)$$

$$Q'_{lL} = -C'_{\text{ox}}(V_{GD} - V_{TD}) \quad (8.2.32b)$$

where

$$V_{TS} = V_{FB} + \phi_0 + \gamma \sqrt{V_{SB} + \phi_0} \quad (8.2.33a)$$

$$V_{TD} = V_{FB} + \phi_0 + \gamma \sqrt{V_{DB} + \phi_0} \quad (8.2.33b)$$

Here we are interested in the case $V_{DS} = 0$. However, we have allowed for independent voltages V_D and V_S , so that we can differentiate the expressions involved in terms of one voltage while keeping the other constant in accordance with the capacitance definitions.

For $V_D \approx V_S$, one can assume a straight-line variation of Q'_l with the position x along the channel:

$$Q'_l = Q'_{l0} + (Q'_{lL} - Q'_{l0}) \frac{x}{L} \quad (8.2.34)$$

One can now find Q_D and Q_S by using (8.2.32) and (8.2.34) in (6.3.9). The result is

$$Q_D = -C_{\text{ox}} \left[\frac{1}{6}(V_{GS} - V_{TS}) + \frac{1}{3}(V_{GD} - V_{TD}) \right] \quad (8.2.35)$$

$$Q_S = -C_{\text{ox}} \left[\frac{1}{3}(V_{GS} - V_{TS}) + \frac{1}{6}(V_{GD} - V_{TD}) \right] \quad (8.2.36)$$

Various capacitance values in (8.2.31) can now be verified easily by using the preceding expressions. In particular, the reason for the denominators 3 and 6 in (8.2.31f) and (8.2.31g) becomes apparent.

All results in (8.2.31) agree exactly with those obtained by using the charges corresponding to the complete strong-inversion model.

Saturation. Using the general capacitance expressions with $V_{DS} = V'_{DS}$ ($\eta = 0$) gives the following results (some of them have already been derived in Sec. 7.3 and are repeated here for convenience):

$$C_{dg} = \frac{4}{15} C_{ox} \quad (8.2.37a)$$

$$C_{gd} = 0 \quad (8.2.37b)$$

$$C_{db} = \frac{4}{15} C_{bc,\text{sat}} \quad (8.2.37c)$$

$$C_{bd} = 0 \quad (8.2.37d)$$

$$C_{sg} = \frac{2}{5} C_{ox} \quad (8.2.37e)$$

$$C_{gs} = \frac{2}{3} C_{ox} \quad (8.2.37f)$$

$$C_{sb} = \frac{2}{5} C_{bc,\text{sat}} \quad (8.2.37g)$$

$$C_{bs} = \frac{2}{3} C_{bc,\text{sat}} \quad (8.2.37h)$$

$$C_{gb} = C_{bg} = \frac{1}{3} \frac{C_{ox} \times C_{bc,\text{sat}}}{C_{ox} + C_{bc,\text{sat}}} \quad (8.2.37i)$$

$$C_{ds} = -\alpha_1 \frac{4}{15} C_{ox} \quad (8.2.37j)$$

$$C_{sd} = 0 \quad (8.2.37k)$$

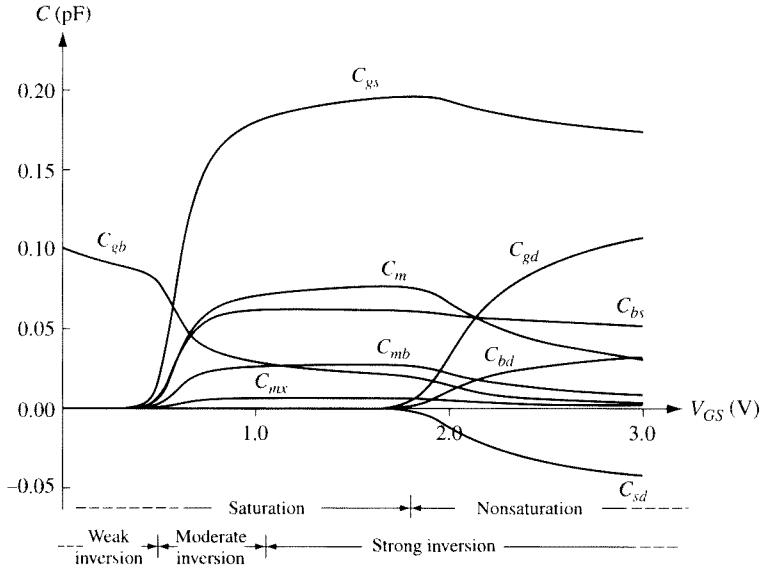
$$C_{gg} = \frac{2}{3} C_{ox} + \frac{1}{3} \frac{C_{ox} \times C_{bc,\text{sat}}}{C_{ox} + C_{bc,\text{sat}}} \quad (8.2.37l)$$

$$C_{bb} = \frac{2}{3} C_{bc,\text{sat}} + \frac{1}{3} \frac{C_{ox} \times C_{bc,\text{sat}}}{C_{ox} + C_{bc,\text{sat}}} \quad (8.2.37m)$$

$$C_{dd} = 0 \quad (8.2.37n)$$

$$C_{ss} = \alpha_1 \frac{2}{5} C_{ox} \quad (8.2.37o)$$

$$C_m = \frac{4}{15} C_{ox} \quad (8.2.37p)$$

**FIGURE 8.9**

The nine intrinsic capacitances in the small-signal model of Fig. 8.5 vs. V_{GS} from an accurate all-region model for a large-gate device with a thick oxide. $t_{ox} = 11 \text{ nm}$, $N_A = 1 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 10 \mu\text{m}$, $V_{DS} = 0.9 \text{ V}$, $V_{SB} = 0 \text{ V}$.

$$C_{mb} = \frac{4}{15} C_{bc,\text{sat}} \quad (8.2.37q)$$

$$C_{mx} = 0 \quad (8.2.37r)$$

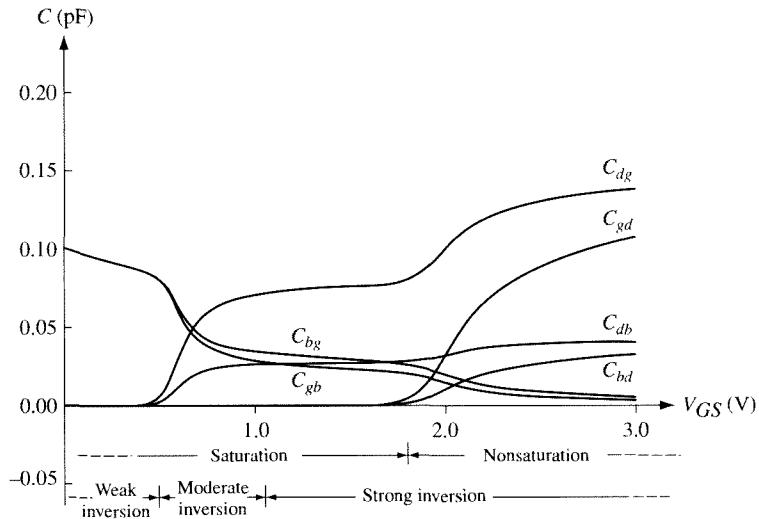
where $C_{bc,\text{sat}}$ is the value of C_{bc} in saturation, found from (7.2.12) and (7.2.13) with $\eta = 0$.

Note that in saturation we have $C_{kl} \neq C_{lk}$ (for $k \neq l$), which is to be expected, for the reasons discussed following (8.2.2).[†] Note also that in saturation $C_{gd} = C_{bd} = C_{sd} = 0$, which is a manifestation of the fact that in this region the drain voltage does not influence the device charges.

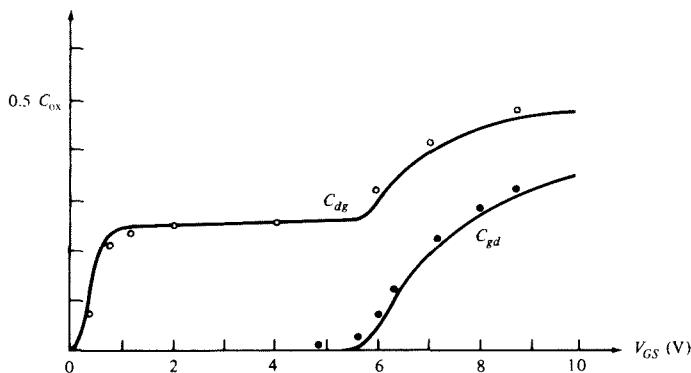
Weak Inversion The most important intrinsic capacitance in weak inversion, C_{gb} , has been discussed in Sec. 7.3. The other intrinsic capacitances in Fig. 8.5 are small, and their effect is usually swamped by that of extrinsic capacitances (Secs. 6.8.1 and 7.4) for all but very long devices.

All-Region Model Capacitances derived by using the accurate charge calculations suggested in Sec. 6.4.5 are shown in Fig. 8.9 vs. V_{GS} .³⁶ All nine capacitances used

[†] The only exception is C_{gb} and C_{bg} , which are found to be equal under our simplifying assumptions; see, however, (8.2.28) and the associated discussion.

**FIGURE 8.10**

A comparison of C_{dg} to C_{gd} , C_{db} to C_{bd} , and C_{bg} to C_{gb} , for the device of Fig. 8.9, plotted vs. V_{GS} for $V_{DS} = 0.9$ V and $V_{SB} = 0$ V, using an accurate surface potential-based model.

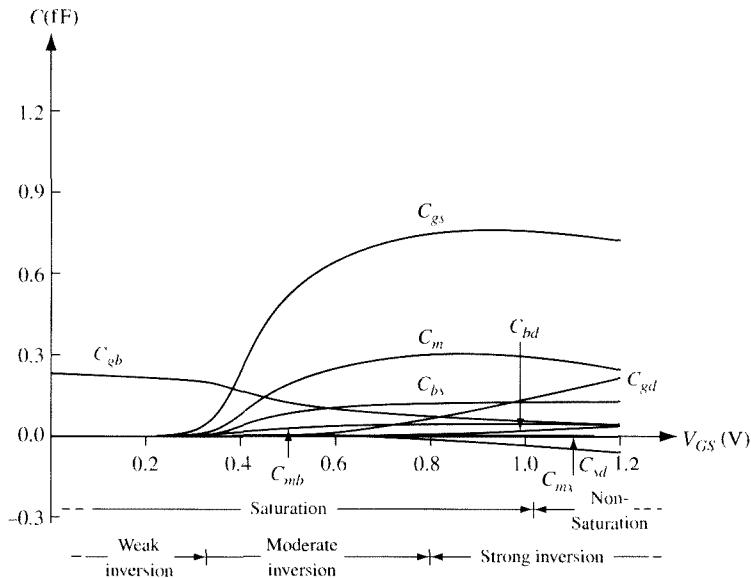
**FIGURE 8.11**

A comparison of C_{dg} to C_{gd} for an older process with $V_{T0} = 0.92$ V plotted vs. V_{GS} for $V_{DS} = 4$ V, and $V_{SB} = 0$ V. Points are measurements;³² lines are accurate calculations.³⁶

in the model of Fig. 8.5 are shown. Figure 8.10 compares C_{gd} to C_{dg} , C_{bd} to C_{db} , and C_{gb} to C_{bg} . As seen, the capacitances in each of these pairs are, in general, different, as predicted by the theory in this section. These predictions agree with measurements.^{29,32,56} An example is shown in Fig. 8.11.

Short-channel devices show qualitatively similar behavior, although the boundaries between regions are generally less distinct. An example is shown in Fig. 8.12, which can be compared with the long-channel plots in Fig. 8.9.

Extrinsic elements can be added as in Sec. 7.4.

**FIGURE 8.12**

The nine intrinsic capacitances in the small-signal model of Fig. 8.5 vs. V_{GS} from an accurate all-region model for a short-channel device with a thin oxide. $t_{ox} = 1.2 \text{ nm}$, $N_A = 1 \times 10^{18} \text{ cm}^{-3}$, $W = 1 \mu\text{m}$, $L = 0.065 \mu\text{m}$, $V_{DS} = 0.4 \text{ V}$. $V_{SB} = 0 \text{ V}$.

8.2.4 Frequency Region of Validity

The inclusion of the four capacitances C_m , C_{mb} , C_{mx} , and C_{sd} (Fig. 8.5) makes the complete quasi-static model better than the simple model of Fig. 7.19, as far as the frequency region of validity is concerned. The improvement depends on bias and on the terminals we are considering. For example, C_m is maximum in saturation, and it is there that the difference between the two models will be maximum, with respect to gate-to-drain action. At $V_{DS} = 0$, $C_m = 0$ and the two models are identical in that respect. On the other hand, it is at $V_{DS} = 0$ that C_{sd} is maximum, and hence it is at this point that the difference between the two models is maximum, as far as drain-to-source action is concerned.

To provide some rough indication of the difference between the two models, consider operation in the saturation region, which is the most important region for small-signal circuit applications. Assume the element values in the complete quasi-static model of Fig. 8.5 have been chosen so that, at very low frequencies, the performance is good. Then the performance will continue to be good, with practically no deterioration up to about $\omega_o/3$, where ω_o is given by (7.3.6).³⁶ This result follows by comparison to more sophisticated, non-quasi-static models considered in Sec. 8.4. At this point, a warning should be given. *Although the frequency region of validity for the complete quasi-static model of Fig. 8.5 is larger, if this region is exceeded, this model can give very wrong results, in fact worse than those of the simple model of Fig. 7.19 in some respects.* This important point is considered in a model comparison in Sec. 8.4.4.

8.3 y-Parameter Models

Design of very-high-frequency circuits (not necessarily using MOS transistors) is often done by using the so-called y parameters. In this section, we develop y -parameter models for the MOS device. We will first derive the general form of such models. In doing so, no assumptions will be made as to the physics of the device. In fact, we do not even have to assume the device is a MOS transistor. We say only that it must have four terminals. For later convenience, we will denote these terminals by D , G , B , and S , but for the present, we do not have to associate these symbols with particular terminals of a specific device; indeed, we might as well have used X , Y , Z , and W instead.

Let us consider the transistor driven by bias and small-signal voltages at each terminal, as shown in Fig. 8.1b. A small-signal equivalent circuit of the transistor driven by the small-signal parts of the voltage excitations is shown in Fig. 8.13a. We assume now that all small-signal voltages are sinusoidal and of the *same* angular frequency ω . Then in the *sinusoidal steady state* all small-signal currents will be sinusoidal and of the same frequency.⁵⁷⁻⁵⁹ The small-signal voltages and currents can be represented by cosine functions, e.g.,

$$v_g(t) = M_{vg} \cos(\omega t + \phi_{vg}) \quad (8.3.1)$$

We will use a *phasor* representation for each small-signal voltage or current, i.e., a complex number with magnitude and angle equal to the amplitude and phase,

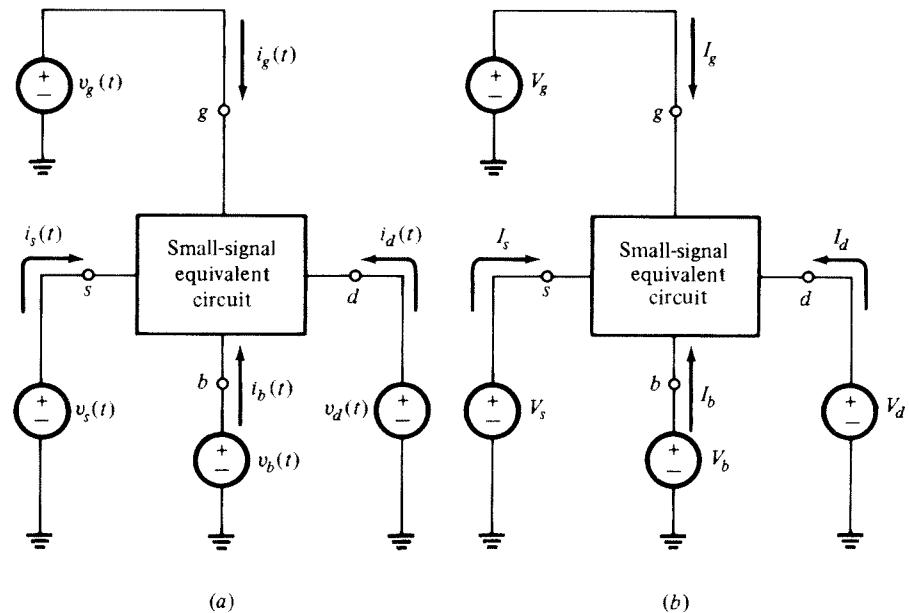


FIGURE 8.13

(a) The small-signal equivalent of Fig. 8.1b in the time domain; (b) corresponding representation in the frequency domain using phasors.

respectively, of the corresponding cosine waveform.⁵⁷⁻⁵⁹ Phasors will be denoted by capital letters with lowercase subscripts. For example, corresponding to $v_g(t)$ in (8.3.1), we have a phasor V_g :

$$V_g = M_{v_g} e^{j\phi_{v_g}} \quad (8.3.2)$$

From now on we will use for brevity the terms “voltage” and “current” instead of the more complete terms “voltage phasor” and “current phasor.” Because the context will be clear, no confusion should arise. The phasor representation for the circuit of Fig. 8.13a is shown in Fig. 8.13b.

Let us assume that we are interested in the effect of V_g , V_b , V_d , and V_s on the current I_d . We will perform four experiments. In each, we will consider only one of the four small-signal voltages by setting the other three equal to zero in Fig. 8.13b. This is equivalent to setting to zero the values of three of the small-signal voltage sources in Fig. 8.1b, but, of course, *leaving all four dc bias sources intact* in that figure. The four experiments are summarized in Fig. 8.14. In each one, the ratio of the current phasor to the voltage phasor is a complex *admittance*.⁵⁷⁻⁵⁹ We will use the symbols shown in Fig. 8.14 for the four admittances.

Small-signal equivalent circuits are *linear* circuits, representing the fact that in an actual transistor with vanishingly small signals, the quantities I_d , I_g , I_b , and I_s are *linear* functions of V_d , V_g , V_b , and V_s . One can thus use *superposition* to find I_d when all four of the small-signal voltages are active (nonzero). This is done by considering one voltage active at a time, evaluating the resulting current, and then adding up the currents:

$$I_d = I_d|_{V_g, V_b, V_s=0} + I_d|_{V_d, V_b, V_s=0} + I_d|_{V_d, V_g, V_s=0} + I_d|_{V_d, V_g, V_b=0} \quad (8.3.3)$$

Using the definitions in Fig. 8.14, the preceding equation can be written as follows:

$$I_d = y_{dd} V_d = y_{dg} V_g + y_{db} V_b + y_{ds} V_s \quad (8.3.4)$$

We can perform similar experiments to determine the currents in each of the other three terminals. If, in each case, we define admittances as follows:

$$y_{kl} = \frac{I_k}{V_l} \Big|_{V_n=0, n \neq l} \quad (8.3.5)$$

we have a total of four equations:

$$I_d = y_{dd} V_d + y_{dg} V_g + y_{db} V_b + y_{ds} V_s \quad (8.3.6a)$$

$$I_g = y_{gd} V_d + y_{gg} V_g + y_{gb} V_b + y_{gs} V_s \quad (8.3.6b)$$

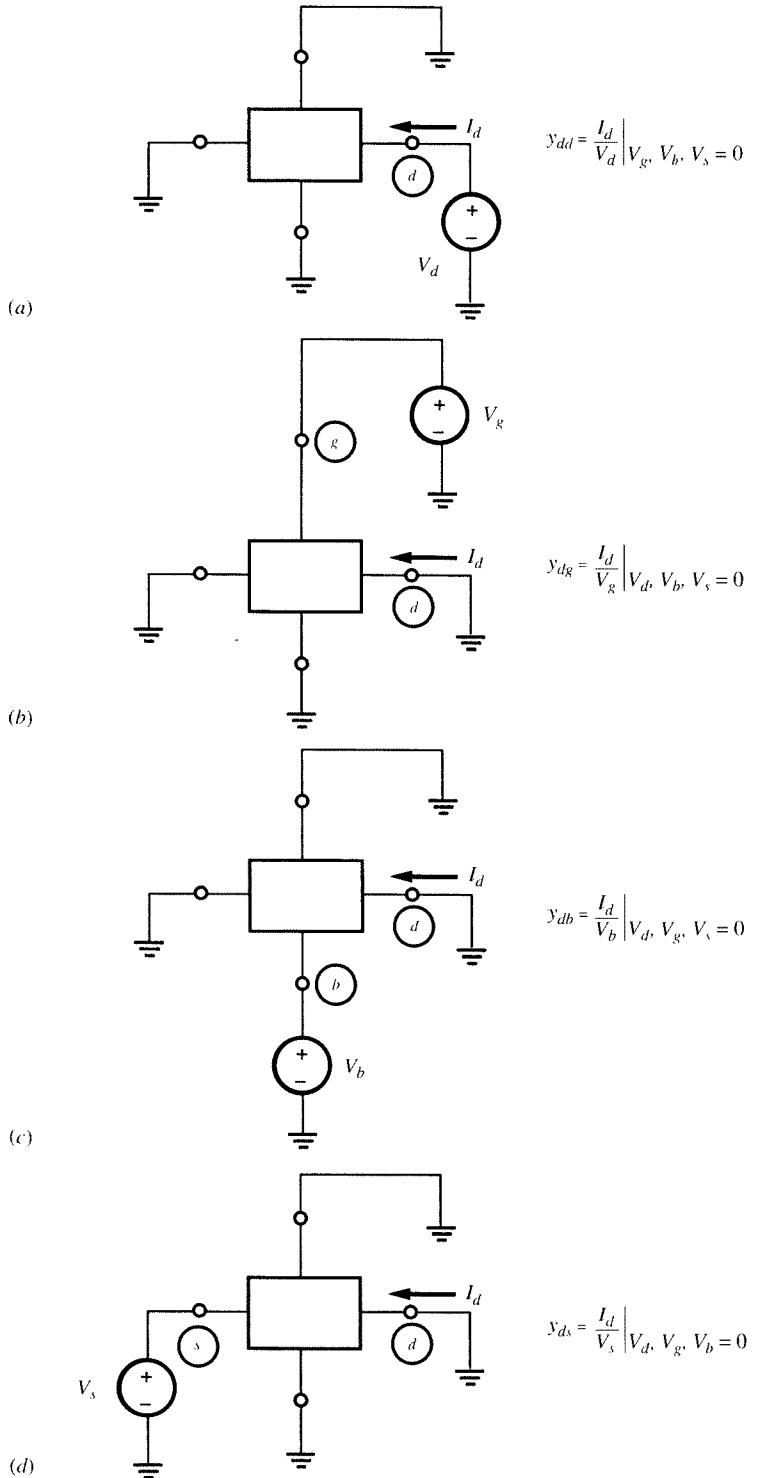


FIGURE 8.14
Definition of y parameters associated with the drain current. The rectangle represents a small-signal equivalent circuit, as in Fig. 8.13. The notation $V_i, V_y, V_z = 0$ means $V_i = 0, V_y = 0, V_z = 0$.

$$I_b = y_{bd}V_d + y_{bg}V_g + y_{bb}V_b + y_{bs}V_s \quad (8.3.6c)$$

$$I_s = y_{sd}V_d + y_{sg}V_g + y_{sb}V_b + y_{ss}V_s \quad (8.3.6d)$$

Note that these equations are similar in structure to (8.2.2), except that here no minus signs are used. This is a consequence of our definition (8.3.5) [which should be compared with (8.2.1)]. This definition of admittance parameters is standard in circuit theory. Equation (8.3.6), expressed in matrix form, is known as a *terminal*, or *indefinite, admittance matrix* representation.⁵⁹

We can follow a reasoning analogous to the one that led to (8.2.8) to get relations between the y parameters (Prob. 8.11):

$$y_{dd} + y_{dg} + y_{db} + y_{ds} = y_{dd} + y_{gd} + y_{bd} + y_{sd} = 0 \quad (8.3.7a)$$

$$y_{gg} + y_{gd} + y_{gb} + y_{gs} = y_{gg} + y_{dg} + y_{bg} + y_{sg} = 0 \quad (8.3.7b)$$

$$y_{bb} + y_{bd} + y_{bg} + y_{bs} = y_{bb} + y_{db} + y_{gb} + y_{sb} = 0 \quad (8.3.7c)$$

$$y_{ss} + y_{sd} + y_{sg} + y_{sb} = y_{ss} + y_{ds} + y_{gs} + y_{bs} = 0 \quad (8.3.7d)$$

Similarly, following a reasoning analogous to the one that led to (8.2.12), we conclude that the fourth equation in (8.3.6) can be omitted (in fact, any one among the four equations could have been chosen for omission) without losing information, and that the remaining three equations can be written as follows:

$$I_d = y_{dd}V_{ds} + y_{dg}V_{gs} + y_{db}V_{bs} \quad (8.3.8a)$$

$$I_g = y_{gd}V_{ds} + y_{gg}V_{gs} + y_{gb}V_{bs} \quad (8.3.8b)$$

$$I_b = y_{bd}V_{ds} + y_{bg}V_{gs} + y_{bb}V_{bs} \quad (8.3.8c)$$

with $V_{kl} = V_k - V_l$. The preceding set of equations can be represented by the circuit of Fig. 8.15, a fact verifiable directly by writing Kirchhoff's law at nodes d , g , and b for the current phasors.

Other three-port y -parameter representations are also possible. For example, if instead of using the s terminal as a potential reference and omitting (8.3.6d), we had used the b terminal as a reference and had omitted (8.3.6c), we would have obtained the representation shown in Fig. 8.16. This representation has an appealing symmetry. For a symmetrically laid out device, the role of source and drain is identical; thus $y_{ss} = y_{dd}$, $y_{sg} = y_{dg}$, and $y_{sd} = y_{ds}$. Using such equal values in Fig. 8.16 makes the symmetry evident. Nevertheless, small-signal models using the substrate as the reference are not in much use for high-frequency work. The relative merits of using the source or the substrate as a reference are discussed in Sec. 4.10.

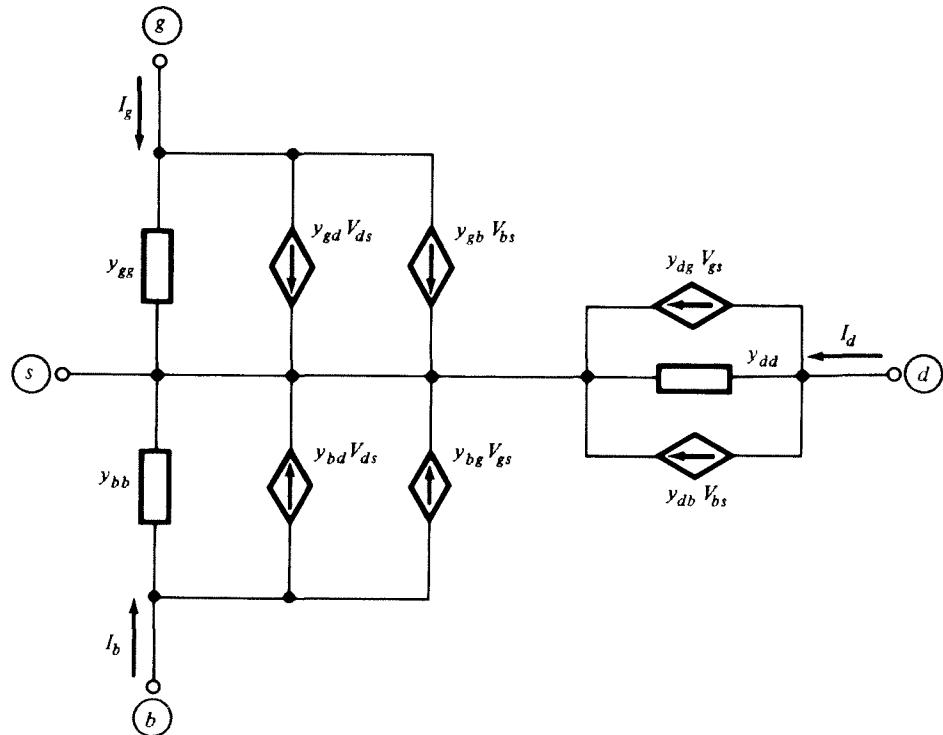


FIGURE 8.15
A general y -parameter model using the source terminal as a reference.

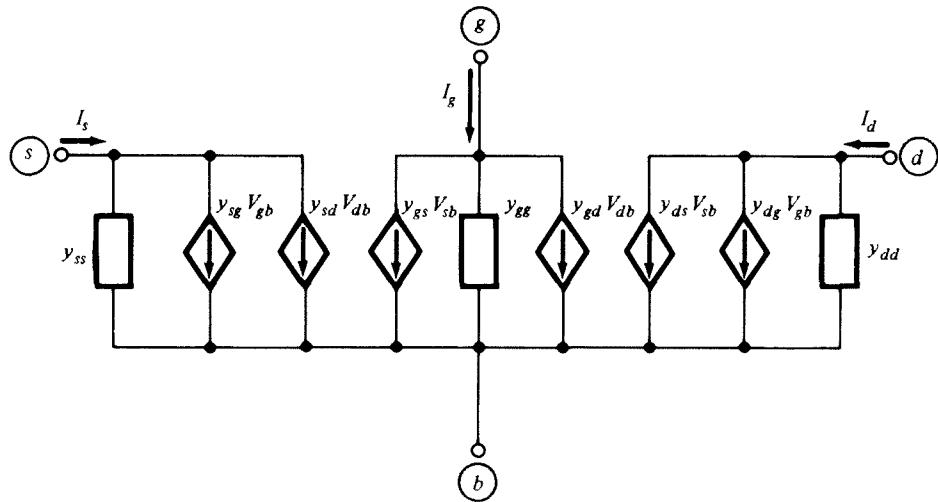


FIGURE 8.16
A general y -parameter model using the substrate terminal as a reference.

We now go one step closer toward relating our present models to the one in Fig. 8.5. Following an approach similar to the one that led to (8.2.19), we can rewrite (8.3.8) as follows (Prob. 8.12):

$$I_d = -y_{gd}V_{dg} - y_{sd}V_{ds} - y_{bd}V_{db} + y_mV_{gs} + y_{mb}V_{bs} \quad (8.3.9a)$$

$$I_g = -y_{gd}V_{gd} - y_{gb}V_{gb} - y_{gs}V_{gs} \quad (8.3.9b)$$

$$I_b = -y_{bd}V_{bd} - y_{gb}V_{bg} + y_{mx}V_{gb} - y_{bs}V_{bs} \quad (8.3.9c)$$

where

$$y_m = y_{dg} - y_{gd} \quad (8.3.10a)$$

$$y_{mb} = y_{db} - y_{bd} \quad (8.3.10b)$$

$$y_{mx} = y_{bg} - y_{gb} \quad (8.3.10c)$$

These equations can be represented by the circuit of Fig. 8.17, as can be verified by writing Kirchhoff's current law for nodes d , g , and b .[†] Note that this topology has the same form as our low-frequency model in Fig. 7.6. If desired, the dc small-signal conductances in Fig. 7.6 can be included in the admittances in Fig. 8.17; we will not do so for simplicity. However, the general derivation of the model in Fig. 8.17 should convince the reader that this model can, in principle, represent *any* small-signal effect (in fact, for *any* four-terminal device), provided that the parameters in it are given appropriate values.

Because the development of the model in Fig. 8.17 was general, the model in Fig. 8.5 should simply be a special case of it. By comparing the two circuits we get,[‡] *for this special case*,

$$-y_{gd} = j\omega C_{gd} \quad (8.3.11a)$$

$$-y_{gs} = j\omega C_{gs} \quad (8.3.11b)$$

$$-y_{bd} = j\omega C_{bd} \quad (8.3.11c)$$

$$-y_{bs} = j\omega C_{bs} \quad (8.3.11d)$$

[†] The minus signs in the two-terminal admittances in the figure are no mystery; they can be traced to the fact that, in the classical definition of y parameters in (8.3.5) and Fig. 8.14, currents are defined in the direction entering the device, rather than leaving it. Consider an analogous situation of a two-terminal admittance y between a point a driven by a voltage V_a , and a point b , which is grounded. The current *entering* node b will be $I_b = -yV_a$; thus if we define y_{ba} as I_b/V_a , we have $y_{ba} = -y$, and the two-terminal admittance y can be labeled as $-y_{ba}$. These minus signs do not imply that values are necessarily negative; for example, as we will see, at low frequencies, y_{sd} turns out to have the value $-g_{sd}$ and thus the two-terminal box between s and d in Fig. 8.17 has an admittance g_{sd} which is in agreement with Fig. 7.19.

[‡] To arrive at (8.3.11), we use the following fact:⁵⁷⁻⁵⁹ The time domain i - v equation for a capacitor, namely $i(t) = C dv(t)/dt$, corresponds to the phasor equation $I = j\omega CV$, where $j\omega C$ is the admittance corresponding to the capacitance C .

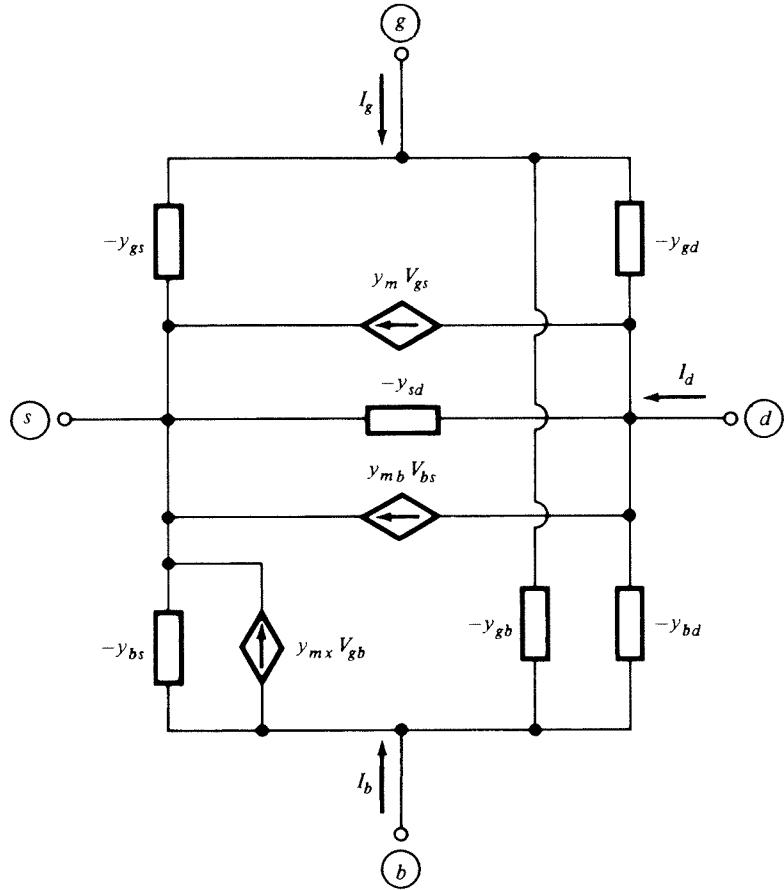


FIGURE 8.17
A general y -parameter model.³⁹

$$-y_{gb} = j\omega C_{gb} \quad (8.3.11e)$$

$$-y_{sd} = g_{sd} + j\omega C_{sd} \quad (8.3.11f)$$

$$y_m = g_m - j\omega C_m \quad (8.3.11g)$$

$$y_{mb} = g_{mb} - j\omega C_{mb} \quad (8.3.11h)$$

$$y_{mx} = -j\omega C_{mx} \quad (8.3.11i)$$

In these equations, observe that: (1) y_m , y_{mb} , and $-y_{sd}$ have a constant, real, positive part, and (2) all y parameters have imaginary parts that are proportional to frequency. These two observations are verified by measurements at frequencies

up to about $\omega_o/3$, with ω_o as given by (7.3.6). Indeed, the capacitance values can be determined as the constant of proportionality in the imaginary part of measured admittances.^{32,38} At frequencies higher than $\omega_o/3$, however, the behavior predicted above fails. Measurements then show, for example, that both the real and the imaginary part of y_m decrease in magnitude with frequency, and that y_{gs} begins to have a nonzero real part. To explain such phenomena, the quasi-static assumption must be abandoned. This is what will be done in the following section.

8.4 Non-Quasi-Static Models

8.4.1 Introduction

In this section, the quasi-static assumption will be dropped and the dynamics of the channel charge at high frequencies will be investigated in detail. Before starting a mathematical analysis, we present a short discussion as to what one might expect intuitively at frequencies where quasi-static behavior is no longer observed.

Let us consider the intrinsic transistor in Fig. 8.18. We assume that, in general, each terminal voltage consists of a dc bias and a small-signal sinusoidal component, as shown. Assume that only one small-signal voltage is nonzero at a time. If v_s is varying

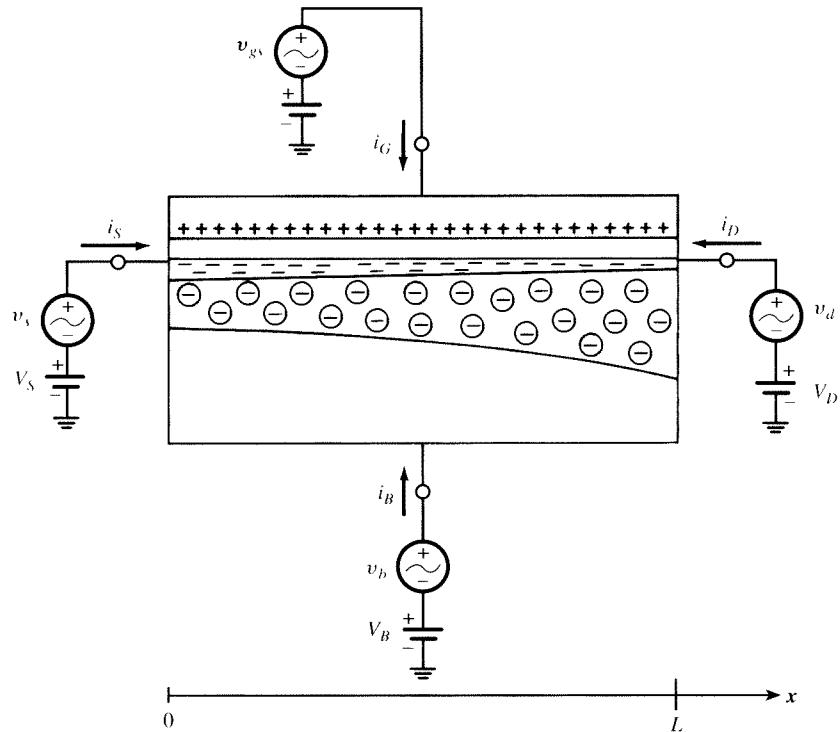


FIGURE 8.18
Intrinsic transistor with bias and small-signal voltages.

very slowly, the inversion layer charge has time to follow with practically no delay. We have seen that the resulting effect on the gate can be modeled by connecting a capacitance C_{gs} from source to gate. However, if the variation of v_s is fast, the “inertia” of the inversion layer becomes nonnegligible, and the effect (gate charge change) will lag behind the cause (source voltage change). A similar effect will be observed between drain and gate. Also, similar conclusions can be drawn for the effect of the source and of the drain on the substrate (the “back gate”). Consider now the effect of the gate voltage. If v_g is varying very fast, the inversion layer charge does not have enough time to respond fully, and thus $|y_{dg}|$, which models this response, will be small. Also, the angle of y_{dg} should be significant and negative, because of the delay between the cause (the variation in the gate voltage) and the effect (the variation in the drain current). Finally, similar observations hold for the effect of the substrate voltage on the inversion layer charge. Measurements⁵¹ on both long- and short-channel devices do, in fact, verify that such effects can influence transistor behavior.

The effects will be observed if the operating frequency exceeds the upper limit of validity for quasi-static modeling. From Sec. 8.2.4, this upper limit is proportional to ω_o , which in turn is proportional to $1/L^2$ in the absence of velocity saturation. One way to model the transistor at frequencies above the limit is, in principle, to split the device into sections (Fig. 6.12). The length of each section is chosen such that, for *it*, a quasi-static model *can* be used. The combination of the models of all sections will then be a valid model for the whole transistor at the frequency of interest. The higher this frequency, the shorter the length needed for each section. One is thus led to carrying the idea to the limit, i.e., considering elemental sections the lengths of which are allowed to approach zero. The following analysis corresponds mathematically to this idea. The analysis is based on well-established principles,^{1,2,4,17,20,21,35} but is generalized to view the transistor as a four-terminal device.³⁹ We want to establish all steps carefully, so the procedure will be rather long; once again we ask for the reader’s patience.

8.4.2 A Non-Quasi-Static Strong-Inversion Model

In this section, we will derive a useful high-frequency model corresponding to the simplified strong-inversion dc model. To end up with manageable expressions, we will use here the same simplifying assumptions we used in Sec. 7.3.³⁹ The first assumption is that, in the expressions for the charge, the quantity α will be assumed to be given by

$$\alpha = \alpha_1 = 1 + \frac{\gamma}{2\sqrt{\phi_0 + V_{SB}}} \quad (8.4.1)$$

The second assumption is that the derivative of α_1 with V_S or V_B is negligible for our purposes, and thus, α_1 will be treated as a constant during differentiation. At a later point, we will verify that these assumptions do indeed produce a useful model by comparison to more exact models.

Many of the expressions we have developed in previous chapters will be needed here. For convenience, we will repeat these in the analysis to follow. We will

take this opportunity and present the various expressions in an organized fashion, so as to present a complete picture of, and the relation between, the following cases:

1. DC (bias) excitation.
2. Time-varying excitation.
3. The special case of (2), where the time variations are small signals.
4. A special case of (3), where the small signals have a form particularly useful for high-frequency model development.

DC (Bias) Excitation In our analysis, we will find it convenient to express the gate, depletion region, and inversion layer charges per unit area in terms of $V_{GS} = V_{GB} - V_{SB}$ and $V_{CS}(x) = V_{CB}(x) - V_{SB}$. Recall that, in strong inversion, $V_{CB}(x)$ can be viewed as the effective reverse bias between the inversion layer at point x and the substrate. Thus, $V_{CS}(x)$ represents the potential drop across the part of the inversion layer contained between point x and the source region. Using the preceding in (4.7.17) and (4.7.16), using (8.4.1), employing the charge conservation equation $Q'_G + Q'_o + Q'_I + Q'_B = 0$, and recalling (6.2.3), we have the following results.

The gate charge per unit area and the total gate charge are, respectively,

$$Q'_G(x) = C'_{ox}[V_{GS} - V_{FB} - \phi_0 - V_{CS}(x)] - Q'_o \quad (8.4.2)$$

$$Q_G = W \int_0^L Q'_G(x) dx \quad (8.4.3)$$

The corresponding charges for the depletion region are

$$Q'_B(x) = -C'_{ox} \left[\gamma \sqrt{\phi_0 - V_{BS}} + (\alpha_l - 1)V_{CS}(x) \right] \quad (8.4.4)$$

$$Q_B = W \int_0^L Q'_B(x) dx \quad (8.4.5)$$

The inversion layer charge per unit area can be written in the form

$$Q'_I(x) = -C'_{ox} U_I(x) \quad (8.4.6)$$

where

$$U_I(x) = V_{GS} - V_{FB} - \phi_0 - \gamma \sqrt{\phi_0 - V_{BS}} - \alpha_l V_{CS}(x) \quad (8.4.7)$$

The current at point x in the channel will be denoted by $I_I(x)$. This current is given by the right-hand side of (4.7.5). Using $V_{CS}(x) = V_{CB}(x) - V_{SB}$, we have

$$I_I(x) = -\mu W Q'_I(x) \frac{dV_{CS}(x)}{dx} \quad (8.4.8)$$

which, from (8.4.7), can be written as

$$I_I(x) = \frac{1}{\alpha_l} \mu W Q'_I(x) \frac{dU_I(x)}{dx} \quad (8.4.9)$$

and, since at dc, the current is the same everywhere along the channel,

$$I_D = I_I(x) \quad (8.4.10)$$

Using (8.4.10) and (8.4.6) in (8.4.9), and integrating both sides from x to L , we get

$$I_D = \frac{W}{L-x} \frac{\mu C'_\text{ox}}{2\alpha_1} [U_I^2(x) - U_I^2(L)] \quad (8.4.11)$$

which for $x = 0$ gives

$$I_D = \frac{W}{L} \frac{\mu C'_\text{ox}}{2\alpha_1} [U_I^2(0) - U_I^2(L)] \quad (8.4.12)$$

From the previous two equations, we can solve for $U_I(x)$:

$$U_I(x) = \left\{ U_I^2(0) + \frac{x}{L} [U_I^2(L) - U_I^2(0)] \right\}^{1/2} \quad (8.4.13)$$

At the source end of the channel, we have

$$V_{CS}(0) = 0 \quad (8.4.14a)$$

thus

$$U_I(0) = V_{GS} - V_{FB} - \phi_0 - \gamma \sqrt{\phi_0 - V_{BS}} \quad (8.4.14b)$$

At the drain end of the channel we have

$$\begin{aligned} V_{CS}(L) &= V_{DS}, \quad V_{DS} \leq V'_{DS} \\ &= V'_{DS}, \quad V_{DS} > V'_{DS} \end{aligned} \quad (8.4.15a)$$

thus,

$$U_I(L) = V_{GS} - V_{FB} - \phi_0 - \gamma \sqrt{\phi_0 - V_{BS}} - \alpha_1 V_{DS}, \quad V_{DS} \leq V'_{DS} \quad (8.4.15b)$$

$$= V_{GS} - V_{FB} - \phi_0 - \gamma \sqrt{\phi_0 - V_{BS}} - \alpha_1 V'_{DS}, \quad V_{DS} > V'_{DS} \quad (8.4.15c)$$

Using the relations for $U_I(0)$ and $U_I(L)$, it is easy to verify that (8.4.12) is identical to the simplified strong-inversion model equation (4.7.24) (with $\alpha = \alpha_1$; this restriction will be removed later). Similarly, (8.4.13) is equivalent to the equation for the potential distribution corresponding to the simplified strong-inversion model, given by (4.7.33). The form we use for these equations here will be found convenient later in this section.

As usual, for the gate and substrate currents under dc excitation we assume

$$I_G = 0 \quad (8.4.16)$$

$$I_B = 0 \quad (8.4.17)$$

Time-Varying Excitation We will denote total (large-signal) quantities by lowercase symbols with capital subscripts. We show in the following how equations (8.4.2) to (8.4.7) have to be modified for the case of time-varying voltages (time dependence is explicitly indicated throughout):

$$q'_G(x, t) = C'_{\text{ox}}[v_{GS}(t) - V_{FB} - \phi_0 - v_{CS}(x, t)] - Q'_o \quad (8.4.18)$$

where Q'_o is assumed fixed.

$$q_G(t) = W \int_0^L q'_G(x, t) dx \quad (8.4.19)$$

$$q'_B(x, t) = -C'_{\text{ox}} \left[\gamma \sqrt{\phi_0 - v_{BS}(t)} + (\alpha_1 - 1) v_{CS}(x, t) \right] \quad (8.4.20)$$

$$q_B(t) = W \int_0^L q'_B(x, t) dx \quad (8.4.21)$$

$$q'_I(x, t) = -C'_{\text{ox}} u_I(x, t) \quad (8.4.22)$$

where

$$u_I(x, t) = v_{GS}(t) - V_{FB} - \phi_0 - \gamma \sqrt{\phi_0 - v_{BS}(t)} - \alpha_1 v_{CS}(x, t) \quad (8.4.23)$$

Equation (8.4.9) must be replaced by

$$i_I(x, t) = \frac{\mu W}{\alpha_1} q'_I(x, t) \frac{\partial u_I(x, t)}{\partial x} \quad (8.4.24)$$

Since we will allow fast variations, (8.4.10) does not have a corresponding equation in the present case. Instead, we must consider the “continuity equation” (6.7.5). That equation, from (8.4.22), can be written as

$$\frac{\partial i_I(x, t)}{\partial x} = -C'_{\text{ox}} W \frac{\partial u_I(x, t)}{\partial t} \quad (8.4.25)$$

Note that for the special case of no variation with time, this equation is consistent with (8.4.10).

The terminal currents now are given by

$$i_D(t) = i_I(L, t) \quad (8.4.26)$$

$$i_G(t) = \frac{dq_G(t)}{dt} \quad (8.4.27)$$

$$i_B(t) = \frac{dq_B(t)}{dt} \quad (8.4.28)$$

Small-Signal Excitation We now assume that the total terminal voltages are of the following form:

$$v_{GS}(t) = V_{GS} + v_{gs}(t) \quad (8.4.29a)$$

$$v_{BS}(t) = V_{BS} + v_{bs}(t) \quad (8.4.29b)$$

$$v_{DS}(t) = V_{DS} + v_{ds}(t) \quad (8.4.29c)$$

where in each right-hand side the first term is a dc bias quantity, and the second term is a small-signal excitation (all small-increment quantities will be represented by lowercase subscripts). As a result of the above form of the voltages, we will have for the other time-varying quantities in our analysis:

$$q'_G(x, t) = Q'_G(x) + q'_g(x, t) \quad (8.4.30a)$$

$$q_G(t) = Q_G + q_g(t) \quad (8.4.30b)$$

$$q'_B(x, t) = Q'_B(x) + q'_b(x, t) \quad (8.4.30c)$$

$$q_B(t) = Q_B + q_b(t) \quad (8.4.30d)$$

$$u_I(x, t) = U_I(x) + u_i(x, t) \quad (8.4.30e)$$

$$v_{CS}(x, t) = V_{CS}(x) + v_{cs}(x, t) \quad (8.4.30f)$$

$$i_I(x, t) = I_I(x) + i_i(x, t) \quad (8.4.30g)$$

$$i_D(t) = I_D + i_d(t) \quad (8.4.30h)$$

$$i_G(t) = I_G + i_g(t) \quad (8.4.30i)$$

$$i_B(t) = I_B + i_b(t) \quad (8.4.30j)$$

Assuming $v_{gs}(t)$, $v_{bs}(t)$, and $v_{ds}(t)$ are very small, all quantities denoted above with capital letters and capital subscripts will be the same as under dc excitation; values for them have already been found. Using quantities from (8.4.29) and (8.4.30) in (8.4.18) to (8.4.28), one can split the resulting expressions into a “bias” and a small-increment part. For example, using (8.4.29a), (8.4.30a), and (8.4.30f) in (8.4.18), we have

$$Q'_G(x) + q'_g(x, t) = C'_{ox}[V_{GS} + v_{gs}(t) - V_{FB} - \phi_0 - V_{CS}(x) - v_{cs}(x, t)] - Q'_o$$

$$\begin{aligned}
&= \{C'_{ox}[V_{GS} - V_{FB} - \phi_0 - V_{CS}(x)] - Q'_o\} \\
&\quad + C'_{ox}[\nu_{gs}(t) - \nu_{cs}(x, t)]
\end{aligned} \tag{8.4.31}$$

The quantity in braces, { }, can be recognized as $Q'_G(x)$ from (8.4.2). Hence (8.4.31) gives

$$q'_g(x, t) = C'_{ox}[\nu_{gs}(t) - \nu_{cs}(x, t)] \tag{8.4.32}$$

Similarly, using (8.4.30b) in (8.4.19), gives

$$\begin{aligned}
Q_G + q_g(t) &= W \int_0^L [Q'_G(x) + q'_g(x, t)] dx \\
&= W \int_0^L Q'_G(x) dx + W \int_0^L q'_g(x, t) dx
\end{aligned} \tag{8.4.33}$$

which, using (8.4.3), gives

$$q_g(t) = W \int_0^L q'_g(x, t) dx \tag{8.4.34}$$

To derive similar expressions for the depletion region charges, we first note that the term containing the square root in (8.4.20) becomes, from (8.4.29b),

$$\gamma \sqrt{\phi_0 - v_{BS}(t)} = \gamma \sqrt{(\phi_0 - V_{BS}) - v_{bs}(t)} \tag{8.4.35}$$

Since $v_{bs}(t)$ is small, the right-hand side can be approximated by the first two terms of a series expansion. This gives

$$\gamma \sqrt{\phi_0 - v_{BS}(t)} = \gamma \sqrt{\phi_0 - V_{BS}} - (\alpha_1 - 1)v_{bs}(t) \tag{8.4.36}$$

with α_1 as given by (8.4.1). Using the preceding in (8.4.20) and proceeding as before, we obtain, for the small-increment part of $q'_B(t)$:

$$q'_b(x, t) = (\alpha_1 - 1) C'_{ox}[\nu_{bs}(t) - \nu_{cs}(x, t)] \tag{8.4.37}$$

Following the approach that led to (8.4.34), for $q_b(t)$ we get

$$q_b(t) = W \int_0^L q'_b(x, t) dx \tag{8.4.38}$$

For $u_i(x, t)$, starting from (8.4.23), using (8.4.36), and proceeding as before, we easily get

$$u_i(x, t) = \nu_{gs}(t) + (\alpha_1 - 1) \nu_{bs}(t) - \alpha_1 \nu_{cs}(x, t) \tag{8.4.39a}$$

$$= [v_{gs}(t) - v_{cs}(x, t)] + (\alpha_1 - 1)[v_{bs}(t) - v_{cs}(x, t)] \quad (8.4.39b)$$

Since $v_{cs}(x, t)$ is zero at the source end and equal to $v_{ds}(t)$ at the drain end, we have, from (8.4.39),

$$u_i(0, t) = v_{gs}(t) + (\alpha_1 - 1)v_{bs}(t) \quad (8.4.40)$$

$$u_i(L, t) = [v_{gs}(t) - v_{ds}(t)] + (\alpha_1 - 1)[v_{bs}(t) - v_{ds}(t)] \quad (8.4.41)$$

The small-signal quantity $i_i(x, t)$ can be obtained by starting from (8.4.24). Using the fact that $u_i(x, t)$ is very small, the result can be put in the form (Prob. 8.13)

$$i_i(x, t) = -\frac{\mu W C'_{\text{ox}}}{\alpha_1} \frac{\partial}{\partial x} [U_I(x) u_i(x, t)] \quad (8.4.42)$$

Starting from (8.4.25), we get, using the facts that $\partial I_I(x)/\partial x = 0$ and $\partial U_I(x)/\partial t = 0$,

$$\frac{\partial i_i(x, t)}{\partial x} = -C'_{\text{ox}} W \frac{\partial u_i(x, t)}{\partial t} \quad (8.4.43)$$

For the drain small-signal current we have, starting from (8.4.26),

$$i_d(t) = i_i(L, t) \quad (8.4.44)$$

For the gate small-signal current, starting from (8.4.27), we obtain

$$i_g(t) = \frac{dq_g(t)}{dt} \quad (8.4.45)$$

We can now use (8.4.34) with (8.4.32) in this equation. If in the result we substitute $v_{cs}(x, t)$ as it ensues from solving (8.4.39), we obtain

$$i_g(t) = WC'_{\text{ox}} \frac{d}{dt} \int_0^L \left\{ \frac{\alpha_1 - 1}{\alpha_1} [v_{gs}(t) - v_{bs}(t)] + \frac{1}{\alpha_1} u_i(x, t) \right\} dx \quad (8.4.46)$$

For the substrate small-signal current we have, starting from (8.4.28),

$$i_b(t) = \frac{dq_b(t)}{dt} \quad (8.4.47)$$

Using in this equation (8.4.38) and (8.4.37), and substituting $v_{cs}(x, t)$ from (8.4.39), we obtain

$$i_b(t) = (\alpha_1 - 1) W C_{ox}' \frac{d}{dt} \int_0^L \left\{ \frac{1}{\alpha_1} [v_{bs}(t) - v_{gs}(t)] + \frac{1}{\alpha_1} u_i(x, t) \right\} dx \quad (8.4.48)$$

It is clear from the preceding expression that to find any of the terminal currents one needs an expression for $u_i(x, t)$, and that must be found from (8.4.42) and (8.4.43). The result depends on the form of the terminal small-signal voltages through the boundary conditions (8.4.40) and (8.4.41). In what follows, we will present the solution when the terminal voltages assume a form of special interest.

Complex Exponential Excitation One could assume that the small-signal voltages are sinusoids and consider the corresponding small-signal terminal currents in the sinusoidal steady state. However, the algebra turns out to be unnecessarily complicated. We will thus follow instead a standard practice and consider a fictitious complex exponential excitation^{57–59} of the form

$$v_{gs}(t) = V_{gs} e^{j\omega t} \quad (8.4.49a)$$

$$v_{bs}(t) = V_{bs} e^{j\omega t} \quad (8.4.49b)$$

$$v_{ds}(t) = V_{ds} e^{j\omega t} \quad (8.4.49c)$$

where capital symbols with lowercase subscripts denote time-independent phasor quantities that can, in general, be complex, and ω is the angular frequency (in rad/s). Since the equations relating the various *small-signal* quantities (derived above) are linear, each small-signal quantity that results as an effect of the excitations in (8.4.49) in the steady state will also be equal to a complex quantity times $e^{j\omega t}$. In particular, we can write

$$u_i(x, t) = U_i(x, \omega) e^{j\omega t} \quad (8.4.50a)$$

$$i_i(x, t) = I_i(x, \omega) e^{j\omega t} \quad (8.4.50b)$$

$$i_d(t) = I_d(\omega) e^{j\omega t} \quad (8.4.50c)$$

$$i_g(t) = I_g(\omega) e^{j\omega t} \quad (8.4.50d)$$

$$i_b(t) = I_b(\omega) e^{j\omega t} \quad (8.4.50e)$$

Equations (8.4.50) are the various “responses” to the excitations (8.4.49). All these *complex* excitations and responses are fictitious, but they are useful for the following reason.^{57–59} If the *real* part of the excitations is used instead to drive the device, all responses in the steady state will be given by the *real* part of the fictitious responses in (8.4.50). Now, the real part of any of the excitations is a sinusoid. [For example, if M and ϕ are the magnitude and phase of V_{gs} , respectively, then the real part of $v_{gs}(t)$ in (8.4.49a) is simply $M \cos(\omega t + \phi)$.] Thus, working with these fictitious exponential functions provides all useful information about the actual

sinusoidal steady state, with real excitations and real responses, only with greater mathematical ease. In addition, if the response to complex exponentials is known, the response to other types of waveforms can be determined by using transform techniques.⁵⁷⁻⁵⁹

The quantities in (8.4.49) and (8.4.50) can now be substituted into (8.4.40), (8.4.41), (8.4.42), (8.4.43), (8.4.44), (8.4.46), and (8.4.48). In all cases, $e^{j\omega t}$ appears as a common factor on both sides. Thus, we easily obtain (Prob. 8.14)

$$I_i(x, \omega) = -\frac{\mu W C'_{\text{ox}}}{\alpha_1} \frac{\partial}{\partial x} [U_i(x) U_i(x, \omega)] \quad (8.4.51a)$$

$$\frac{\partial I_i(x, \omega)}{\partial x} = -j\omega C'_{\text{ox}} W U_i(x, \omega) \quad (8.4.51b)$$

$$U_i(0, \omega) = V_{gs} + (\alpha_1 - 1)V_{bs} \quad (8.4.52a)$$

$$U_i(L, \omega) = (V_{gs} - V_{ds}) + (\alpha_1 - 1)(V_{bs} - V_{ds}) \quad (8.4.52b)$$

$$I_d(\omega) = I_i(L, \omega) \quad (8.4.53a)$$

$$I_g(\omega) = j\omega C'_{\text{ox}} W \left[L \frac{\alpha_1 - 1}{\alpha_1} (V_{gs} - V_{bs}) + \frac{1}{\alpha_1} \int_0^L U_i(x, \omega) dx \right] \quad (8.4.53b)$$

$$I_b(\omega) = j\omega (\alpha_1 - 1) C'_{\text{ox}} W \left[L \frac{1}{\alpha_1} (V_{bs} - V_{gs}) + \frac{1}{\alpha_1} \int_0^L U_i(x, \omega) dx \right] \quad (8.4.53c)$$

In the preceding equations, note the following: μ , W , L , and C'_{ox} are known device parameters, α_1 is known for a given bias V_{SB} from (8.4.1), and $U_i(x)$ is a known function of x , from (8.4.13). V_{gs} , V_{bs} , and V_{ds} are known phasors representing the excitation. Thus, for a given ω , (8.4.51) is a system of two differential equations in two unknown functions, $I_i(x, \omega)$ and $U_i(x, \omega)$. This system can be solved by using Bessel or Kelvin functions, with the boundary conditions given in (8.4.52);^{1,7,17,21,35,39,45,52,55} an alternate solution uses symbolic iterative techniques.²⁰ Once the functions $U_i(x, \omega)$ and $I_i(x, \omega)$ have been determined, they can be substituted in (8.4.53) to give $I_d(\omega)$, $I_g(\omega)$, and $I_b(\omega)$. The mathematical details are long (the procedure is outlined in the statement of Prob. 8.15) and will not be presented here. In the following, we summarize the form of the results. The final expressions are in the form

$$I_d(\omega) = \frac{N_{dd}(\omega)V_{ds} + N_{dg}(\omega)V_{gs} + N_{db}(\omega)V_{bs}}{D(\omega)} \quad (8.4.54)$$

$$I_g(\omega) = \frac{N_{gd}(\omega)V_{ds} + N_{gg}(\omega)V_{gs} + N_{gb}(\omega)V_{bs}}{D(\omega)} \quad (8.4.55)$$

$$I_b(\omega) = \frac{N_{bd}(\omega)V_{ds} + N_{bg}(\omega)V_{gs} + N_{bb}(\omega)V_{bs}}{D(\omega)} \quad (8.4.56)$$

where the quantities $N_{kl}(\omega)$ ($k, l = d, g, b$) and $D(\omega)$ are infinite series in $j\omega$:

$$N_{kl}(\omega) = n_{kl0} + (j\omega)n_{kl1} + (j\omega)^2n_{kl2} + \dots \quad (8.4.57a)$$

$$D(\omega) = d_0 + (j\omega)d_1 + (j\omega)^2d_2 + \dots \quad (8.4.57b)$$

The coefficients in these series up to second order as well as for all $N_{kl}(\omega)$ in (8.4.54) to (8.4.56), are given in Appendix I. One is finally able to find the y parameters by comparing (8.4.54) to (8.4.56) with (8.3.8):

$$y_{dd} = \frac{N_{dd}(\omega)}{D(\omega)}, \quad y_{dg} = \frac{N_{dg}(\omega)}{D(\omega)}, \quad y_{db} = \frac{N_{db}(\omega)}{D(\omega)} \quad (8.4.58)$$

$$y_{gd} = \frac{N_{gd}(\omega)}{D(\omega)}, \quad y_{gg} = \frac{N_{gg}(\omega)}{D(\omega)}, \quad y_{gb} = \frac{N_{gb}(\omega)}{D(\omega)} \quad (8.4.59)$$

$$y_{bd} = \frac{N_{bd}(\omega)}{D(\omega)}, \quad y_{bg} = \frac{N_{bg}(\omega)}{D(\omega)}, \quad y_{bb} = \frac{N_{bb}(\omega)}{D(\omega)} \quad (8.4.60)$$

For example, using (8.4.57) in (8.4.59), we have

$$y_{gd} = \frac{n_{gd0} + (j\omega)n_{gd1} + (j\omega)^2n_{gd2} + \dots}{d_0 + (j\omega)d_1 + (j\omega)^2d_2 + \dots} \quad (8.4.61)$$

The y parameters can be computed, for a given frequency, to any desired accuracy by keeping an appropriate number of terms in the numerator and the denominator. The values thus obtained can be substituted in the small-signal equivalent circuit of Fig. 8.15.

Consider now the equivalent circuit in Fig. 8.17. In this circuit, only three of the parameters determined above appear directly: y_{gd} , y_{gb} , and y_{bd} . The rest of the parameters can be trivially found from (8.3.7) and (8.3.10):

$$y_{gs} = -y_{gg} - y_{gd} - y_{gb} \quad (8.4.62a)$$

$$y_{bs} = -y_{bb} - y_{bd} - y_{bg} \quad (8.4.62b)$$

$$y_{sd} = -y_{dd} - y_{gd} - y_{bd} \quad (8.4.62c)$$

$$y_m = y_{dg} - y_{gd} \quad (8.4.62d)$$

$$y_{mb} = y_{db} - y_{bd} \quad (8.4.62e)$$

$$y_{mx} = y_{bg} - y_{gb} \quad (8.4.62f)$$

where the quantities in the right-hand side are given by (8.4.58) to (8.4.60).

We will write the expressions for the parameters of the model in Fig. 8.17 in a way that will help relate the model to the one developed in Sec. 7.3. We start from (8.4.58) to (8.4.60). In each y -parameter expression, we factor out the first nonzero term of the numerator. For example, consider y_{gd} in (8.4.61). From Appendix 1 we have $n_{gd0} = 0$ and $d_0 = 1$. Thus, we can write

$$y_{gd} = j\omega n_{gd1} \frac{1 + j\omega(n_{gd2}/n_{gd1}) + \dots}{1 + j\omega d_1 + \dots} \quad (8.4.63)$$

A look at Appendix I reveals that $-n_{gd1}$ has exactly the same expression as C_{gd} in (7.3.9). Thus we can write

$$y_{gd} = -j\omega C_{gd} \frac{1 + j\omega(n_{gd2}/n_{gd1}) + \dots}{1 + j\omega d_1 + \dots} \quad (8.4.64)$$

Proceeding in a similar manner, we can find expressions for all parameters in Fig. 8.17. As each is being developed, part of the expression can be recognized as a familiar small-signal quantity discussed in Chap. 7.[†] The results are summarized here (minus signs are used in front of the y parameters, corresponding to Fig. 8.17):

$$-y_{gs} = j\omega C_{gs} \frac{1 + j\omega\tau_2 + \dots}{1 + j\omega\tau_1 + \dots} \quad (8.4.65a)$$

$$-y_{bs} = j\omega C_{bs} \frac{1 + j\omega\tau_2 + \dots}{1 + j\omega\tau_1 + \dots} \quad (8.4.65b)$$

$$-y_{gd} = j\omega C_{gd} \frac{1 + j\omega\tau_3 + \dots}{1 + j\omega\tau_1 + \dots} \quad (8.4.65c)$$

$$-y_{bd} = j\omega C_{bd} \frac{1 + j\omega\tau_3 + \dots}{1 + j\omega\tau_1 + \dots} \quad (8.4.65d)$$

$$-y_{gb} = j\omega C_{gb} + \frac{(j\omega)^2 C_{gb,\text{sat}} \tau_4 + \dots}{1 + j\omega\tau_1 + \dots} \quad (8.4.65e)$$

$$-y_{sd} = \frac{g_{sd}}{1 + j\omega\tau_1 + \dots} \quad (8.4.65f)$$

[†] Some of the expressions obtained will be valid with best accuracy at low V_{DS} values because they include α_1 , which can be traced to our initial, simplifying assumption in (8.4.1). This same restriction ($\alpha = \alpha_1$) was encountered at various points in Sec. 7.3 and was subsequently removed. This restriction will soon be removed from the present model also.

$$y_m = \frac{g_m}{1 + j\omega\tau_1 + \dots} \quad (8.4.65g)$$

$$y_{mb} = \frac{g_{mb}}{1 + j\omega\tau_1 + \dots} \quad (8.4.65h)$$

$$y_{mx} = 0 \quad (8.4.65i)$$

where (after lengthy algebra)

$$\tau_1 = \frac{4}{15} \frac{1}{\omega_o} \frac{1 + 3\eta + \eta^2}{(1 + \eta)^3} \quad (8.4.66a)$$

$$\tau_2 = \frac{1}{15} \frac{1}{\omega_o} \frac{2 + 8\eta + 5\eta^2}{(1 + \eta)^2(1 + 2\eta)} \quad (8.4.66b)$$

$$\tau_3 = \frac{1}{15} \frac{1}{\omega_o} \frac{5 + 8\eta + 2\eta^2}{(1 + \eta)^2(2 + \eta)} \quad (8.4.66c)$$

$$\tau_4 = \frac{2}{15} \frac{1}{\omega_o} \frac{2 + 13\eta + 30\eta^2 + 13\eta^3 + 2\eta^4}{(1 + \eta)^5} \quad (8.4.66d)$$

with η as given by (4.7.25) and plotted in Fig. 4.25, and

$$\omega_o = \frac{\mu(V_{GS} - V_T)}{\alpha L^2} \quad (8.4.67)$$

We note that the numerators in (8.4.65f) to (8.4.65h) do not contain frequency-dependent terms. This is because of cancellations that occur when (8.4.62c) to (8.4.62e) are used. For $-y_{gb}$ in (8.4.65e), we did not use the form of the relations immediately before it because, for this parameter, such a form causes numerical difficulties at $\eta = 1$. In the formula given, $C_{gb,sat}$ is the value of C_{gb} in saturation (Sec. 7.3). The preceding approach can be extended to models valid in all regions of operation.^{39,60,61}

Consider now low frequencies, so that $\omega \ll \omega_o$. Then (8.4.65) gives $-y_{gs} \approx j\omega C_{gs}$, $-y_{bs} \approx j\omega C_{bs}$, $-y_{gd} \approx j\omega C_{gd}$, $-y_{bd} \approx j\omega C_{bd}$, $-y_{sd} \approx g_{sd}$, $y_m \approx g_m$, and $y_{mb} \approx g_{mb}$. If, in addition, we use[†] $-y_{gb} \approx j\omega C_{gb}$, and since $y_{mx} = 0$, the model in Fig. 8.17 reduces to the model in Fig. 7.19. Thus the same model results, starting from different premises.

[†] This approximation will be discussed shortly.

The quantity η as it results in the preceding development is given by (4.7.25), where $V'_{DS} = (V_{GS} - V_T)/\alpha$ with $\alpha = \alpha_1$. This value for α is not the best one to use, as explained in Sec. 4.7.3 (unless V_{DS} is small). It simply appears here because of our initial simplifying assumption in (8.4.1), which was necessary in order to obtain manageable results.[†] It is desirable, however, that our present model be consistent with the corresponding dc model in (4.7.24), where the restriction that $\alpha = \alpha_1$ is not used. Thus we will remove this restriction from the present model, and we will allow V'_{DS} to have the same appropriate value as in the dc equations (see the discussion in Sec. 4.7.3).

Similar comments apply to other quantities in (8.4.65) to (8.4.67) (C_b , g_{sd} , etc.). Thus, the approach we have followed leads to expressions for these quantities which are familiar from Chap. 7, only with $\alpha_1 - 1$ in place of a somewhat different quantity. [For example, we find $g_{mb} = (\alpha_1 - 1)g_m$, which agrees with (7.2.34), but which is not accurate for large V_{DS} , as can be seen from (7.2.31) and (7.2.32).] Also, in the saturation region, the value of g_{sd} resulting from the preceding development is zero, since we have not taken channel length modulation into account. All such limitations can be removed by adopting the following approach: *Any quantity in (8.4.65) to (8.4.67), which has already been encountered in Chap. 7, will be assumed to have the value given in that chapter.* In this way, at low frequencies, our present model will reduce to the one in Sec. 7.3, not only in topology (Fig. 7.19) but also in the values for its elements. This is a very desirable property. It makes possible the incorporation of all the refinements known for low-frequency, small-signal parameters into the present model. Thus, the model will be very well behaved at low frequencies (and will also provide useful results at frequencies where low-frequency models fail).

Using (7.3.19) and (8.4.65), we obtain

$$\frac{y_{bs}}{y_{gs}} \approx \frac{y_{bd}}{y_{gd}} \approx \frac{y_{mb}}{y_m} \approx \frac{dV_T}{dV_{SB}} = \alpha_1 - 1, \text{ small } V_{DS} \text{ or small } V_{GS} \quad (8.4.68)$$

We now derive some very useful approximations for the relations in (8.4.65). Consider y_{gs} as an example. If the frequency of operation satisfies $\omega\tau_2 \ll 1$, we can write $1 + j\omega\tau_2 \approx 1/(1 - j\omega\tau_2)$. Using this in (8.4.65a) and neglecting high-order terms, we obtain $-y_{gs} \approx j\omega C_{gs}/[1 + j\omega(\tau_1 - \tau_2)]$ and similarly for $-y_{bs}$, $-y_{gd}$, and $-y_{bd}$. The validity of these approximations will be considered shortly. Thus, we have

$$-y_{gs} \approx \frac{j\omega C_{gs}}{1 + j\omega(\tau_1 - \tau_2)}, \quad \omega\tau_2 \ll 1 \quad (8.4.69a)$$

$$-y_{bs} \approx \frac{j\omega C_{bs}}{1 + j\omega(\tau_1 - \tau_2)}, \quad \omega\tau_2 \ll 1 \quad (8.4.69b)$$

$$-y_{gd} \approx \frac{j\omega C_{gd}}{1 + j\omega(\tau_1 - \tau_3)}, \quad \omega\tau_3 \ll 1 \quad (8.4.69c)$$

[†] The reader may want to try developing the model by starting from a different value of α in order to appreciate the ensuing problems.

$$-y_{bd} \approx \frac{j\omega C_{bd}}{1 + j\omega(\tau_1 - \tau_3)}, \quad \omega\tau_3 \ll 1 \quad (8.4.69d)$$

Consider now $-y_{gb}$ in (8.4.65e). We have

$$-y_{gb} = j\omega C_{gb} + y_a \quad (8.4.69e)$$

where, neglecting high-order terms, y_a is given by

$$y_a \approx (j\omega)^2 \frac{C_{gb,\text{sat}} \tau_4}{1 + j\omega\tau_1} \quad (8.4.69f)$$

In saturation, and at frequencies where the other approximations we have been making are accurate, y_a can be assumed zero with little penalty. In nonsaturation, and especially with very small V_{DS} , this term can be the dominant one in y_{gb} , but the magnitude of y_{gb} is then very small anyway. The small currents that can be contributed by it are almost invariably masked by other larger currents (e.g., those contributed by the extrinsic gate-substrate capacitance). Thus, y_a can be omitted for many applications (Prob. 8.17).

For the rest of the parameters in (8.4.65) we simply drop the high-order terms in the denominators:

$$-y_{sd} \approx \frac{g_{sd}}{1 + j\omega\tau_1}, \quad \omega\tau_1 \ll 1 \quad (8.4.69g)$$

$$y_m \approx \frac{g_m}{1 + j\omega\tau_1}, \quad \omega\tau_1 \ll 1 \quad (8.4.69h)$$

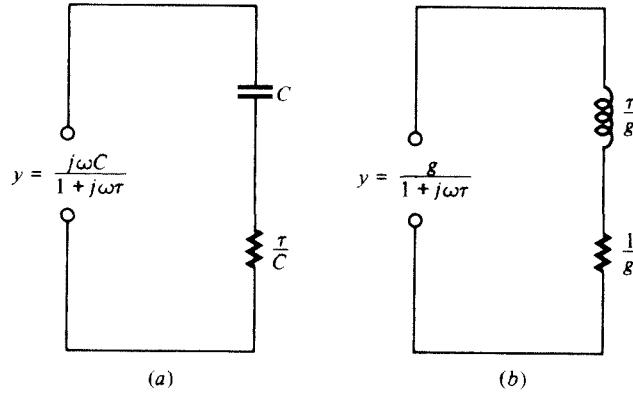
$$y_{mb} \approx \frac{g_{mb}}{1 + j\omega\tau_1}, \quad \omega\tau_1 \ll 1 \quad (8.4.69i)$$

$$y_{mx} = 0 \quad (8.4.69j)$$

The admittances in the right-hand side of (8.4.69a) to (8.4.69d) are of the general form $j\omega C/(1 + j\omega\tau)$. Figure 8.19a shows a simple circuit that realizes such an admittance. Figure 8.19b shows a circuit that realizes the admittance in the right-hand side of (8.4.69g). These circuit representations in Fig. 8.19 can be verified through simple circuit analysis (Prob. 8.18).

With the help of Fig. 8.19, it is easy to see that using (8.4.69) the equivalent circuit of Fig. 8.17 takes the form shown in Fig. 8.20. The box shown in broken lines is y_a in (8.4.69f). In many applications this box can be omitted, for reasons already mentioned. As follows from Fig. 8.19a and (8.4.69a) to (8.4.69d), we have

$$R_{gs}C_{gs} = R_{bs}C_{bs} = \tau_1 - \tau_2 \quad (8.4.70a)$$

**FIGURE 8.19**

Simple circuits used to represent the admittances (a) in (8.4.69a) to (8.4.69d); (b) in (8.4.69g).

$$R_{gd}C_{gd} = R_{bd}C_{bd} = \tau_1 = \tau_3 \quad (8.4.70b)$$

from which the resistance values can be calculated.[†] For the inductor we have, as follows from Fig. 8.19b and (8.4.69g),

$$L_{sd}g_{sd} = \tau_1 \quad (8.4.70c)$$

Plots for the resistances and the inductance are shown in Fig. 8.21.[‡]

Despite the conditions of the form $\omega\tau_i \ll 1$ in (8.4.69), the model just presented is found satisfactory to about $\omega = \omega_o$. A comparison of the various models will be made in Sec. 8.4.4.

The various resistors and the inductor in Fig. 8.20 (in cooperation with the elements in series with them) can be viewed as representing some of the effects of the inversion layer's inertia when responding to rapid changes. In the following discussion, we assume that, in Fig. 8.18, only one small-signal voltage is nonzero at a time. If the source voltage is changing fast, the inversion layer will "hesitate" to respond, and the corresponding changes in the gate and substrate currents will lag behind the source voltage changes; in Fig. 8.20, this is modeled by R_{gs} , C_{gs} and R_{bs} , C_{bs} , respectively. The combinations R_{gd} , C_{gd} and R_{bd} , C_{bd} model the corresponding effects if the drain voltage is changing fast instead (in the nonsaturation region). The combination of L_{sd} and g_{sd} can be viewed as representing the inertia of the inversion layer in changing the source current when a fast-varying

[†] Note that since the capacitances are proportional to $C_{ox} = C'_{ox}WL$ (Sec. 7.3), whereas τ_1 , τ_2 , and τ_3 are inversely proportional to ω_o [see (8.4.66) and (8.4.67)], the resistances will be proportional to $(\omega_o C_{ox})^{-1}$, which is equal to $(L/W)\alpha[\mu C'_{ox}(V_{GS} - V_I)]^{-1}$.

[‡] It is seen that R_{gd} , R_{bd} , and L_{sd} become infinite in saturation, just as do the impedances of the elements in series with them (assuming no CLM). If the model is implemented as part of a circuit simulation program, one should exercise care in order to avoid numerical difficulties.

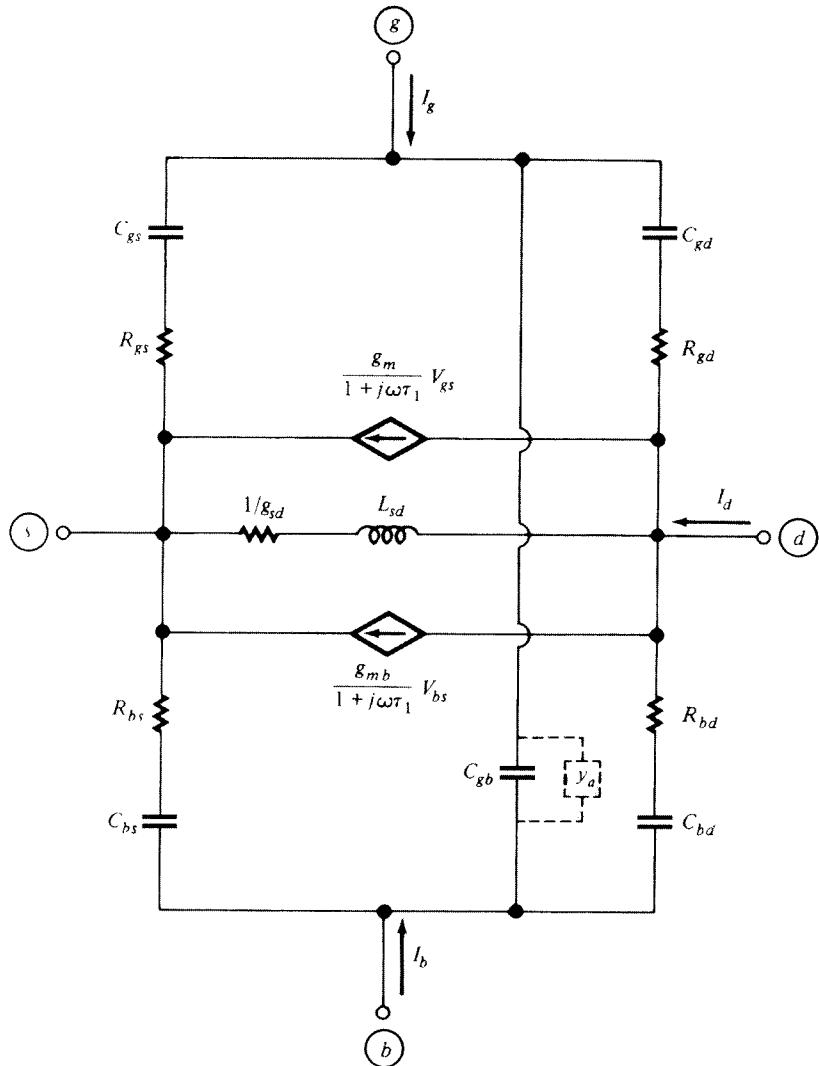
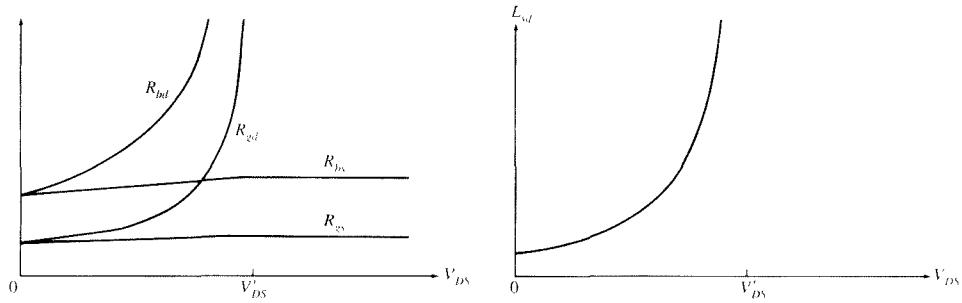


FIGURE 8.20
Small-signal non-quasi-static equivalent circuit model for the intrinsic part of a MOS transistor, corresponding to (8.4.69).

drain voltage demands such a change (in the nonsaturation region); this will be discussed shortly. Finally, the denominators in the values of the two current sources, in Fig. 8.20 model the inertia of the inversion layer in changing the drain current when the gate or the substrate voltages are varying.

The appearance of an inductor in Fig. 8.20 may appear strange. To remove the “mystery” surrounding this element, consider the experiment shown in Fig. 8.22a. A transistor is biased in strong inversion with $V_{DS} = 0$, and a small signal is applied at the drain end of the channel. We chose $V_{DS} = 0$ to simplify things, since with such

**FIGURE 8.21**

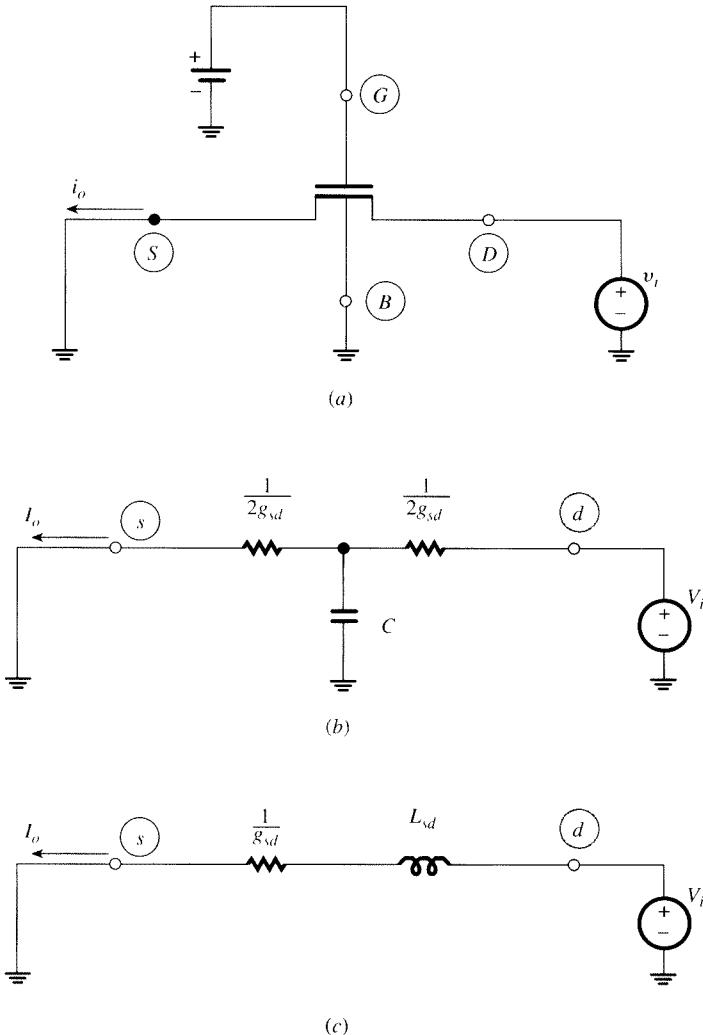
Typical behavior of (a) the resistances R_{gs} , R_{gd} , R_{bs} , and R_{hd} ; (b) the inductance L_{sd} .

bias the channel can be thought of as a uniform resistor. The channel has a parasitic capacitance to the gate and the substrate. Although this capacitance, along with the channel resistance, forms a distributed RC element, to first order we can represent them with the lumped approximation shown in Fig. 8.22b. The capacitance C is shown grounded, since both the gate and the substrate capacitance in (a) are connected to fixed voltages (i.e., “small-signal ground”). The proper value to use for C in this model is less than the total capacitance, since the capacitance close to the ends of the channel is practically shunted out (Prob. 8.26). Let us consider the small-signal current exiting from the source, shown in Fig. 8.22a. This current, and the driving voltage, are represented by complex phasors^{57–59} in Fig. 8.22b. Analysis of this circuit gives

$$I_o = \frac{g_{sd}}{1 + j\omega\tau} V_i \quad (8.4.71)$$

where $\tau = C/(4g_{sd})$. Let us now, instead of using the circuit in Fig. 8.22b, apply the small-signal equivalent circuit of Fig. 8.20 to the case of Fig. 8.22a. The result is shown in Fig. 8.22c, where branches shunted by the voltage sources or by short circuits are omitted, since they do not affect the outcome. The controlled sources are not shown, since at the bias point $V_{DS} = 0$, we have $g_m = g_{mb} = 0$. It is easily seen that an analysis of this circuit gives exactly (8.4.71) again, with $\tau = L_{sd}g_{sd}$ (see Prob. 8.26 for a quantitative comparison). Thus, the presence of the inductor in the circuit in Fig. 8.22c is no mystery: this circuit is just an equivalent circuit that produces the same results for I_o as the circuit in Fig. 8.22b. Notice that the series RC circuits in Fig. 8.20 cannot model the effect we are discussing in relation to Fig. 8.22a, since those circuits are shunted by voltage sources and short circuits and do not affect the value of I_o ; it is the presence of the inductor that provides the proper modeling in this case.[†] Notice

[†] The presence of the inductor in the model, in association with the capacitances in the latter, might suggest the possibility of “overvoltages,” associated with series RLC circuits.^{57,58} This can, in fact, be observed to some extent. Consider a transistor with the drain not connected to anything (and thus with $I_{DS} = 0$ and $V_{DS} = 0$), and driven by a bias and small-signal voltage between gate and source. In this case, the channel is uniform, and for small signals, it behaves as a passive resistance with associated distributed capacitance to the gate. Such

**FIGURE 8.22**

(a) A transistor with zero drain-source bias, and with a small voltage signal at the drain terminal. (b) A lumped model for modeling the current exiting from the source terminal. (c) Part of the model in Fig. 8.20, relevant to modeling the same current.

a device can exhibit, under certain conditions, an ac drain voltage somewhat larger than the input signal for a range of frequencies. This effect can be predicted by the model of Fig. 8.20 (Prob. 8.27) and can be verified by measurements. The effect can even be predicted by using a simple, passive lumped *RC* model of the distributed *RC* behavior of the channel, using a few resistors and capacitors (certain passive *RC* circuits can exhibit a voltage gain somewhat larger than unity, although they cannot exhibit power gain). The interested reader is referred to Prob. 8.27.

from (8.4.71) that the higher the frequency, the smaller the magnitude of I_o , and the larger its phase lag with respect to V_i . This is what one would expect by considering the “inertia” of the channel to respond to the variations of $v_i(t)$. In fact, *small-signal non-quasi-static effects and distributed effects are two ways of looking at the same thing.*

It is clear that, as frequency is increased further and further, the representation of distributed effects by the lumped approximation in Fig. 8.22b will eventually fail; thus, the model in Fig. 8.22c will fail, too, and so will the model of Fig. 8.20 in general. One can extend the frequency limit of validity with higher-order (but more complicated) models,^{39,44,45,48–50,52,55} or by using several sections as in Fig. 6.12.

If the frequency of the small-signal voltages applied to the terminals becomes low enough, the inertia mentioned becomes negligible. Indeed, with decreasing frequency the magnitude of the impedances of C_{gs} , C_{bs} , C_{gd} , and C_{bd} increases. Relative to them, the series resistances in Fig. 8.20 become unimportant and can be omitted. Also, with decreasing frequency the impedance of the inductor decreases, and this element can be omitted in comparison to the resistance in series with it. Finally, the denominators in the current source values become approximately 1 at low frequencies. In this case, the model in Fig. 8.20 is thus seen to reduce to the model of Fig. 7.19.

The model can also be related^{35,41} to the complete quasi-static model of Sec. 8.2. At low frequencies, the series RC combinations in Fig. 8.20 reduce to the corresponding capacitances in Fig. 8.5, for the reasons discussed earlier. In addition, assuming $\omega\tau_1 \ll 1$, we can use the approximation $1/(1+j\omega\tau_1) \approx 1-j\omega\tau_1$, and we can write (8.4.69g) to (8.4.69i) as follows:

$$-y_{sd} \approx g_{sd} - j\omega\tau_1 g_{sd}, \quad \omega\tau_1 \ll 1 \quad (8.4.72a)$$

$$y_m \approx g_m - j\omega\tau_1 g_m, \quad \omega\tau_1 \ll 1 \quad (8.4.72b)$$

$$y_{mb} \approx g_{mb} - j\omega\tau_1 g_{mb}, \quad \omega\tau_1 \ll 1 \quad (8.4.72c)$$

A comparison with $-y_{sd}$, y_m , and y_{mb} for the complete quasi-static model [(8.3.11f) to (8.3.11h)] shows that the form is the same (recall that C_{sd} was a *negative* quantity). Furthermore an examination of the expressions for the various parameters in these equations reveals that $\tau_1 g_{sd} = -C_{sd}$, $\tau_1 g_m = C_m$, and $\tau_1 g_{mb} = C_{mb}$; therefore, (8.4.72a) to (8.4.72c) are in fact *identical* to (8.3.11f) to (8.3.11h), not only in form but even in *value* (Prob. 8.19). Thus, the model of Fig. 8.20 reduces to the complete quasi-static model of Fig. 8.5, assuming C_{mx} is negligible. Note in particular that the series resistance-inductance combination of Fig. 8.20 reduces to a parallel combination of a resistance and a *negative* capacitance. With further reduction in frequency, the terms containing ω in $-y_{sd}$, y_m , and y_{mb} become negligible, and the model reduces to the simple model of Fig. 7.19.

The coefficients of the controlled sources in Fig. 8.20 are complex. This might make it impossible to use this model *directly* in some computer ac analysis programs. The problem is circumvented by noting that we can write

$$\frac{g_m}{1+j\omega\tau_1} V_{gs} = g_m V_1 \quad (8.4.73a)$$

$$\frac{g_{mb}}{1+j\omega\tau_1}V_{bs} = g_{mb}V_2 \quad (8.4.73b)$$

where

$$V_1 = \frac{1}{1+j\omega\tau_1}V_{gs} \quad (8.4.74a)$$

$$V_2 = \frac{1}{1+j\omega\tau_1}V_{bs} \quad (8.4.74b)$$

and that it is very easy to develop V_1 from V_{gs} and V_2 from V_{bs} by using two simple circuits. This idea is exploited in Fig. 8.23, where it is easy to verify that (8.4.74) holds as long as $R_1C_1 = \tau_1$, $g_{n1}R_1 = 1$, $R_2C_2 = \tau_1$, and $g_{n2}R_2 = 1$.[†]

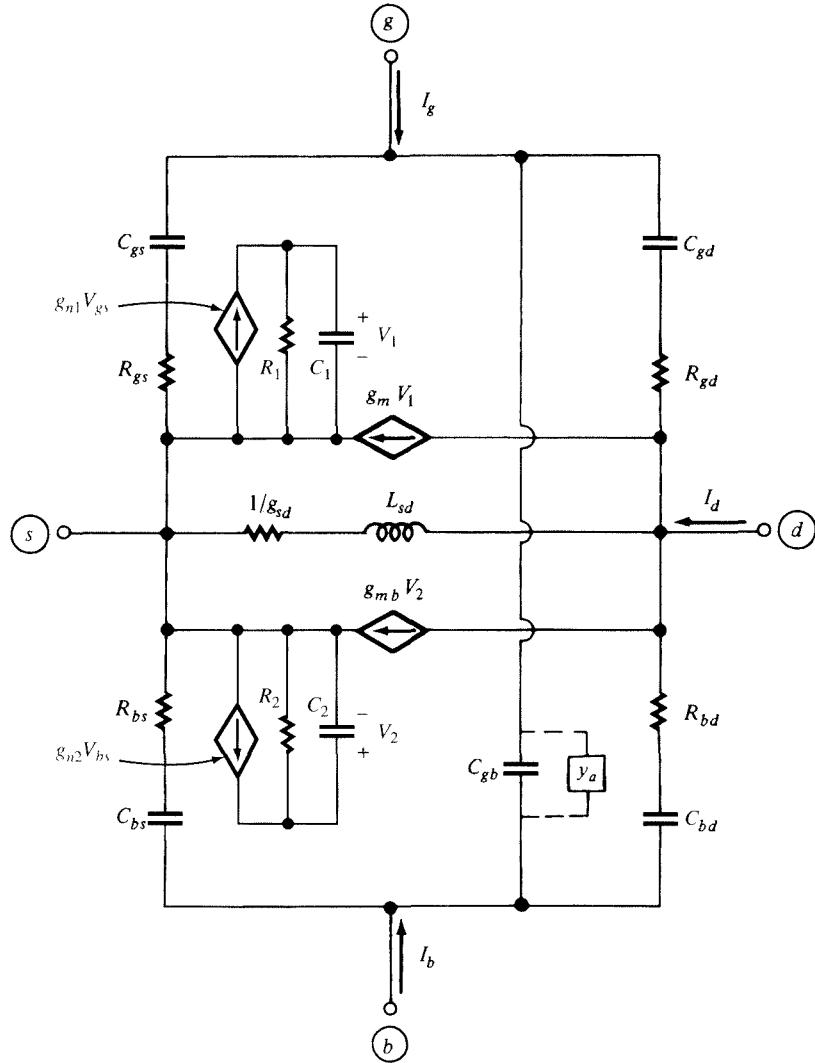
Similarly, it is possible to replace y_a (if it is decided to include this element in the model) by a combination of circuit elements with real values.

In the saturation region, the model of Fig. 8.20 assumes a simple form, easily derived from the relations given previously, and shown in Fig. 8.24. In this region, g_{sd} models CLM (Sec. 5.3). This effect has largely been studied at low frequencies and was not included in the derivations of the present model. Thus, there is no reason to assume that L_{sd} , as it results from (8.4.70c), will be the correct value to use in the saturation region. In fact, in saturation models, L_{sd} in Fig. 8.24 is often replaced by a short circuit. This is not likely to be a problem in practice, since in a circuit capacitances invariably exist between drain and source (e.g., because of another transistor connected to the one under consideration, or parasitics). At high frequencies, where the impedance of L_{sd} could become comparable to $1/g_{sd}$, the impedance of the capacitances is low, and relatively large small-signal currents can flow in them. Compared with these, the minute small-signal current flowing through $1/g_{sd}$ and L_{sd} is likely to be negligible.

8.4.3 Other Approximations and Higher-Order Models

The model we have presented is valid up to about $\omega = \omega_o$, as already mentioned. In developing it, we used a certain type of approximation, e.g., (8.4.65a) was approximated by (8.4.69a). One may wonder why not simply drop the high-order terms in (8.4.65a) and use $-y_{gs} = j\omega C_{gs}(1+j\omega\tau_2)/(1+j\omega\tau_1)$ without any further manipulation. The answer is that such an approach would not be wise for two reasons: (1) The region of validity for a model based on this type of approximation turns out to be about the same as for the one we presented, whereas the equivalent circuit becomes more complicated; and (2) the degradation of such a model with frequency would not be “graceful.” This means that, as the frequency is increased above the limit of validity, the error in some parameters (especially in their phase) soon would become quite large. In contrast to this, the simple model we have presented degrades “gracefully,”

[†] Unfortunately, we could not have used the voltages across C_{gs} and C_{bs} themselves, since $R_{gs}C_{gs} = R_{bs}C_{bs} = \tau_1 - \tau_2$ instead of being equal to the desired τ_1 .

**FIGURE 8.23**

The model of Fig. 8.20 modified to avoid complex coefficients in the controlled current sources.

with the errors increasing rather mildly as the frequency limit of validity is exceeded. The explanation for this somewhat unexpected behavior is as follows. Once the frequency is high enough for the missing high-order terms to become important, the new approximation we are considering is inadequate and large errors can result, since the numerator is of second order whereas the denominator is of first order. In (8.4.69a), the approximation made turns out to be in such a direction as to partially compensate for the effect of the missing terms; hence the more graceful degradation. Graceful degradation of models is important (although one is not supposed to

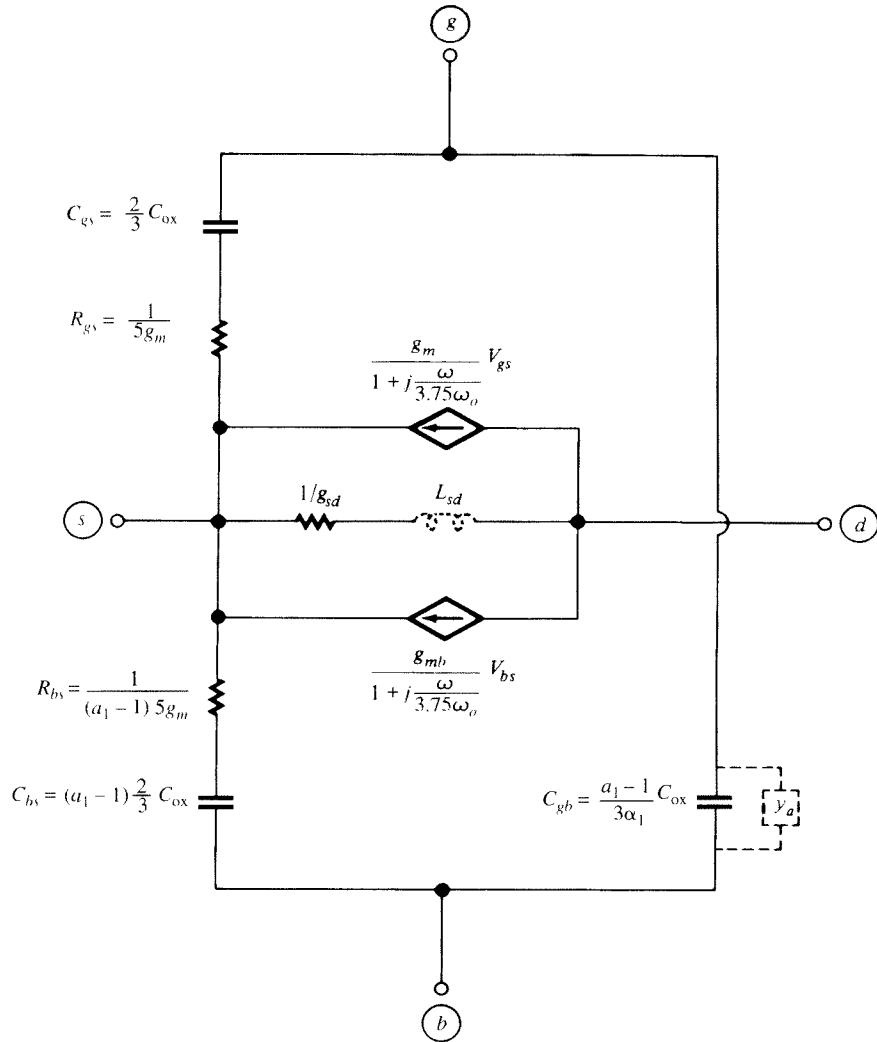


FIGURE 8.24
The model of Fig. 8.20 for the special case of operation in the saturation region.

use models outside the specified limits of validity). For example, a user might not be aware of such limits (which is usually the case in circuit simulation) or might have to use a model outside the region of validity, just because there is no better model available. In such cases, it is desirable for the resulting error not to be excessive. A good circuit simulation program, though, should warn the user when limits of validity are exceeded.

Models valid up to frequencies higher than ω_o can be developed by keeping an appropriate number of high-order terms in the y -parameter expressions.^{39,40,43-45,48-50,52,55} This must be done with care to ensure graceful degradation in the sense

discussed earlier. One might wonder what need there could be for models valid at such high frequencies when the model of Fig. 8.20 is capable of predicting device performance all the way to ω_o , which is near the intrinsic transition frequency of the device (7.3.32). Consider, however, a circuit with devices operating well below *their* ω_o . In the same circuit there may be *other* devices with *longer* channels for which ω_o is much *smaller* [see (8.4.67)].[†] Unless these devices are modeled correctly, the predicted performance for the whole circuit might be in error.

In such cases, an alternative to using very complicated high-order models may be considered for the devices with the longer channels. Let v_{highest} be the highest operating frequency of interest for a given circuit. Evaluate ω_o for all relevant devices in the circuit. Those with $\omega_o > \omega_{\text{highest}}$ can be modeled as discussed in this section. The others can be divided into “subtransistors” (Fig. 6.12), such that ω_o of each subtransistor is larger than ω_{highest} . The subtransistors can then also be modeled as discussed in this section. Notice that, even though the internal subtransistors may have a small channel length, these subdevices do not have real sources and drains; thus, the model used for them should be implemented carefully. No junction and overlap capacitances should be associated with the internal points. Short-channel effects in the subtransistors need to be handled carefully. Velocity saturation is automatically handled, as although the section’s length is a fraction of the total length, so the section’s “drain-source” voltage is a fraction of the total drain-source voltage, and the electric field distribution in each section is correct. However, other effects such as DIBL require special consideration.

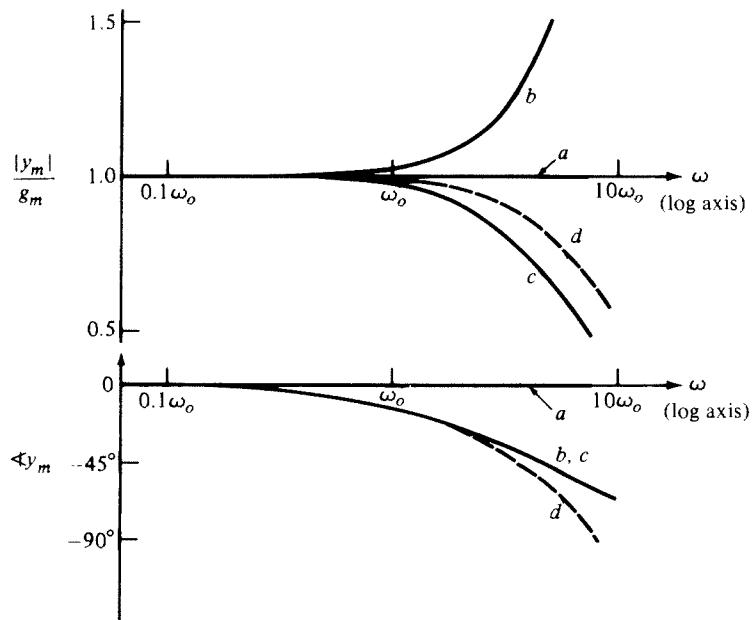
The model we have derived in this section is valid in strong inversion only. It is possible to derive a non-quasi-static model valid in all regions of operation, starting from the general all-region model concepts discussed in Sec. 4.3.³⁹ The topology of such a model is exactly the same as in Fig. 8.20, but, the expressions for the model parameters are different. Unfortunately, these expressions are complicated.

8.4.4 Model Comparison

Figure 8.25 shows plots³⁹ of $|y_m|/g_m$ and of the phase of y_m vs. frequency (log scale) in strong inversion, for $\eta = 0.5$ ($V_{DS} = V'_{DS}/2$). The frequency ω_o is given by (7.3.6), repeated here for convenience:

$$\omega_o = \frac{\mu(V_{GS} - V_T)}{\alpha L^2} \quad (8.4.75)$$

[†] For example, long channels may be used for certain devices to keep their source-drain small-signal conductance, g_{sd} , low. This may be needed, for example, to make possible high dc gain in a “bias loop” used to bias RF circuits. Although these devices are supposed to be operating at dc, parasitic paths may exist that can couple RF signals to them, and at such frequencies, the long-channel devices will be operating non-quasi-statically. Further, channel lengths longer than the minimum may be needed to reduce mismatch between devices; the shorter a device is, the more sensitive it is to mismatch (Chap. 9). Keeping mismatch low can be particularly important, for example, for RF circuits that process signals using two paths that should ideally exhibit identical behavior, which process in-phase (I) and quadrature (Q , or 90° phase-shifted) channels.

**FIGURE 8.25**

Normalized magnitude and phase of y_m vs. frequency for a transistor operating in nonsaturation with $\eta = 0.5$.³⁹ (a) Simple model of Fig. 7.19; (b) complete quasi-static model of Fig. 8.5; (c) model of Fig. 8.20; (d) numerical result (© 1985 by IEEE).

In each plot, curve *a* is for the simple model of Fig. 7.19, curve *b* is for the complete quasi-static model of Fig. 8.5, and curve *c* is for the model of Fig. 8.20. Finally, curve *d* is for a model resulting by keeping many terms in the numerators and denominators of (8.4.65). This model is valid even beyond $10\omega_o$.³⁹ It is seen in the phase plots that going from *a* to *c* produces a drastic improvement in the region of validity. The region of validity for *b* is limited by the fact that, at high frequencies the error in the magnitude becomes severe. This is because y_m contains a right-half-plane zero for this model (8.3.11g) in contrast to the left-half-plane pole in y_m for *c* (8.4.69h). The upward-going magnitude for *b* at high frequencies is clearly unrealistic, since it suggests an enhancement in the forward gate-to-drain action, contrary to one's expectation that, at high frequencies, control of the gate on the drain current is gradually lost due to the inversion layer's inertia. In fact, as far as magnitude is concerned, *a* is better than *b*, although it corresponds to a simpler model! Going to *c* eliminates the problem of *b*, and now both magnitude and phase are predicted satisfactorily up to about ω_o .

Similar plots can be constructed for other operating points and for other parameters in order to compare the various models and determine upper frequency limits of validity.³⁹ It is found that the upper-frequency limit of validity for a parameter of a given model depends on which parameter it is, what the operating point is, what accuracy is desired, whether magnitude or phase is of most interest, etc. Furthermore, one can always construct pathological cases where a model will fail in some way

(Prob. 8.17). Thus, the frequency limits of validity we have been providing should be considered in the above light. They are only rough indications of the regions within which a given model will perform satisfactorily in most cases. To summarize, the frequency limits of validity we have suggested for strong-inversion models in this and the previous chapters are as follows:

1. Quasi-static model without transcapacitors (Fig. 7.19): $\frac{\omega_o}{10}$
2. Quasi-static model with transcapacitors (Fig. 8.5): $\frac{\omega_o}{3}$
3. First-order non-quasi-static model (Fig. 8.20): ω_o

8.5 High-Frequency Noise

Non-quasi-static effects influence the power spectral density of the drain current noise at very high frequencies.^{62–71} It turns out, though, that this happens mainly at frequencies higher than ω_o , which is the limit of validity of our first-order non-quasi-static model. Thus, the drain current power spectral density in the model of Fig. 7.40, can be calculated as in Sec. 7.5. There is, however, another effect that should be considered at high frequencies, as explained in the following.

Thermal noise in strong inversion was seen in Sec. 7.5 to be the result of random potential fluctuations in the channel. These fluctuations are coupled to the gate terminal through the oxide capacitance and cause a gate noise current to flow even if all terminal voltages are fixed.^{45,62–71} This current is called *induced gate noise*. At high frequencies, the impedance of the gate capacitance becomes smaller, and this effect becomes more pronounced. Thus, non-quasi-static models, being appropriate for high-frequency work, should include a model for this noise.

To obtain a feel for this effect, consider, as an example, a transistor in the strong-inversion saturation region, connected in a common-source configuration as shown in Fig. 8.26a, and ignore extrinsic elements. The corresponding intrinsic small-signal model can be taken from Fig. 8.24 and is given in Fig. 8.26b with only the most important elements shown. To model the drain noise current, a noise current source i_{nd} in the drain port is used as discussed in Sec. 7.5. Consider now the resistance R_{gs} . The presence of this element in the model has been seen before to occur due to the resistance of the channel, which causes distributed *RC* effects in conjunction with the gate capacitance. If a complete non-quasi-static analysis is carried out for the influence of such effects on the induced gate noise,^{62,63,66–69} a surprisingly simple, and by no means obvious, result is found: it turns out that the induced gate current can be estimated by assuming that R_{gs} produces the same amount of thermal noise an actual resistor of value R_{gs} would produce.⁶⁵ This noise is modeled in Fig. 8.26b as a noise voltage source v_{ng} . According to (7.5.3a), the power spectral density of this noise will be $S_{vng} = 4kTR_{gs}$. Various accurate (and very complicated) calculations^{60–63,66–69} give practically the same result, within a factor close to unity. According to the results of Ref. 69, this factor is 4/3; thus we have

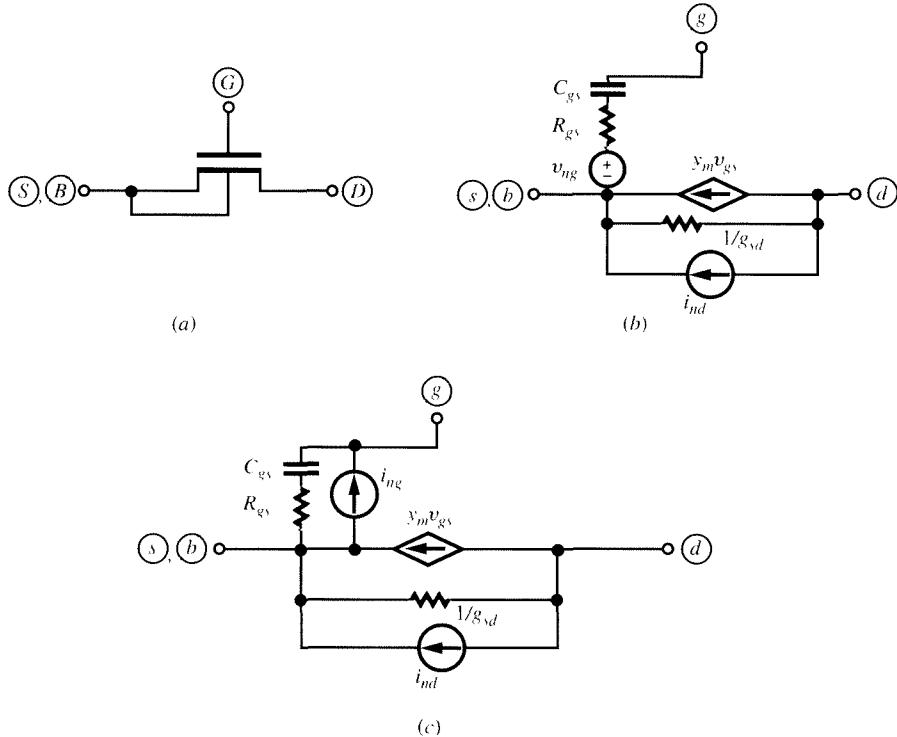


FIGURE 8.26

(a) A transistor with substrate shorted to source, assumed to operate in the saturation region. (b) Equivalent small-signal circuit for (a), including noise sources. (c) Alternative representation of the circuit in (b).

$$S_{vng} = 4kT \left(\frac{4}{3} R_{gs} \right), \quad \text{saturation} \quad (8.5.1)$$

Another common representation for the induced gate noise uses a “Norton equivalent” of the gate branch in Fig. 8.26b.⁶⁵ Such a circuit can be found by replacing the series voltage source by a parallel current source, as shown in Fig. 8.26c. From classical Thévenin-to-Norton conversion,^{57–59} it can be concluded that the power spectral density S_{img} of this source will be S_{vng} times the square of the magnitude of the admittance of the C_{gs} - R_{gs} branch (the square comes from the fact that power spectral densities are associated with mean square values; see Sec. 7.5). From the value given in Fig. 8.24, it is easily seen that for frequencies below ω_o the admittance of this branch is dominated by C_{gs} , and its magnitude is thus approximately given by ωC_{gs} ; thus

$$S_{img} = 4kT \left(\frac{4}{3} R_{gs} \right) \omega^2 C_{gs}^2, \quad \text{saturation, } \omega < \omega_o \quad (8.5.2)$$

Using for R_{gs} and C_{gs} the saturation values given in Fig. 8.24, with $g_m = (W/L) \mu C'_{ox} (V_{GS} - V_T)$ (assuming $\alpha \approx 1$) the preceding equation becomes (Prob. 8.28)

$$S_{ing} = 4kT \frac{\omega^2 (C'_{ox} WL)^2}{\frac{W}{L} \mu C'_{ox} (V_{GS} - V_T)} \frac{16}{135}, \text{ saturation, } \omega < \omega_o \quad (8.5.3)$$

This result agrees exactly with those obtained using very complicated calculations in which infinitesimal elements of the channel are considered, their contribution to the current noise is found, and their combined effect is calculated.^{45,62-69} In fact, this development can be carried out directly in the nonsaturation region, giving an equation of the same form:^{63,68}

$$S_{ing} = 4kT \frac{\omega^2 (C'_{ox} WL)^2}{\frac{W}{L} \mu C'_{ox} (V_{GS} - V_T)} K_2(V_{DS}), \quad \omega < \omega_o \quad (8.5.4)$$

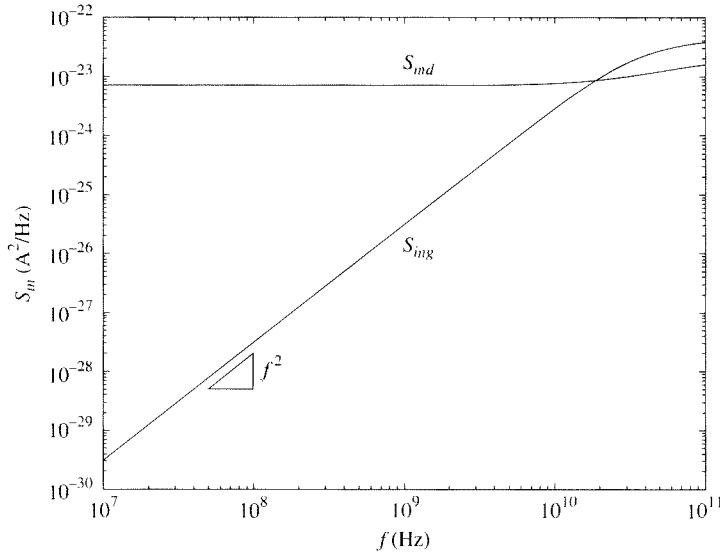
where the factor $K_2(V_{DS})$ is $1/12 \approx 0.083$ at $V_{DS} = 0$ and gradually rises to $16/135 \approx 0.12$ in saturation.⁶⁸ In the preceding formulas, the quantity in the denominator can be recognized, from (7.2.37), to be the source-drain small-signal conductance at $V_{DS} = 0$:

$$\frac{W}{L} \mu C'_{ox} (V_{GS} - V_T) = g_{ds} \Big|_{V_{DS}=0} \quad (8.5.5)$$

This quantity can thus easily be measured or determined by small-signal simulation. Noise expressions are often expressed in terms of this quantity.^{69,72,73}

Figure 8.27 shows drain and gate noise spectral densities, simulated with a 10-section ideal symmetric linearization charge sheet model, with no parasitics. Such a model approximates distributed effects of the intrinsic transistor (see Fig. 6.12 and associated discussion). As expected from the preceding discussion, because the capacitive coupling is directly proportional to frequency, the induced gate noise (in A²/Hz) increases as the square of the frequency. In the plot, the frequency is driven to values higher than the transition frequency f_T (6.6 GHz for the device simulated) to show the behavior there, although in practice a transistor would typically not be operated at such high frequencies relative to f_T . Figure 8.28 shows the gate noise spectral density at one frequencies vs. V_{DS} . The points are as calculated from (8.5.4) with the values given below it and using (8.5.5).

Similar calculations can be used to determine the degree of correlation between the gate and drain thermal-induced noise currents (a partial correlation is to be expected, since both currents have the same physical origin). In circuit calculations, caution is needed to take this correlation properly into account. One way to do this is to characterize the correlation by what is called the “cross-spectral density” (a quantity analogous to the power spectral density, but defined based on the product of the variations in the two currents, rather than on the square of the variation in a single current^{68,69}). This quantity turns out to be imaginary, a fact that can be traced to the phase shift in the capacitively induced gate current compared with the fluctuations in the channel current. For readers familiar with circuit noise calculations, we mention that the cross-spectral density of the gate and drain thermal-induced currents in saturation is given by^{63,68}

**FIGURE 8.27**

Drain and induced gate noise vs. frequency for an idealized device operating in strong-inversion saturation, $t_{ox} = 2.5 \text{ nm}$, $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $W = 1 \mu\text{m}$, $L = 1.2 \mu\text{m}$, $V_{GS} - V_T = 1.5 \text{ V}$, $V_{DS} = 5 \text{ V}$. $f_i = 6.6 \text{ GHz}$.

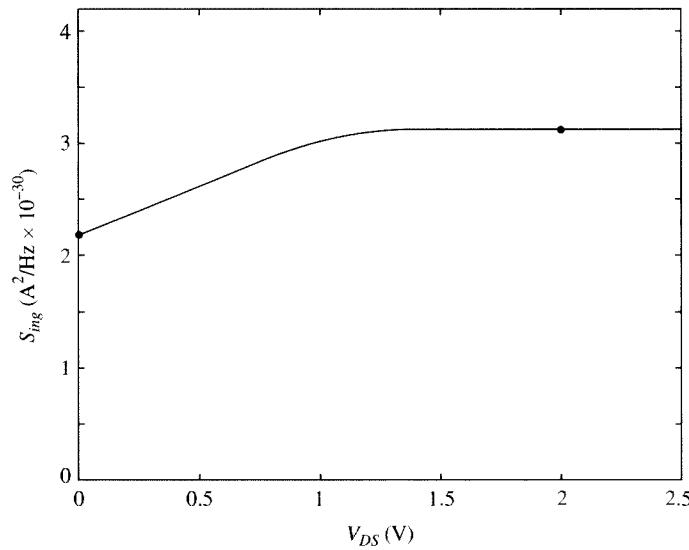
$$S_{ing,ind} = 4kT \frac{1}{9} j\omega (C'_{ox} WL), \quad \text{saturation}, \quad \omega < \omega_o \quad (8.5.6)$$

This quantity is often normalized to give the so-called (cross-) correlation coefficient, defined as^{68,69}

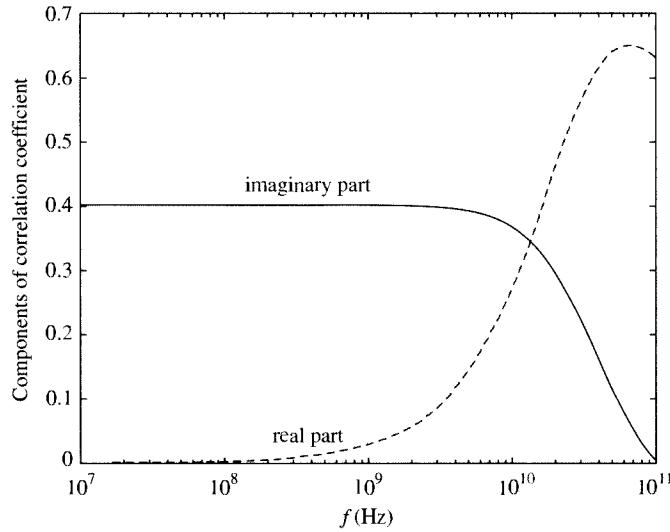
$$c = \frac{S_{ing,ind}}{\sqrt{S_{ing} S_{ind}}}, \quad \text{saturation}, \quad \omega < \omega_o \quad (8.5.7)$$

From the results given here and in Sec. 7.5, we find that this correlation coefficient in saturation should be $\sqrt{5/32}j \approx 0.395j$. Figure 8.29 shows the real and imaginary parts of the cross-correlation coefficient vs. frequency, for operation in strong-inversion saturation, from the 10-section ideal model mentioned above. At low frequencies, the simulated correlation coefficient is very close to the predicted value. At high frequencies, as higher-order frequency terms, not included in the above results, become important, the detailed simulations differ from the above value. Figure 8.30 shows the value of the correlation coefficient vs. V_{DS} . It is seen that this value goes to zero at $V_{DS} = 0$.

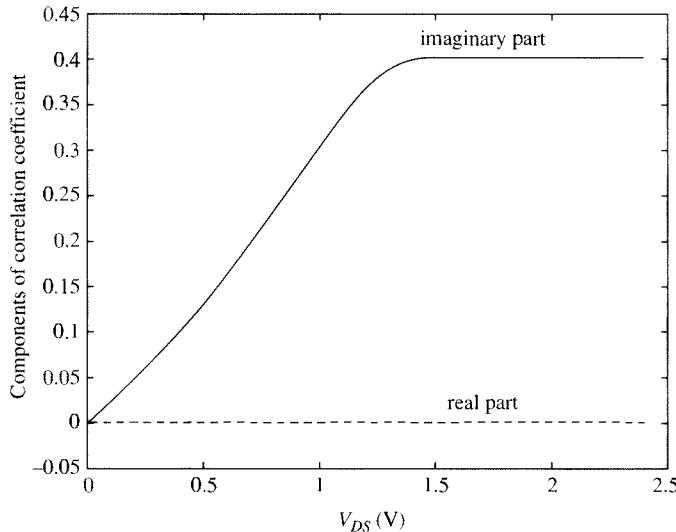
We note that, although a multisection model is useful for investigation purposes as previously discussed, for the purpose of circuit simulation it is not feasible to use such a model because of computational inefficiency. Therefore, induced gate noise is better implemented directly,⁷⁴ based on equations such as the ones we have already presented.

**FIGURE 8.28**

Induced gate noise vs. V_{DS} for the case of Fig. 8.27, at $f = 10$ MHz. Line is from distributed model simulation; points are from (8.5.4).

**FIGURE 8.29**

Correlation coefficient between gate and drain noise vs. frequency for the case of Fig. 8.27, $f_T = 6.6$ GHz.

**FIGURE 8.30**

Correlation coefficient between gate and drain noise vs. V_{DS} for the case of Fig. 8.27, at $f=10$ MHz.

8.6 Considerations in MOSFET Modeling for RF Applications

Model Topologies The frequencies used for radio communications cover a wide range. Usually, radio-frequency (RF) modeling of MOSFETs refers to modeling at frequencies up to the transition (cutoff) frequency of the devices. For such applications, small-signal non-quasi-static modeling becomes indispensable for the intrinsic part, and Secs. 8.3 and 8.4 are of direct relevance. However, the extrinsic part must be considered, too (Sec. 6.8). Distributed effects in that part can be modeled using lumped approximations. An example is shown in Fig. 8.31a (noise sources are not shown for simplicity). A non-quasi-static model is used for the intrinsic part; this makes sense, since both long- and short-channel devices have been found to exhibit non-quasi-static effects.⁵¹ The resistances are those of the extrinsic gate, substrate, and source-drain regions. As seen, each is divided into subresistances connected to a common node, so that the appropriate capacitances can be connected to that node. For example, the elements R_{se1} , R_{se2} , and C_{bse} model, to first order, what is actually a distributed RC effect of the source resistance and its capacitance to the substrate. For noise modeling, each resistance can be modeled as in Fig. 7.30, and of course noise sources for the intrinsic part should be used as discussed in the previous section.

Although the level of modeling exhibited by Fig. 8.31a is very desirable, in practice it is difficult to use because the individual values of the subresistances are difficult to determine. This is particularly true for the substrate subresistances. In such cases, one may have to resort to the model shown in Fig. 8.31b. At times,

530 OPERATION AND MODELING OF THE MOS TRANSISTOR

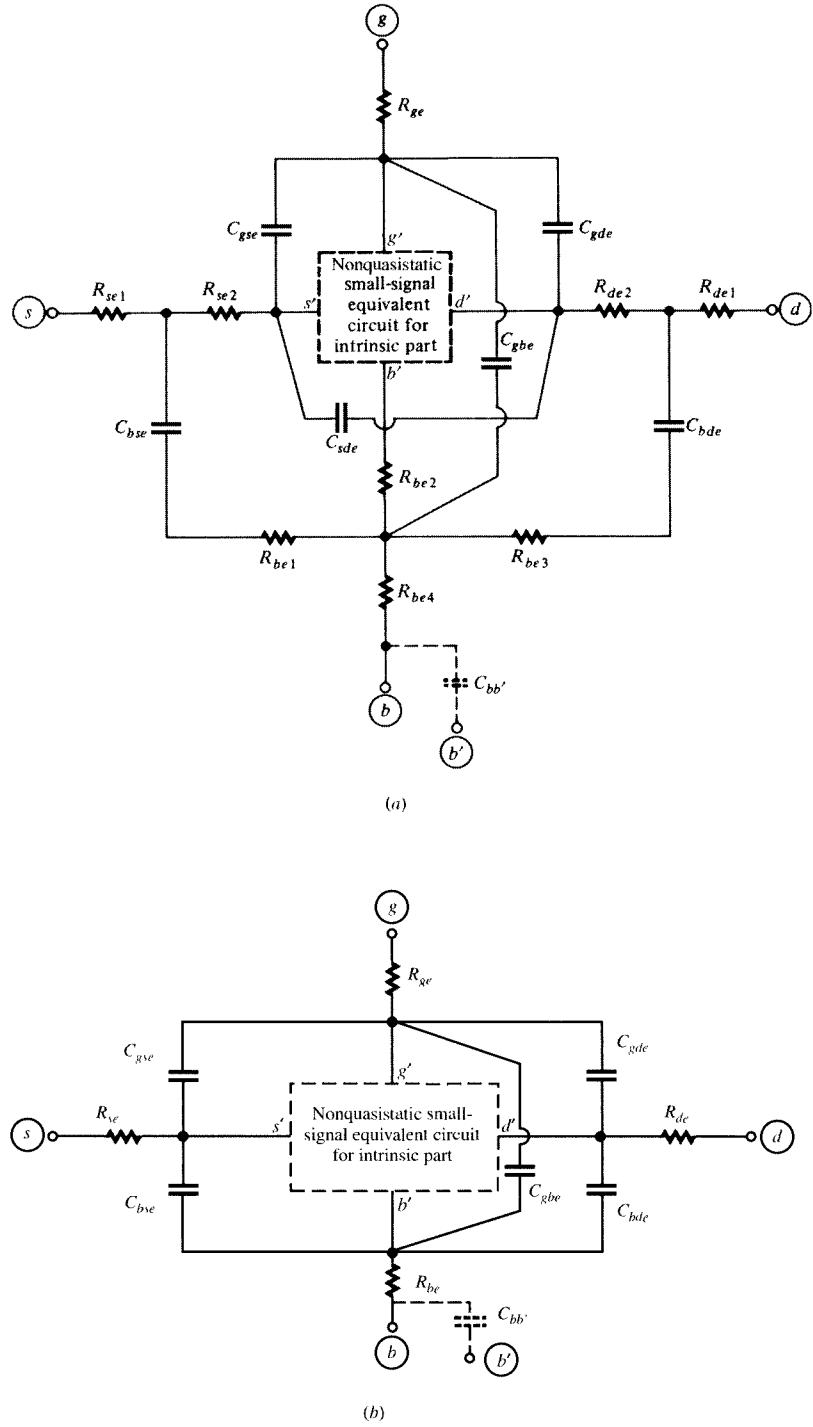


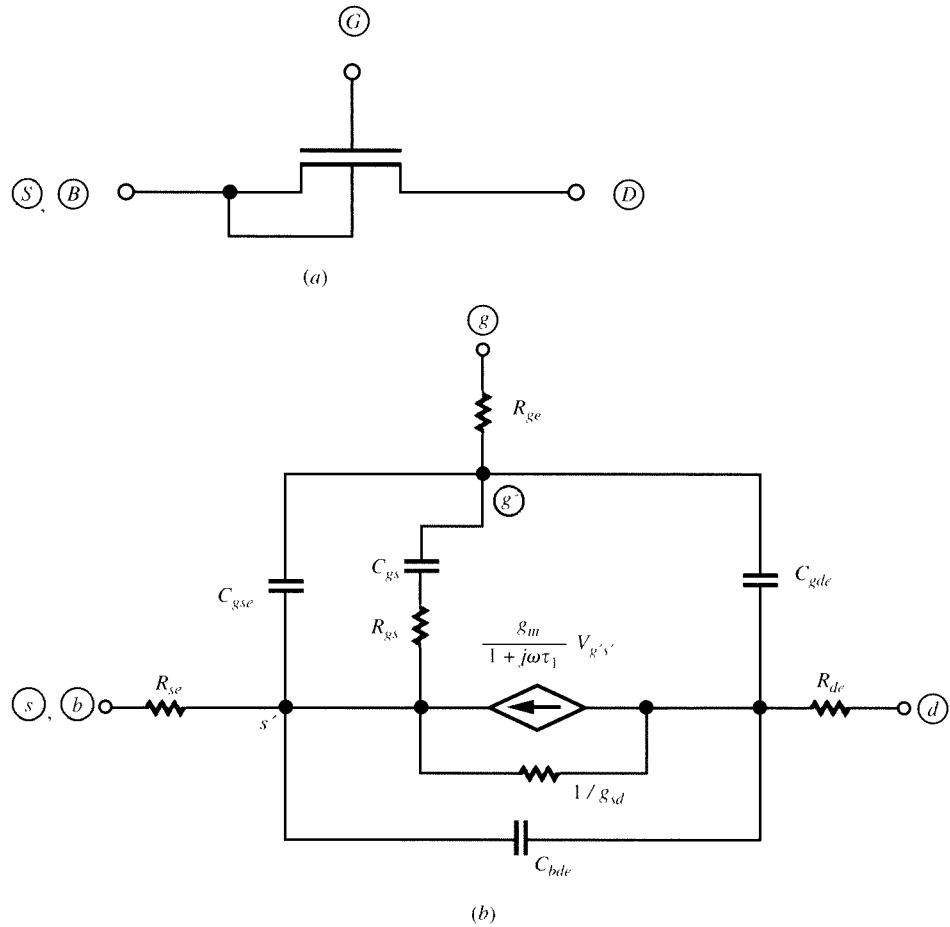
FIGURE 8.31
Small-signal models for the complete transistor: (a) is more accurate; (b) is more practical.

models even simpler than that are used.^{75–85} An example is shown in Fig. 8.32. Figure 8.32a shows a transistor with the source and substrate shorted together (a very common connection). For operation in the saturation region, it is claimed in the literature that the model in Fig. 8.32b is valid. The intrinsic part of the model is taken from Fig. 8.24. Note, however, that this model cannot be derived from that in Fig. 8.31b since, although the external source and substrate terminal are shorted together, R_{se} and R_{be} prevent the *intrinsic* source and substrate terminals (s' and b' in Fig. 8.31b) from being short-circuited. At times, an even simpler model is used; for example, R_{gs} and/or the frequency dependence of the gate transadmittance are omitted, or the intrinsic device is modeled using the complete quasi-static model.

At this point, one may reasonably wonder, "How can such highly simplified models be of any use at RF?" There are several answers to this question. One is related to the extrinsic parasitics; these can dominate device behavior in some cases, and thus limit application to frequencies lower than those at which the effect of R_{gs} or of τ_l would be felt. Another answer is related to parameter extraction. The elements are often assigned whatever values make the model give results close to measurements. The values one ends up with in this way are nonphysical, but are used to make up for the inadequacies of the model. This is a dangerous practice, as those same nonphysical values may result in a very wrong prediction of a different performance parameter, not considered during parameter extraction; this problem will be discussed in more detail in Chap. 10. In general, the fact that a grossly simplified model happens to give satisfactory results for some cases is no reason for trusting such a model; there can be other cases where the same model will prove very inadequate. Simplified models can totally fail to predict certain phenomena that more complete models such as the one in Fig. 8.31a would predict. For example, at high frequencies, the impedances of C_{bse} and C_{bde} in that model become very small, and they tend to shunt the channel with the substrate resistances R_{be1} and R_{be3} . This can affect the value of the admittance y_{dd} seen at the drain terminal. This effect cannot be predicted by the model of Fig. 8.32b. In general, models developed for low-frequency work will have difficulty in modeling RF operation for all situations. If such models have to be used, one may at least extend their range of validity by using the approach suggested in Fig. 6.12. This has been attempted for RF applications.⁸⁶

In the derivations for the various parameters in this chapter, short-channel effects were not considered. The topologies for small-signal non-quasi-static equivalent circuits derived in this chapter are still useful in the presence of such effects, but the values of the parameters in them must be modified. The influence of some short-channel effects has been considered, in the literature.^{49,50,52,96–99} Simplified models like the one in Fig. 8.32b are still used in association with short-channel devices, with element values determined from measurements.

It should be noted that, rather than try to stretch simplified models like the one in Fig. 8.32b to fit measurements, one can in principle use a general y -parameter representation, such as the ones shown in Figs. 8.15 to 8.17, since these were derived for an arbitrary four-terminal structure without *any* assumptions as to channel length, substrate uniformity, type of extrinsic effects, etc. In other words, those representations

**FIGURE 8.32**

(a) A transistor with substrate shorted to source. (b) Small-signal model sometimes used for the connection in (a) in the saturation region.

were completely general and can in principle take care of all distributed effects, at any frequency, provided appropriate values for the admittances in them can be provided as a function of frequency. However, calculating such values is very involved even for long-channel devices.⁴⁵ If values directly extracted from measurements are used, the models do not have any predictive power for situations other than those measured. In addition, y -parameter models are not supported in many simulation programs.

Gate Resistance In many RF applications, devices with large gate widths W are used. In such cases, gate resistance effects can be very significant.^{80,87–91} As we have discussed in Sec. 6.8, the gate resistance forms a distributed RC circuit with the

gate-channel capacitance. We can gain a first-order intuitive understanding of the effect of the gate resistance by thinking of it, as well as of the gate capacitance, as lumped elements. At low frequencies, there is negligible ac current through the gate capacitance, and thus negligible voltage drop across the gate resistance; so the complete ac gate voltage is available to generate ac drain current through the transconductance g_m . At higher frequencies, the magnitude of the admittance of the gate capacitance increases, and considerable current flows through it; this current causes a significant voltage drop across the gate resistance, thus diminishing the effective gate voltage of the device, effectively decreasing its transadmittance.

To minimize the gate resistance, in practice RF MOSFET layouts are “folded” or multifinger structures, as shown in Fig. 8.33. Alternate fingers are connected in parallel through a metal layer (not shown) to form the source and drain regions. A particular overall device width W is generally required to provide the desired current drive or transconductance. If the width of each finger is too big, then the gate resistance will limit device performance due to the previous effects. If the width of each finger is too small, then the parasitic capacitances associated with the metal interconnect necessary to contact each finger will become excessive and will again limit device performance. Care is taken in RF MOSFET layouts to minimize both gate resistance and parasitic capacitance to the extent possible. Depending on the details of a particular technology, maximum gate finger widths vary from about 1 to 3 μm , and gates are contacted at both sides to minimize gate resistance. As follows from our discussion in Sec. 6.8, the main part of the gate resistance, with contacts at both ends of the gate, is one-quarter of that when only one end is contacted (compare Figs. 6.18a and 6.19).

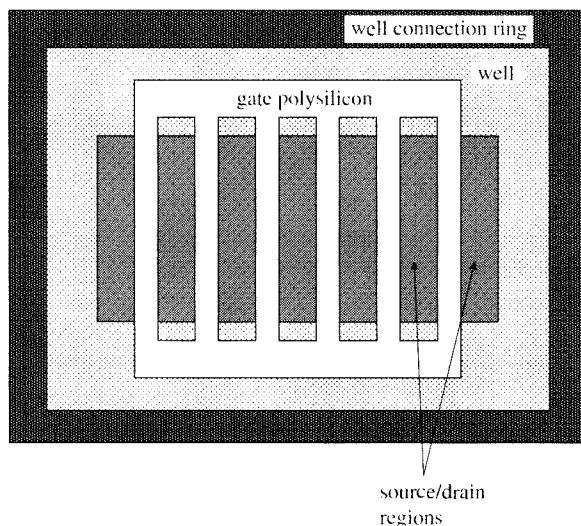


FIGURE 8.33

Multifinger layout used to minimize gate resistance. Interconnect metallization, which affects extrinsic parasitic capacitance, is not shown.

The gate resistance, in addition to affecting the frequency response, also contributes noise. At low frequencies, this noise can be modeled by using a single effective value for the gate resistance, discussed in Sec. 6.8.^{90,92,93} Depending on geometrical dimensions and gate resistivity, this noise can in some cases be significantly higher than that produced by the intrinsic part of the device. At high frequencies, the gate resistance noise tends to be “filtered out” by the gate capacitance, and the total noise approaches that produced by the intrinsic part.⁹¹

Transition Frequency In Sec. 7.3.2, we considered the intrinsic part of the transistor, and we derived an expression for the “intrinsic transition frequency.” For the complete device, including extrinsic elements, we can similarly calculate a *transition frequency* (or *cutoff frequency*), as that frequency at which the “short-circuit current gain” of the transistor (measured in saturation, with a dc source between drain and source, as in Fig. 7.22a) becomes unity. An estimate can be obtained by using a greatly simplified equivalent circuit, as shown in Fig. 8.34; the short between drain and source corresponds to the dc source in the actual circuit. Voltage and current phasors are indicated in the figure. A current source input is assumed for convenience. The total capacitance between the internal node g' and ground is seen to be

$$C_g = C_{gs} + C_{gb} + C_{gd} \quad (8.6.1)$$

where the individual capacitances include both intrinsic *and* extrinsic components. Thus $V_{g's}$ in the figure is $I_i/(j\omega C_g)$. Elements $1/g_{sd}$ and C_{bd} have a zero voltage

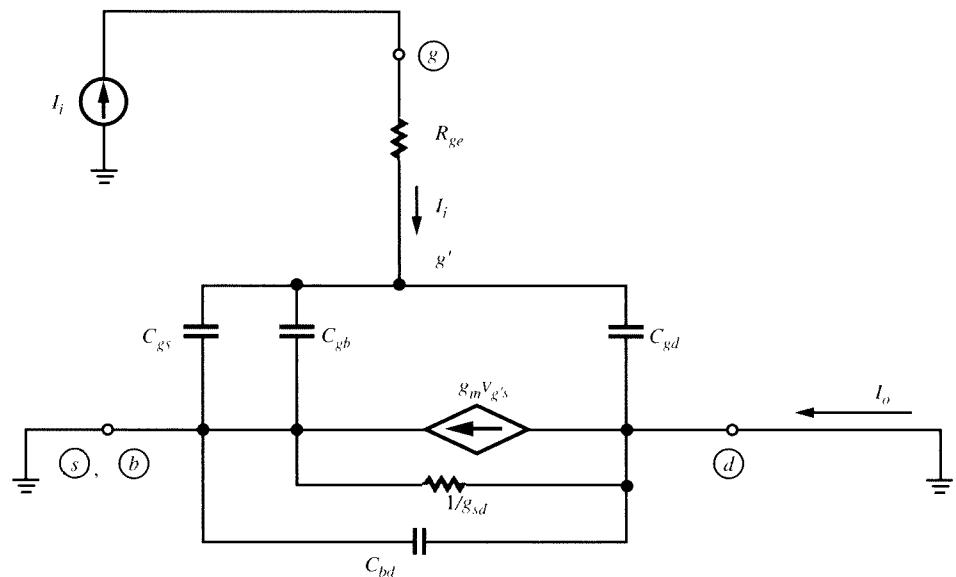


FIGURE 8.34
Circuit for estimating the transition frequency ω_T .

across them, and their current is zero; neglecting the small current through C_{gd} , the output current is $I_o = g_m V_{g's} = g_m I_i / (j\omega C_g)$, and thus the short-circuit gain is $I_o/I_i = g_m / (j\omega C_g)$. The transition frequency ω_T is defined as the frequency at which the magnitude of I_o/I_i becomes unity. Thus

$$\boxed{\omega_T = \frac{g_m}{C_g}} \quad (8.6.2)$$

Let us consider two special cases of interest. Assume first that no velocity saturation is present. Then, g_m can be calculated from (7.2.29a). Using values given for the various capacitances in Secs. 7.3 and 7.4, it can be seen that a simple estimate for C_g is the total oxide capacitance, $C'_{ox} WL$. Using these expressions in (8.6.2), we obtain

$$\omega_T = \frac{\mu(V_{GS} - V_T)}{\alpha L^2} = \omega_o, \quad \text{no velocity saturation} \quad (8.6.3)$$

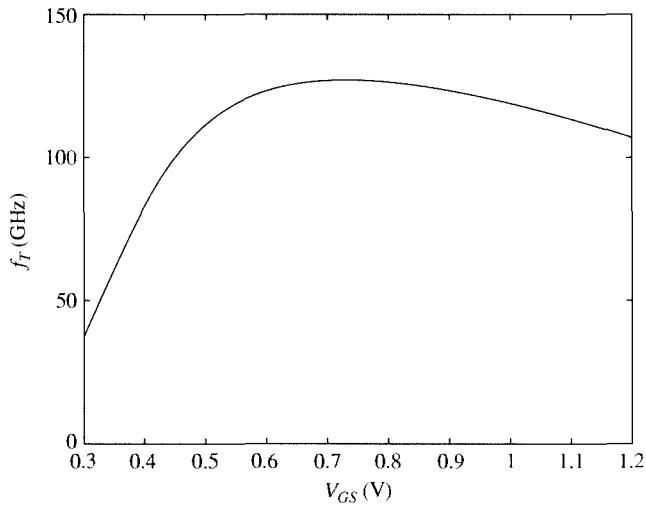
where ω_o is the quantity defined in (8.4.67). If, instead, complete velocity saturation is assumed, g_m can be calculated from (7.2.30b); from this, and using the same estimate as before for C_g , we obtain

$$\omega_T = \frac{|v_d|_{\max}}{L}, \quad \text{velocity saturation} \quad (8.6.4)$$

where $|v_d|_{\max}$ is the magnitude of the saturated carrier velocity (Sec. 5.2).

Example 8.1. For a transistor with a 0.1-μm gate length, assuming a saturation velocity of 10^7 cm/s, (8.6.4) gives $\omega_T = 1,000$ Grad/s, corresponding to $f_T = \omega_T/(2\pi) = 159$ GHz.

Several observations can now be made. Consider first the effect of reducing L on ω_T . As seen from the above development, for long-channel devices, reducing the channel length increases ω_T drastically, due to the presence of L^2 in (8.6.3). When the channel becomes very short, though, velocity saturation can occur and thus the increase in ω_T will be less drastic, as seen from (8.6.4). The effect of the gate bias voltage is also noteworthy. In the absence of velocity saturation effects, ω_T can be increased by increasing V_{GS} as seen in (8.6.3). The increase will not be linear, though, since increasing V_{GS} decreases the effective mobility. For large V_{GS} , the saturation voltage V'_{DS} can be large enough for velocity saturation to set in, and thus the effect of V_{GS} diminishes, as can be expected from (8.6.4). This is illustrated in Fig. 8.35. In fact, if the influence of V_{GS} on mobility and of all the various short-channel effects on g_m and C_g in (8.6.2) are considered, it is often found that further increases in V_{GS} can actually cause a decrease in ω_T , as shown toward the right in Fig. 8.35.

**FIGURE 8.35**

Transition frequency as a function of $t_{ox} = 1.5$ nm, $N_A = 3 \times 10^{17} \text{ cm}^{-3}$, $W = 10 \mu\text{m}$, $L = 0.09 \mu\text{m}$, $V_{DS} = 1$ V.

More accurate models give more complicated expressions for ω_T , but the corresponding values are usually not too different from those found from (8.6.2). Thus, this equation is widely used. Notice that R_{ge} does not enter into this expression, since R_{ge} in Fig. 8.34 is in series with a current source, and thus the current entering node g' is I_s , independent of the value of R_{ge} . However, R_{ge} is known to have a detrimental effect in RF circuits, so its absence from the expression for ω_T simply points to the fact that ω_T is not an adequate figure of merit for general RF performance. An additional figure of merit, which makes the effect of R_{ge} clear, is now discussed.

Maximum Frequency of Oscillation A common figure of merit for RF work is the *maximum frequency of oscillation* ω_{max} , also called *unity power gain frequency*. Power gain in a circuit is the ratio of the load power to the input power. Transistors take power from the power supply and convert it to output power delivered to a load and controlled by the input. A power gain larger than unity is what qualifies these devices as “active.” Power gain can be maximized by choosing the circuitry connected to the transistor in an optimum way: The transistor input is conjugate-matched to the input signal source, the load is conjugate-matched to the transistor output impedance, and an appropriate network is used to cancel the effect of feedback from the output to the input.⁹⁴ The power gain obtained in this way is called *unilateral* power gain. As the frequency of operation is raised, the unilateral power gain decreases and eventually becomes less than unity; the device, then, behaves as a “passive” device. The frequency at which the unilateral power gain drops to unity is defined as ω_{max} and is a quantity of fundamental importance. This quantity is called the *maximum oscillation frequency*, since this is the maximum frequency at which the device can

be made to oscillate by feeding its output back to its input. Since R_{ge} dissipates input power, the quantity ω_{\max} captures the effect of this element.

Using circuit theory, one can calculate ω_{\max} for a given model.⁹⁵ For complete models, the resulting expressions can be very complicated, so usually simple estimates are obtained by using simple models. For example, assuming that the source extrinsic resistance R_{se} is very small, one can use the model in Fig. 8.34 (without the output short circuit and embedded in appropriate circuitry). For this model it can be found that using optimum external circuitry,^{94,95} the unilateral power gain is approximately $\omega_T^2/[4R_{ge}\omega^2(g_{sd} + \omega_T C_{gd})]$. The value of ω_{\max} can be found by setting this gain equal to 1 and solving for $\omega = \omega_{\max}$; a consideration of distributed effects⁹⁶ shows that, to take such effects into account, R_{ge} in the resulting expression should be replaced by $R_{ge,\text{eff}}$, as calculated in Sec. 6.8. Thus we find

$$\omega_{\max} \approx \frac{\omega_T}{\sqrt{4R_{ge,\text{eff}}(g_{sd} + \omega_T C_{gd})}}, \quad R_{se} \ll R_{ge} \quad (8.6.5)$$

This can be larger or smaller than ω_T . The value of the gate resistance is of obvious importance here; unless it is small, it can make ω_{\max} significantly less than ω_T , thus limiting the usefulness of the device for many applications.

Example 8.2. Consider the device of the previous example, and assume that it has $R_{ge,\text{eff}} = 40 \Omega$, $g_{sd} = 2 \text{ mA/V}$, and $C_{gd} = 3 \text{ fF}$. Then (8.6.5) gives $\omega_{\max} = 1,118 \text{ Grad/s}$, or $f_{\max} = \omega_{\max}/(2\pi) = 178 \text{ GHz}$.

To keep $R_{ge,\text{eff}}$ small, one needs silicided gates (Sec. 4.12), multiple contacts (Fig. 6.19), or a device split into several subdevices connected in parallel as in Fig. 8.33; while devising appropriate layouts for this, though, care has to be taken not to increase the parasitic capacitance values significantly.

Because of the decisive effect of R_{ge} on ω_{\max} , the inclusion of only this extrinsic resistance in the above calculation is justified in many cases. However, if R_{ge} is made small using the techniques just mentioned, ω_{\max} increases to the point where the effect of other elements (e.g., of the source series resistance R_{se} , or of R_{gs}) is felt. In such cases, more complete models may have to be used to calculate ω_{\max} . We note that in the literature on RF devices, sometimes the values of ω_{\max} quoted are calculated from a simplified version of (8.6.5), in which either C_{gd} or g_{sd} are assumed to be negligible. This can give artificially high ω_{\max} values in some cases.

Calculating ω_T and ω_{\max} from more complete models than the ones used in the previous discussion leads to very complicated expressions. Thus, the simple expressions we have presented are widely used. The quantities ω_T and ω_{\max} are often measured by extrapolation from the transistor behavior at low frequencies; thus, the expressions shown previously, which were derived using low-frequency models, are consistent with this practice. More on parasitic elements, and on their simulation and measurement for RF applications, can be found in Chap. 10 and in the references.^{96–100}

References

1. J. A. Geurst, "Calculation of high-frequency characteristics of thin film transistors," *Solid-State Electronics*, vol. 8, pp. 88–90, 1965.
2. D. B. Candler and A. G. Jordan, "A small-signal, high-frequency analysis of the insulated-gate field-effect transistor," *International Journal of Electronics*, vol. 19, pp. 181–196, 1965.
3. M. H. White and R. C. Gallagher, "Metal oxide semiconductor (MOS) small-signal equivalent circuits," *Proceedings of the IEEE*, vol. 53, pp. 314–315, 1965.
4. J. R. Hauser, "Small-signal properties of field-effect devices," *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 605–618, 1965.
5. J. A. Guerst and H. J. C. A. Nunnink, "Numerical data on the high-frequency characteristics of thin-film transistors," *Solid-State Electronics*, vol. 8, pp. 769–771, 1965.
6. W. Fischer, "Equivalent circuit and gain of MOS field-effect transistors," *Solid-State Electronics*, vol. 9, pp. 71–81, 1966.
7. R. Paul, "Hochfrequenzerhalten von Feldeffektransistoren mit isolierter Steuer Elektrode," *AEU*, vol. 20, pp. 317–328, 1966.
8. R. Paul, "Die Ersatzschaltung von Feldeffektransistoren mit isoliertem Gate," *Nachrichtentechnik*, vol. 16, pp. 243–249, 1966.
9. R. Paul, "Einfluss einer nichtidealen Gateisolation auf die Vierpolparameter des Feldeffekttransistors," *Nachrichtentechnik*, vol. 16, pp. 278–285, 1966.
10. R. Paul, "Frequenzabhängigkeit der Vierpoleigenschaften von MOS-Transistoren," *Nachrichtentechnik*, vol. 16, pp. 401–406, 1966.
11. Z. S. Girbnikov and Yu. A. Tkhonik, "Calculation of the transient processes in field-effect triodes with an insulated gate for the saturated mode of operation," *Radio Engineering and Electronic Physics*, vol. 11, pp. 776–781, 1966.
12. H. Johnson, "A high-frequency representation of the MOS transistor," *Proceedings of the IEEE*, vol. 54, pp. 1970–1971, 1966.
13. D. H. Treleaven and F. N. Trofimenkoff, "MOSFET equivalent circuit at pinchoff," *Proceedings of the IEEE*, vol. 54, pp. 1223–1224, 1966.
14. H. C. DeGraaff, "High frequency measurements of thin-film transistors," *Solid-State Electronics*, vol. 10, pp. 51–56, 1967.
15. R. Paul, "Experimentelles Hochfrequenzverhalten von MOS-Transistoren," *Nachrichtentechnik*, vol. 17, pp. 255–260, 1967.
16. M. B. Das, "Generalized high-frequency network theory of field-effect transistors," *IEE Proceedings*, vol. 114, pp. 50–59, 1967.
17. I. R. Burns, "High-frequency characteristics of the insulated gate field-effect transistors," *RCA Review*, vol. 28, pp. 385–418, 1967.
18. F. A. Lindholm, R. J. Balda, and J. L. Clements, "Characterization of the four-terminal MOS transistor for digital and linear applications," *Proceedings of the IEEE International Electronics Conference*, Toronto, 1967, pp. 116–117.
19. M. B. Das, "High-frequency network properties of MOS transistors including the substrate resistivity effects," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 1049–1069, 1969.
20. J. A. Van Nielen, "A simple and accurate approximation to the high-frequency characteristics of IGFETs," *Solid-State Electronics*, vol. 12, pp. 826–829, 1969.
21. J. W. Haslett and F. N. Trofimenkoff, "Small-signal, high-frequency equivalent circuit for the metal-oxide semiconductor field-effect transistor," *IEE Proceedings*, vol. 116, pp. 699–702, 1969.
22. R. S. C. Cobbold, *Theory and Applications of Field-Effect Transistors*, Wiley-Interscience, New York, 1970.
23. E. M. Cherry, "Small-signal high-frequency response of the insulated gate field-effect transistor," *IEEE Transactions on Electron Devices*, vol. ED-17, pp. 569–577, 1970.
24. M. V. Balakirev and V. M. Bogachev, "Frequency and transient characteristics of metal oxide semiconductor transistors and synthesis of their equivalent circuits," *Radio Engineering and Electronics Physics*, vol. 16, pp. 1884–1897, 1971.

25. M. Reiser, "A two-dimensional numerical FET model for dc, ac and large-signal analysis," *IEEE Transactions on Electron Devices*, vol. ED-20, pp. 35–45, 1976.
26. U. Kumar and S. C. Dutta Roy, "A simple small-signal two-part MOST model for the pre-pinchoff region," *Solid-State Electronics*, vol. 20, pp. 1021–1022, 1977.
27. U. Kumar, "A simple two-part model of the metal oxide semiconductor transistor," *Microelectronics Journal*, vol. 10, pp. 50–53, 1978.
28. J. I. Arreola, "Equivalent circuit modeling of the large signal transient response of four-terminal MOS field-effect transistors," doctoral dissertation, University of Florida, 1978.
29. D. E. Ward and R. W. Dutton, "A charge-oriented model for MOS transistor capacitances," *IEEE Journal of Solid-State Circuits*, vol. SC-13, pp. 703–707, 1978.
30. J. A. Robinson, Y. A. El-Mansy, and A. R. Boothroyd, "A general four-terminal charging-current model for the insulated-gate field effect transistor," *Solid-State Electronics*, vol. 23, parts I, II, pp. 405–414, 1980.
31. S. Y. Oh, "A simplified two-dimensional numerical analysis of MOS devices including transient phenomena," Technical Report G201-10, Integrated Circuits Laboratory, Stanford University, Stanford, Calif., June 1981.
32. D. E. Ward, "Charge-based modeling of capacitance in MOS transistors," Technical Report G201-11, Integrated Circuits Laboratory, Stanford University, Stanford, Calif., June 1981.
33. R. Conilogue and C. Viswanathan, "A complete large and small signal charge model for a MOS transistor," *Proceedings of the IEEE International Electron Devices Meeting*, San Francisco, 1982, pp. 654–657.
34. K. Y. Tong, "AC model for MOS transistors from transient-current computations," *IEE Proceedings*, vol. 130, part I, pp. 33–36, February 1983.
35. J. J. Paulos and D. A. Antoniadis, "Limitations of quasi-static capacitance models for the MOS transistor," *IEEE Electron Device Letters*, vol. EDL-4, pp. 221–224, July 1983.
36. C. Turchetti, G. Masetti, and Y. Tsividis, "On the small-signal behavior of the MOS transistor in quasi-static operation," *Solid-State Electronics*, vol. 26, pp. 941–949, 1983.
37. Y. Tsividis and G. Masetti, "Problems in precision modeling of the MOS transistor for analog applications," *IEEE Transactions on Computer-Aided Design*, vol. CAD-3, pp. 72–79, 1984.
38. J. J. Paulos, "Measurement and modeling of small-geometry MOS transistor capacitances," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, September 1984.
39. M. Bagheri and Y. Tsividis, "A small-signal dc-to-high-frequency non-quasi-static model for the four-terminal MOSFET valid in all regions of operation," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2383–2391, 1985.
40. M. Bagheri, "Improving the non-quasi-static weak-to-strong-inversion four-terminal MOSFET model," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 2558–2560, 1987.
41. K.-C. Chai and J. J. Paulos, "Comparison of quasi-static and non-quasi-static capacitance models for the four-terminal MOSFET," *IEEE Electron Device Letters*, vol. EDL-8, pp. 377–379, 1987.
42. U. Kumar, "Modified charge-control model for MOS transistors in pre-saturation region," *Solid-State Electronics*, vol. 31, pp. 999–1001, 1988.
43. K.-W. Chai and J. J. Paulos, "Unified nonquasi-static modeling of the long-channel four-terminal MOSFET for large- and small-signal analyses in all operating regimes," *IEEE Transactions on Electron Devices*, vol. 36, pp. 2513–2520, 1989.
44. P. J. V. Vandelooy and W. M. C. Sansen, "Modeling of the MOS transistor for high frequency analog design," *IEEE Transactions on Computer-Aided Design*, vol. 8, pp. 713–723, 1989.
45. L.-J. Pu and Y. Tsividis, "Small-signal parameters and thermal noise of the four-terminal MOSFET in non-quasi-static operation," *Solid-State Electronics*, vol. 33, p. 521, 1990.
46. L.-J. Pu and Y. Tsividis, "Harmonic distortion of the four-terminal MOSFET in non-quasi-static operation," *IEE Proceedings*, vol. 137, pp. 325–332, 1990.
47. K. A. Sakallah, Y.-T. Yen, and S. S. Greenberg, "A first-order charge conserving MOS capacitance model," *IEEE Transactions on Computer-Aided Design*, vol. 9, pp. 99–108, 1990.
48. P. Roblin, S. C. Kang, and W.-R. Liou, "Improved small-signal equivalent circuit model and large-signal state equations for the MOSFET/MODFET wave equation," *IEEE Transactions on Electron Devices*, vol. 38, pp. 1706–1718, 1991.

49. H. J. Park, P. K. Ko, and C. Hu, "A non-quasi-static MOSFET model for SPICE-AC analysis," *IEEE Transactions on Computer-Aided Design*, vol. 11, pp. 1247–1257, 1992.
50. S. C. Kang and P. Roblin, "Optimal second-order small-signal model for long- and short-channel three-terminal MOSFET/MODFET wave equation," *IEEE Transactions on Electron Devices*, vol. 39, pp. 1909–1915, August 1992.
51. R. Singh, A. Juge, R. Joly, and G. Morin, "An investigation into the non-quasi-static effects in MOS devices with on-wafer s-parameter techniques," *Proceedings of the International Conference on Microelectronic Test Structures*, pp. 21–25, 1993.
52. T. Smedes and F. M. Klaassen, "An analytical model for the non-quasi-static small-signal behaviour of submicron MOSFETs," *Solid-State Electronics*, vol. 38, pp. 121–130, 1995.
53. J. Katzenelson and A. Unikovski, "A network charge-oriented MOS transistor model," *International Journal of High Speed Electronics and Systems*, vol. 6, pp. 285–316, 1995.
54. V. I. Kol'dyacv, A. Clerix, and L. Deferm, "Closed-form frequency dependent gate-to-channel capacitance model for submicron MOSFETs," *Proceedings of the 36th European Solid-State Device Research Conference*, Bologna, Italy, September 1996, pp. 679–682.
55. Y. Niitsu, "Simple small-signal model for 3-port MOS transistors," *IEICE Transactions on Electronics*, vol. E79-C, pp. 1760–1765, 1996.
56. J. J. Paulos, D. A. Antoniadis, and Y. P. Tsividis, "Measurement of intrinsic capacitances of MOS transistors," *Proceedings of the IEEE International Solid-State Circuits Conference*, San Francisco, February 1982, pp. 238–239.
57. C. A. Desoer and E. S. Kuh, *Basic Circuit Theory*, McGraw-Hill, New York, 1969.
58. W. H. Hayt, Jr., and J. E. Kemmerly, *Engineering Circuit Analysis*, McGraw-Hill, New York, 1993.
59. W. H. Kim and H. E. Meadows, Jr., *Modern Network Analysis*, John Wiley, New York, 1971.
60. C. Galup-Montoro and M. C. Schneider, *MOSFET Modeling for Circuit Analysis and Design*, World Scientific, Singapore, 2007.
61. C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling*, Wiley, Hoboken, N.J., 2006.
62. M. Shoji, "Analysis of high-frequency thermal noise of enhancement mode MOS field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 520–524, 1966.
63. H. E. Halladay and A. Van Der Ziel, "On the high frequency excess noise and equivalent circuit representation of the MOS-FET with n-type channel," *Solid-State Electronics*, vol. 12, pp. 161–176, 1969.
64. F. M. Klaassen, "A combination of the high-frequency noise quantities of a MOS-FET," *Philips Research Reports*, vol. 24, pp. 559–571, 1969.
65. F. M. Klaassen and J. Prins, "Noise of field-effect transistors at very high frequencies," *IEEE Transactions on Electron Devices*, vol. ED-16, pp. 952–957, 1969.
66. P. S. Rao and A. Van Der Ziel, "Noise and y-parameters in MOSFETs," *Solid-State Electronics*, vol. 14, pp. 939–944, 1971.
67. E. W. Kirk and A. van Der Ziel, "Induced gate noise in MOSFETs," *Solid-State Electronics*, vol. 14, pp. 945–948, 1971.
68. A. Ambrozy, *Electronic Noise*, McGraw-Hill, New York, 1982.
69. A. van der Ziel, *Noise in Solid-State Devices and Circuits*, Wiley, New York, 1986.
70. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
71. D. P. Triantis, A. N. Birbas, and S. E. Plevridis, "Induced gate noise MOSFETs revisited: The submicron case," *Solid-State Electronics*, vol. 41, pp. 1937–1942, 1997.
72. R. P. Jindal, "Effect of induced gate noise at zero drain bias in field-effect transistors," *IEEE Transactions on Electron Devices*, vol. 52, no. 3, pp. 432–434, 2005.
73. J. C. J. Paasschens, A. J. Scholte, and R. van Langevelde, "Generalizations of the Klaassen-Prins equations for calculating the noise of semiconductor devices," *IEEE Transactions on Electron Devices*, vol. 52, no. 11, pp. 2463–2472, 2005.
74. PSP website. Available at <http://pspmodel.asu.edu/>. Accessed Oct. 1, 2009.
75. C. Raynaud, J. Gautier, G. Guegan, M. Lerme, E. Playez, and G. Dambrine, "High-frequency performance of submicrometer channel-length silicon MOSFETs," *IEEE Electron Device Letters*, vol. 12, pp. 667–669, 1991.

76. A. E. Schmitz, R. H. Walden, L. E. Larson, S. E. Rosenbaum, R. A. Metzger, J. R. Behnke, and P. A. Macdonald, "A deep-submicrometer microwave/digital CMOS/SOS technology," *IEEE Electron Devices Letters*, vol. 12, pp. 16–17, 1991.
77. N. Camilleri, J. Costa, D. Lovelace, and D. Ngo, "Silicon MOSFETs, the microwave device technology for the 90s," *IEEE MTT-S Digest*, pp. 545–548, 1993.
78. J. C. Costa, D. Lovelace, D. Ngo, and N. Camilleri, "Modeling a new generation of RF devices: MOSFETs for L-band applications," *IEEE MTT-S Digest*, pp. 293–296, 1993.
79. D. Lovelace, J. Costa, and N. Camilleri, "Extracting small-signal model parameters of silicon MOSFET transistors," *IEEE-MTT-S Digest*, pp. 865–868, 1994.
80. R. R. J. Vanoppen, J. A. M. Geelen, and D. B. M. Klaassen, "The high-frequency analogue performance of MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 173–176, December 1994.
81. J. M. Collantes, J. J. Raoux, J. P. Villotte, R. Quere, G. Montoriol, and F. Dupis, "A new large-signal model based on pulse measurement techniques for RF power MOSFET," *IEEE MTT-S Digest*, pp. 1553–1556, 1995.
82. D. B. M. Klaassen, B. Nauta, and R. R. J. Vanoppen, "RF modelling of MOSFETs," in *Analog Circuit Design: MOST RF Circuits, Sigma-Delta Converters, and Translinear Circuits*, W. Sansen, J. H. Huijsing, and R. J. v. d. Plassche (editors), Kluwer Academic Publishers, Boston, 1996.
83. S. Lee, H. K. Yu, C. S. Kim, J. G. Koo, and K. S. Nam, "A novel approach to extracting small-signal model parameters of silicon MOSFETs," *IEEE Microwave and Guided Wave Letters*, pp. 75–77, March 1997.
84. M. C. Ho, K. Green, R. Culbertson, J. Y. Yang, D. Ladwig, and P. Ehnis, "A physical large signal Si MOSFET model for RF circuit design," *IEEE MTT-S Digest*, pp. 391–394, 1997.
85. M. Miller, T. Dinh, and E. Shumate, "A new empirical large signal model for silicon RF LDMOS FETs," *IEEE MTT-S Digest*, Vancouver, 1997, pp. 19–22.
86. M. C. Ho, F. Brauchler, and J. Y. Yang, "Scalable RF Si MOS-FET distributed lumped element model based on BSIM3v3," *Electronics Letters*, vol. 33, pp. 1992–1993, 1997.
87. H. C. Lin, Y. F. Arzoumanian, J. L. Halsor, M. N. Giuliano, and H. F. Benz, "Effect of silicon gate resistance on the frequency response of MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 255–264, 1975.
88. T. Sakurai and T. Iizuka, "Gate electrode RC delay effects in VLSIs," *IEEE Journal of Solid-State Circuits*, vol. SC-20, pp. 290–294, 1985.
89. L.-S. Kim and R. W. Dutton, "Modeling of the distributed gate RC effect in MOSFETs," *IEEE Transactions on Computer-Aided Design*, vol. 8, pp. 1365–1367, 1989.
90. B. Razavi, R.-H. Yan, and K. F. Lee, "Impact of distributed gate resistance on the performance of MOS devices," *IEEE Transactions on Circuits and Systems—I*, vol. 41, pp. 750–754, 1994.
91. E. Abou-Allan and T. Manku, "A small-signal MOSFET model for radio frequency IC applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 16, pp. 437–447, 1997.
92. K. K. Thornber, "Resistive-gate-induced thermal noise in IGFETs," *IEEE J. Solid-State Circuits*, vol. 16, no. 4, pp. 414–415, 1981.
93. R. P. Jindal, "Noise associated with distributed resistance of MOSFET gate structures in integrated circuits," *IEEE Transactions on Electron Devices*, vol. 31, no. 10, pp. 1505–1509, 1984.
94. I. Bahl and P. Bhartia, *Microwave Solid-State Circuit Design*, Wiley, New York, 1988.
95. E. S. Kuh and R. A. Rohrer, *Theory of Linear Active Networks*, Holden-Day, San Francisco, 1967.
96. C. C. Enz and Y. Cheng, "MOS transistor modeling for RF IC design," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, pp. 186–201, 2000.
97. Y. Cheng, M. J. Deen, and C.-H. Chen, "MOSFET modeling for RF IC design," *IEEE Transactions on Electron Devices*, vol. 52, no. 7, pp. 1286–1303, 2005.
98. M. J. Deen and T. A. Fjeldly, *CMOS RF Modeling, Characterization and Applications*, World Scientific, Singapore, 2002.
99. T. Ytterdal, Y. Cheng, and T. A. Fjeldly, *Device Modeling for Analog and RF CMOS Circuit Design*, Wiley, Hoboken, N.J., 2003.
100. W. Grabinski, B. Nauwelaers, and D. Schreurs (editors), *Transistor Level Modeling for Analog/RF IC Design*, Springer, Boston, 2006.

Problems

- 8.1.** For the case of Fig. P8.1 (where $q_2 = -q_1$), relate the various quantities by writing equations analogous to (6.3.16) and (8.2.2). Show that in this case the signs used in the definition (8.2.1) represent a “natural” choice.

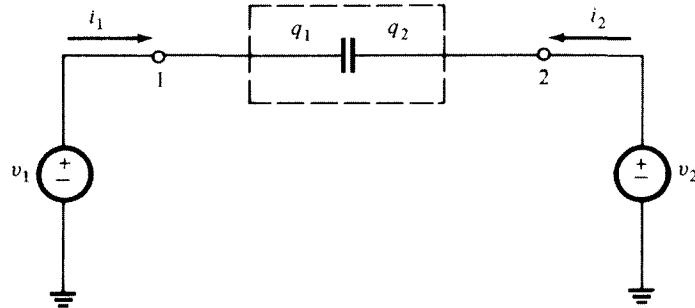


FIGURE P8.1

- 8.2.** We define the “nonreciprocity” between two capacitance parameters C_{kl} and C_{lk} ($l \neq k$) as the quantity $C_{kl} - C_{lk}$. Using (8.2.8), derive relations among the various nonreciprocities and comment on the result.
- 8.3.** Using Kirchhoff’s current law for all terminals, check the validity of the model in Fig. 8.4b.
- 8.4.** Prove in detail that (8.2.19) and (8.2.20) are equivalent to (8.2.12).
- 8.5.** Consider the model of Fig. 8.5. For each of 16 ordered pairs of terminals (k, l) perform the following experiment. Apply a voltage v_l between terminal l and ground, keeping all other terminals grounded, and determine the current i_k entering terminal k . Show that the capacitive part of this current is given by $C_{kl}(dv_l/dt)$ if $l = k$, and by $-C_{kl}(dv_l/dt)$ if $l \neq k$. Note that six of the C_{kl} appear directly in the model, whereas others do not, in which case (8.2.8) or (8.2.20) will have to be evoked.
- 8.6.** Show that results identical to those in Prob. 8.5 would be obtained if the model of Fig. 7.19 were used instead for five of the terminal pair combinations (those that correspond to the subscripts of the capacitors in Fig. 7.19). Show that for other terminal pairs the results will differ from those in Prob. 8.5; this points to the incompleteness of the simple model of Fig. 7.19. Explain why this incompleteness does not matter at low frequencies.
- 8.7.** Prove (8.2.21) to (8.2.27).
- 8.8.** An intuitive argument was given following (8.2.24) concerning the fact that C_{sd} is negative in nonsaturation. Using similar arguments, explain why C_{ds} is also negative, whereas all other capacitances C_{kl} are positive.
- 8.9.** Using the simplified strong-inversion model, $C_{mx} = 0$ was found. Allow for a nonzero C_{mx} , and show that the expressions for the capacitances would have to be modified so that (8.2.8) remains valid.
- 8.10.** This problem continues the comparison between the simple model of Fig. 7.19 and the complete quasi-static model of Fig. 8.5, started in Prob. 8.6. Consider operation in the saturation region. Using the small-signal quantity values found for the simple model in Secs. 7.2 and 7.3, justify the upper frequency limit of validity given for the model of Fig. 7.19 following (7.3.6). Concentrate on the effect of V_g on i_d in saturation.

- 8.11.** Prove (8.3.7).
- 8.12.** Prove (8.3.9).
- 8.13.** Provide all the details in the proofs of (8.4.36) to (8.4.48).
- 8.14.** Prove (8.4.51) and (8.4.53) in detail.
- 8.15.** (*Note:* This problem requires *extremely lengthy* derivations. It is not meant as a homework problem.) Solve (8.4.51) and (8.4.52) using the following iterative procedure.²⁰ Replace x in (8.4.51) with a dummy variable \hat{x} . Develop an integral form of (8.4.51a) and (8.4.51b), by integrating from x to L . Use as an initial approximation $I_i(x, \omega) = I_i(L, \omega)$ (which is actually exact for $\omega = 0$). Substitute this in the integral form of (8.4.51a) and solve for the corresponding approximation of $U_i(x, \omega)$. Use this in the integral form of (8.4.51b) to find a new approximation for $I_i(x, \omega)$. Use this in the integral form of (8.4.51a) and solve for the new approximation of $U_i(x, \omega)$, etc. Show that the procedure produces a solution from which one can obtain (8.4.58) to (8.4.60), with the values given in Appendix 1.
- 8.16.** Prove (8.4.65) to (8.4.67), using results from Appendix 1. (*Note:* Very lengthy derivations are involved.)
- 8.17.** (a) Plot the magnitude and phase of $-y_{gb}$ vs. ω (for ω up to ω_o on a log axis) in the saturation region and at $V_{DS} = 0$, using (8.4.69e) with and without including y_a . Use an oxide thickness of 2 nm, $\alpha = 1.1$, $W = 1 \mu\text{m}$, $L = 0.2 \mu\text{m}$, $\mu = 450 \text{ cm}^2/(\text{V}\cdot\text{s})$, and $V_{GS} - V_T = 0.5 \text{ V}$. Comment on the validity of approximating $-y_{gb}$ by $j\omega C_{gb}$ (which is equivalent to deleting the box shown in broken lines in Figs. 8.20, 8.23, and 8.24). Comment on currents that, in practice, are likely to mask the current due to y_{gb} .
- (b) Assume $V_{DS} = 0$. Show instances where a terminal current being observed is only due to y_{gb} (and to extrinsic elements). Explain why such instances are unlikely to be of significance in practice and that, even then, the magnitude of the observed current is very small and likely to be dominated by extrinsic effects.
- 8.18.** Verify the two circuit representations in Fig. 8.19.
- 8.19.** Using the approximation in (8.4.72), show that the model in Fig. 8.20 reduces at low frequencies to the model of Fig. 8.5, in both form and element values (y_{m_1} is assumed zero), and that at even lower frequencies it further reduces to the model of Fig. 7.19.
- 8.20.** The intrinsic cutoff frequency of a transistor was defined in Sec. 7.3, and in the strong-inversion saturation region was found to be given by (7.3.32), by using the model of Fig. 7.19. As mentioned following that equation, the model of Fig. 7.19 is not accurate at frequencies as high as $\omega = \omega_T$, and therefore the result (7.3.32) should be checked by using a more accurate model. Show that, by using the model of Fig. 8.23, one obtains practically the same value for ω_T .
- 8.21.** Prove all the values given in the model in Fig. 8.24.
- 8.22.** Provide plots vs. ω for normalized magnitude and phase of all parameters in Fig. 8.17 in the saturation region, for the non-quasi-static model of Sec. 8.4.
- 8.23.** Consider a transistor with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $t_{ox} = 2 \text{ nm}$, $V_{T0} = 0.4 \text{ V}$, $\phi_0 = 1.0 \text{ V}$, $\mu = 450 \text{ cm}^2/(\text{V}\cdot\text{s})$, $W = 1 \mu\text{m}$, and $L = 0.2 \mu\text{m}$, connected as shown in Fig. 8.1b with $V_B = 0.4 \text{ V}$, $V_S = 0.8 \text{ V}$, $V_G = 2 \text{ V}$, and $V_D = 1.2 \text{ V}$. Calculate all parameters for the model of Fig. 8.23.
- 8.24.** For the device of Prob. 8.23, find the magnitude and phase of all terminal currents (Fig. 8.1b) if the small-signal voltages $v_b(t)$, $v_s(t)$, $v_g(t)$, and $v_d(t)$ are of the form $a \cos(\omega t + \phi)$, with magnitudes a of 2, 1, 4, and 3 mV, respectively, and phases of 0° , 0° , 40° , and 50° , respectively. The frequency ω is $2\pi \times 100 \text{ GHz}$.

- 8.25.** The circuit in Fig. P8.2 has little to do with a transistor, but it will help in getting a feeling for the various levels of approximation involved in modeling.
- Find the admittance y_{AB} , its real part y_{ABr} , and its imaginary part y_{ABI} .
 - Find the conditions under which C can be neglected in each of the preceding three expressions, and give a simplified model.
 - Find the conditions under which R_2 can be neglected in each of the three expressions in (a), and give a simplified model.
 - Show that neglecting R_2 is equivalent to assuming that the charge Q on the top plate depends quasi-statically on the voltage V defined in the figure. Show that at high frequencies this quasi-static dependence breaks down.

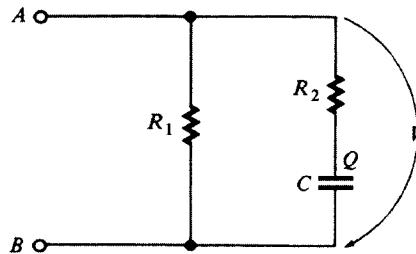
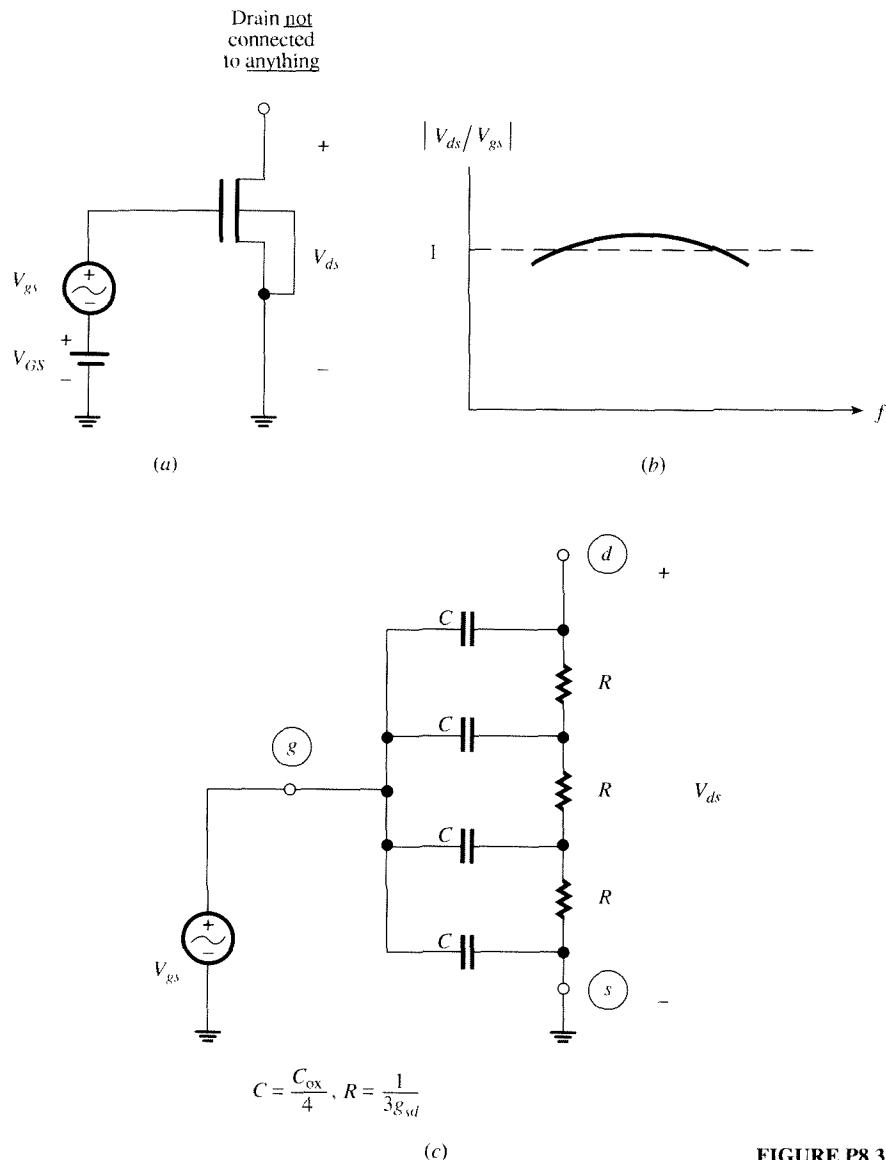


FIGURE P8.2

- 8.26.** Show that the circuits of Fig. 8.22b and c both predict the behavior in (8.4.71), and find the proper value to be used for C in order for this behavior to be identical even quantitatively. Attempt to explain the value found for C .
- 8.27.** Consider a MOS transistor with its body connected to source, driven by a bias plus small-signal voltage between gate and source, and with its drain not connected to anything (Fig. P8.3a). In this connection $I_D = 0$, $V_{DS} = 0$, and the channel is uniform. Derive an expression for the small-signal voltage gain of this device, V_{ds}/V_{gs} , where V_{ds} and V_{gs} are phasors corresponding to small-signal voltages, assuming that the substrate doping is very light, and thus the substrate parasitics can be neglected; also, assume that W and L are very large, so that intrinsic effects dominate and extrinsic parasitics can be neglected. Show that $|V_{ds}/V_{gs}|$ behaves as in Fig. P8.3b; i.e., it can exceed unity in a range of frequencies. Show that this behavior is also predicted by the simple lumped model of Fig. P8.3c. Are the findings in this problem strange? Why, or why not?
- 8.28.** Derive (8.5.3) from (8.5.2).

**FIGURE P8.3**

546 OPERATION AND MODELING OF THE MOS TRANSISTOR

- 8.29.** Shown in Fig. P8.4 is the model of Fig. 8.24 in saturation with three important extrinsic parasitics added, assuming that the substrate is short-circuited to the source and that C_{gb} and L_{sd} are negligible. Assume $R_{de} \ll 1/g_{sd}$ and $\omega \ll 1/(C_{bde}R_{de})$. Derive expressions for the y -parameters y_{gg} , y_{gd} , y_{dd} , and y_{dg} . Show qualitative plots for their real and imaginary parts vs. ω on log-log axes for (a) zero C_{gde} , C_{bde} , and R_{de} , and (b) nonzero C_{gde} , C_{bde} , and R_{de} . Discuss the important effects caused by the extrinsic elements.

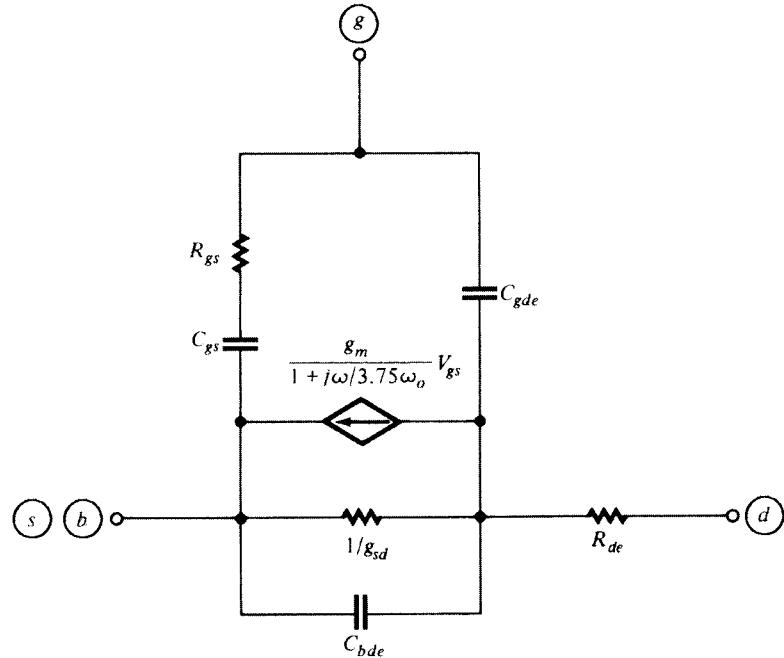


FIGURE P8.4

Substrate Nonuniformity And Other Structural Effects

9.1 Introduction

Most of the fundamental physics that underlie MOS transistor operation have been discussed in the previous chapters of this book. Along the way, we have made several simplifying assumptions, such as substrate uniformity (with a few exceptions). We have also implicitly assumed that the transistors are not affected by nearby structures, and that the physical parameters that characterize their behavior are known exactly. In this chapter, we will discuss additional considerations that arise when such assumptions are relaxed. We will take a closer look at the actual structure of a realistic device and its surroundings, and we will discuss how the models we have developed so far may have to be modified to account for the details of this structure. We will call the effects discussed in this chapter “structural effects.” Fortunately, it turns out that the models developed so far can still be used in the presence of such effects, albeit with certain modifications to the value or behavior of some parameters in them.

The first section below briefly discusses the process of ion implantation, and the next two sections discuss how this process is used to introduce substrate non-uniformity, separately along the direction perpendicular to the surface, and along the direction parallel to it, to improve transistor characteristics. Then, we discuss the influence on device characteristics of nearby structures, specifically the edge of the well in CMOS structures and the isolation region in STI processes, which results in mechanical stress. Finally, we consider the variability of transistor parameters and discuss how it can be described using simple statistics.

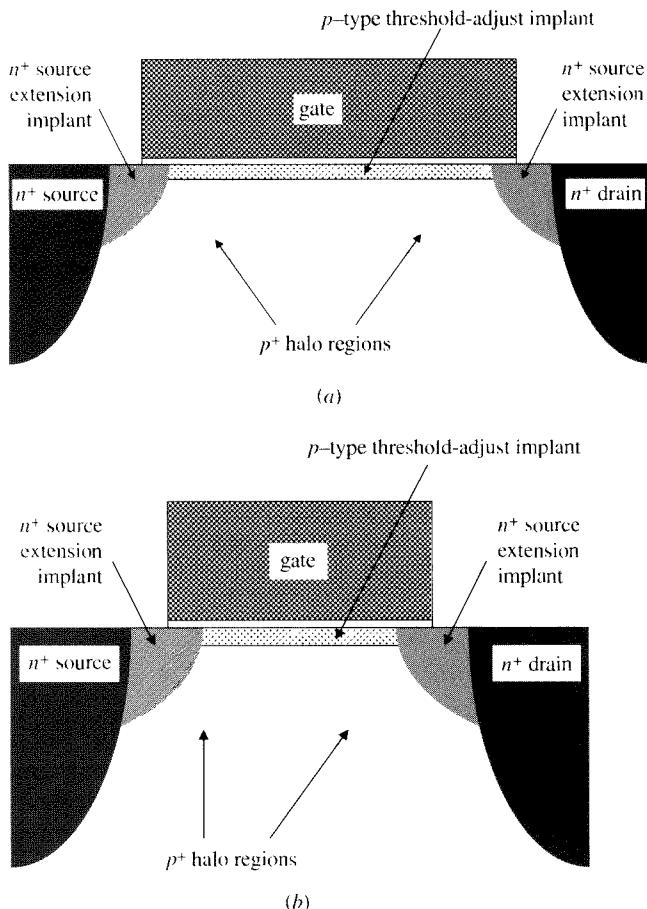


FIGURE 9.1
Cross section of the semiconductor part of a typical device (a) long channel, (b) short channel.

9.2 Ion Implantation and Substrate Nonuniformity

In the ion implantation process, which takes place during transistor fabrication, atoms of a particular dopant are ionized and then accelerated by an electric field and directed at the surface of a wafer. (Often boron or BF₂ is used for *p*-type doping, and arsenic or phosphorus for *n*-type doping, but other elements are also used.) The ions, which have kinetic energy from the acceleration by the electric field, penetrate into the silicon in regions where there is nothing to block the ion flux (such as photoresist, gate polysilicon, and dielectric material used for isolation). The implantation leads to a vertically nonuniform doping profile. The implanted ions statistically lose energy by scattering events with the silicon atoms that make up the bulk, so there is a characteristic depth to which the ions on average penetrate the silicon bulk, and a spread around this depth that depends on the statistics of the scattering events, giving a roughly Gaussian-shaped doping profile. The depth, spread, and density of the implant depends on several factors. One is the average kinetic energy of the implanted ions; the higher

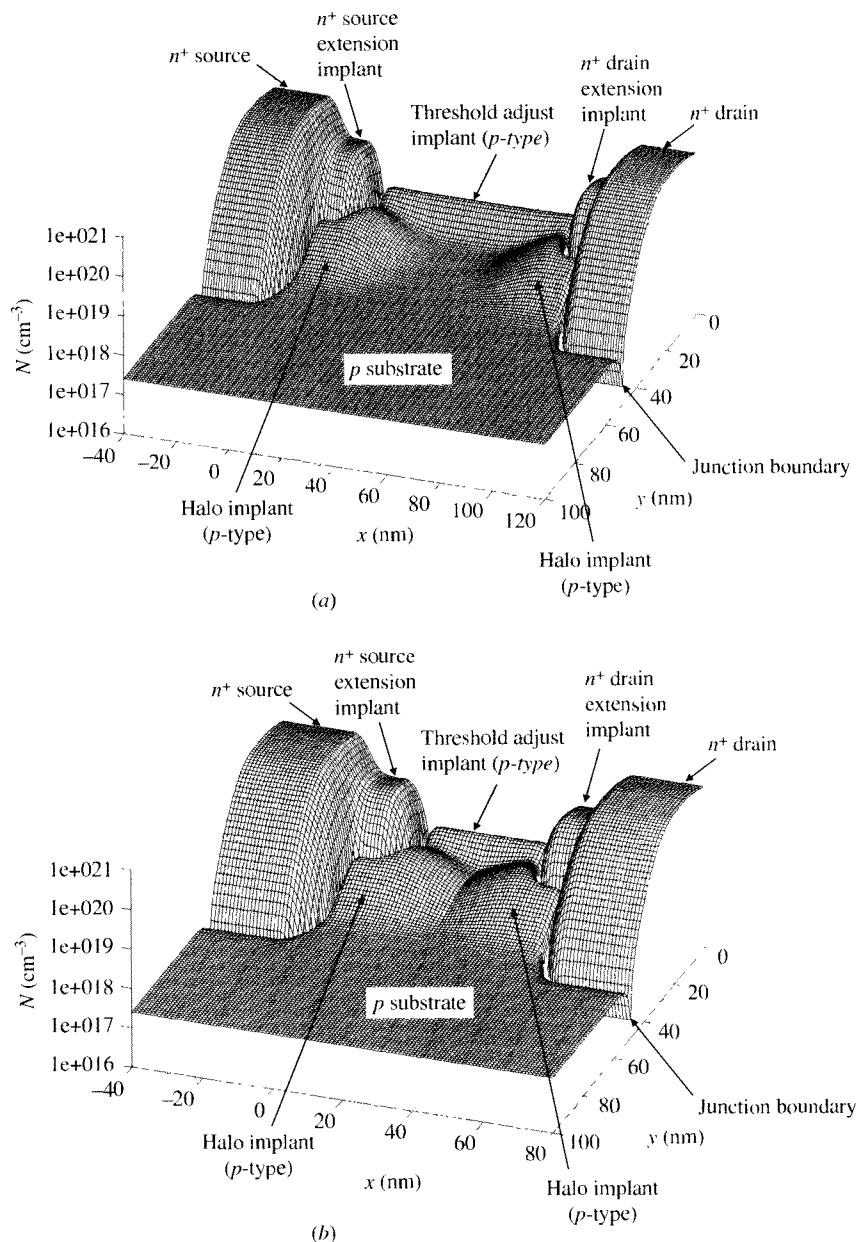
the energy, the deeper the ions penetrate into the silicon. Energies used are typically between 5 and 300 keV. Another factor is the angle of the implant. For vertical implants, the ions can “channel” between the silicon atoms in the lattice, which allows some implanted ions to avoid scattering events with the silicon lattice and so to penetrate deeply into the silicon; this is undesirable, so implants are commonly done at some “tilt” angle to avoid channeling. Finally, the result depends on the flux density and implantation time of the ion beam, both of which result in the *dose* (number of implanted ions within the semiconductor per unit area, typically 10^{11} to 10^{13} per cm^2); other things being equal, a greater dose means a higher final doping concentration. The implant *profile* (the resulting concentration of ions as a function of depth) is modified by subsequent annealing steps in the semiconductor manufacturing process, whereby a wafer is heated to a certain temperature for a specific amount of time. This causes the implanted dopants to migrate due to diffusion, which broadens the doping profile.

Since its early application to MOS transistors,^{1–4} ion implantation has been used extensively and is currently a standard part of device fabrication.^{5–8} Ion implantation is used for more than doping the channel; for example, *n*MOS devices are formed by using ion implantation to dope the source and drain regions n^+ ;² during this process, the polysilicon gate acts as a mask, preventing the ions from landing in the channel area. In this process, the gate itself receives the ions and is thus doped n^+ . However, here we will concentrate on the effects of ion implantation of the *channel* region. The characteristics of devices with ion-implanted channels have been the subject of extensive studies.^{9–97}

Tailoring Device Characteristics through Ion Implantation The process of ion implantation is used to selectively modify substrate doping to tailor several characteristics. A cross section of a modern device is shown in Fig. 9.1a. Note that in addition to the *p* substrate and the source/drain n^+ regions, we have the following:

- *Source/drain extensions:* These are used to provide shallow junctions adjacent to the channel region. The main source/drain regions are deep to ensure low resistance, thus limiting the series resistance effects discussed in Sec. 4.12. At the same time, these regions are kept away from the channel with the help of the extensions, and thus the depletion regions along the depth of the main regions are kept at a safe distance from the intrinsic part of the transistor. Thus, the use of the extensions helps limit short-channel effects such as charge sharing and DIBL, discussed in Chap. 5.
- *Halo regions:* To further limit the extent of the depletion regions around source and drain, halo regions are used around them. Halo regions are doped heavier than the main substrate, as discussed in Sec. 5.4.2. The heavy doping means that the width of the depletion regions is limited, as follows from (1.5.16), which keeps short-channel effects in check (Sec. 5.4.2). The modeling of halo implants effects is discussed in Sec. 9.4.
- *Threshold-adjust implants:* These are used to set the value of the threshold voltage. At the same time, the main substrate doping is kept low, which is desirable for low body effect (Fig. 2.11) and low junction capacitance [see (1.5.24)]. The use of this implant will be discussed in detail in the following sections.

Other considerations in the use of implants involves the modification of the surface electric field, which, as can be inferred from (2.6.5), (2.6.6), and Fig. 4.32, affects the effective mobility.

**FIGURE 9.2**

Magnitude of net doping concentration vs. position (x) and depth from the surface (y) for devices representative of a 45-nm process: (a) long channel, (b) short channel.

When the channel length is reduced to the minimum allowed by the fabrication technology, the two halo regions can merge, as shown in Fig. 9.1*b*. This tends to make the substrate doping approximately uniform, with concentration corresponding to that of the halo regions.

Figure 9.2*a* and *b* shows the net doping concentration vs. length (*x*) and depth (*y*) for the cases corresponding to Fig. 9.1*a* and *b*. More precisely, what is plotted is $N = |N_A - N_D|$, where N_A is the acceptor concentration and N_D the donor concentration at a given point. Note that the *pn* junctions involved are gradual; i.e., the doping concentration changes gradually from *n*-type to *p*-type. Thus, N must pass through zero at the point where N_A and N_D become equal; because the axis is logarithmic, this causes a dip in the plot, as seen, which defines the boundary of the junction. One such boundary is indicated on the plot. The values shown for net concentration, length, and depth are realistic for a transistor in a 45-nm process.

As is obvious from Figs. 9.1 and 9.2, the body in the channel region is in general nonuniform in both the transverse direction (i.e., with depth from the Si-SiO₂ interface) and in the longitudinal direction (i.e., along the direction of current flow from source to drain). In order to be able to discuss the results of nonuniformity in a simple manner, and to study separately the various effects involved, we will consider nonuniformity in each direction separately. We begin by assuming that the substrate is nonuniform only in the transverse direction.

9.3 Substrate Transverse Nonuniformity

9.3.1 Preliminaries

There are two primary reasons why the substrate doping concentration of a MOS transistor is not uniform in the transverse direction. First, many transistors are formed in regions, called “wells” or “tubs” (Sec. 1.7) that are formed by ion implantation. The effective substrate doping concentration is changed in the areas where these ions end up and thus varies as one goes from the surface toward the bulk.[†] Second, ion implantation is used to selectively modify the bulk doping concentration at various depths in order to modify the transistor characteristics to advantage.

We now discuss the use of ion implantation to optimize the value of the threshold voltage. Let us first consider an unimplanted *n*MOS device. The zero- V_{SB} threshold voltage, V_{T0} , is given by (2.6.21), repeated here:

$$V_{T0} = \phi_{MS} - \frac{Q'_o}{C'_{ox}} + \phi_0 - \frac{Q'_B}{C'_{ox}} \quad (9.3.1)$$

where Q'_B is the depletion region charge in strong inversion, given by (2.6.19):

$$Q'_B = -\sqrt{2q\epsilon_s N_A \phi_0} \quad (9.3.2)$$

[†] Even for unimplanted channels the assumption of a uniform substrate is not strictly correct because of some impurity redistribution that takes place during oxide growth.

This equation can thus be written as in (2.6.22):

$$V_{T0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0} \quad (9.3.3)$$

with, from (2.2.6)

$$V_{FB} = \phi_{MS} - \frac{Q'_0}{C'_{ox}} \quad (9.3.4)$$

and, from (2.4.26a)

$$\gamma = \frac{\sqrt{2q\epsilon_s N_A}}{C'_{ox}} \quad (9.3.5)$$

If V_{SB} is not zero, the value of the threshold voltage V_T is changed from the preceding value due to the body effect; it is now given by (4.7.19) or (4.7.20a), repeated here:

$$V_T = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0 + V_{SB}} \quad (9.3.6a)$$

$$= V_{T0} + \gamma(\sqrt{\phi_0 + V_{SB}} - \sqrt{\phi_0}) \quad (9.3.6b)$$

For the enhancement-mode devices used in digital circuits, the value of V_{T0} should be large enough to make sure the device current will be negligible with $V_{GS} = 0$ (for which the device will be in weak inversion). This is needed in order to keep the static power dissipation of large digital circuits low, or to make sure that the charges used for capacitative storage in dynamic memories will not leak out. At the same time, V_{T0} should not be too high, so that with the maximum possible value of V_{GS} , large currents can be provided to charge parasitic capacitances fast (as required for high-speed operation). The maximum possible value of V_{GS} is normally equal to the power supply voltage, and the latter is only about 1 V in modern technologies (Chap. 5). Thus, a compromise value must be chosen for V_{T0} , say 0.4 V.

To see what values of V_{T0} can be expected from (9.3.6), one can estimate V_{FB} in (9.3.4) as shown in Sec. 2.2. For modern devices, Q'_0 is small and C'_{ox} is large (because the effective oxide thickness in modern technologies can be as small as 1.2 nm); thus, Q'_0/C'_{ox} can be neglected to first order. The value of ϕ_{MS} depends to some degree on the substrate doping (Secs. 2.2 and 1.4), but not strongly; for n^+ poly gates, and common substrate doping concentrations, typical values for ϕ_{MS} , and thus of V_{FB} , are close to -1 V (see Example 2.2 in Sec. 2.2). If we use for ϕ_0 a typical value of 0.85 V, we can see that $V_{FB} + \phi_0$ in (9.3.3) is about -0.15 V. Thus for this equation to give a sufficiently positive V_{T0} , a large value of γ will be needed. In modern technologies, in which oxides are thin and thus C'_{ox} is large, this means that the substrate doping will have to be high, as follows from (9.3.5). A high substrate doping would

also help keep the depletion region widths around source and drain narrow, thus limiting short-channel effects (Chap. 5).[†]

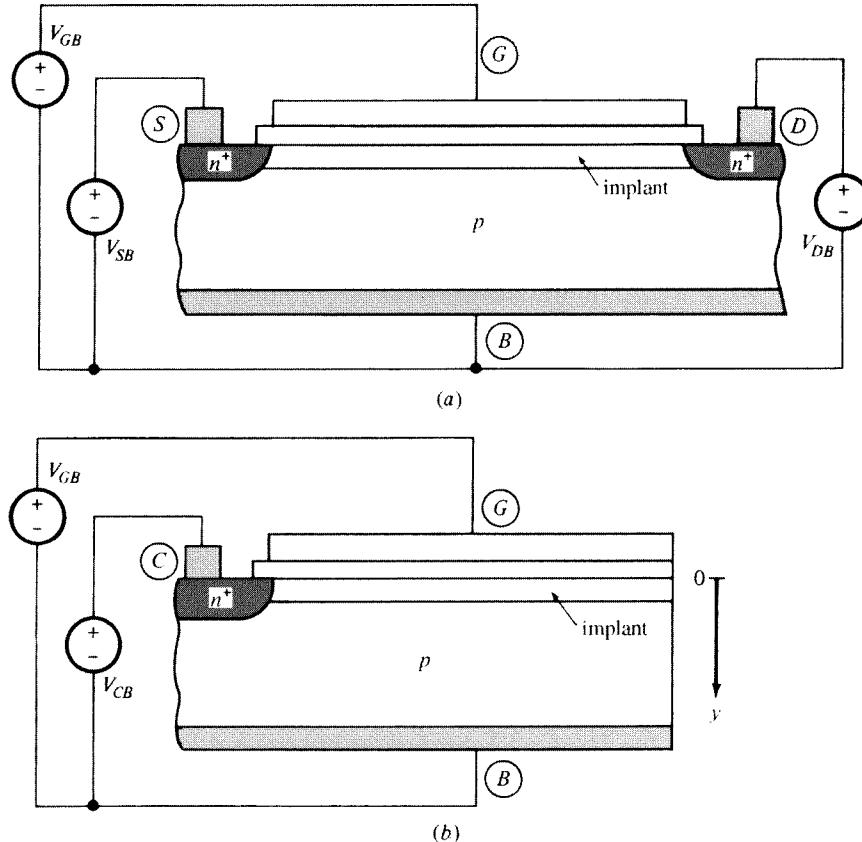
The use of a large γ , though, creates problems. As follows from Sec. 4.8, such a value results in a large value for the quantity n , and this causes a small value for the slope of $\log I_{DS}$ vs. V_{GS} in weak inversion. Thus, when V_{GS} is made zero, it may not be possible to turn the device adequately off in digital circuits. In analog applications, a large γ causes large variations of V_T when V_{SB} changes [see (9.3.6)], which can cause complications in circuit design. Finally, the high substrate doping needed in this solution causes large junction capacitances (Sec. 1.5), which reduces circuit speed (Chaps. 6 to 8). It also increases the electric field at the surface, which reduces mobility.

It is thus seen that there are conflicting considerations as to the proper value of substrate doping concentration. One can solve this problem by realizing that the various issues discussed previously are associated with effects that occur at different depths below the surface. Assume, for example, that the value of N_A chosen from body effect and junction capacitance considerations is low, which by itself would result in too small a value for V_{T0} , as already discussed. To augment the value of V_{T0} , one implants acceptor ions (e.g., boron) very close to the surface. Ionized acceptor atoms have a negative charge, which would be effectively added to Q'_o in (9.3.4), thus changing the effective V_{FB} and providing the required threshold increase. On the other hand, the substrate concentration would not be increased at points below the *p*-type implant where most of the depletion region is located. Thus the body effect coefficient and the junction capacitances would be kept small.

If, instead, the optimum value for the substrate doping from other considerations turns out to be high, e.g., from scaling considerations (Sec. 5.15), the value of V_{T0} that would result would be too large. This now calls for an ion implantation of the *opposite* type from the preceding. Assume that donor ions are implanted very close to the surface. Because ionized donor atoms have a positive charge, they are effectively added to Q'_o in (9.3.4), thus *reducing* the threshold, while not affecting the doping concentration deeper into the substrate.

It can be seen from the previous example that the details of how the substrate doping should be modified are in general different, depending on which of the characteristics in the preceding list one is trying to optimize. Thus, multiple ion implants are often used, each addressing a particular issue, alone or in combination, to tailor the final doping profile and optimize device performance. The trade-offs involved are part of what is known as “device engineering,”⁹⁷ which falls outside the scope of this book. In this chapter, we will mostly discuss how device characteristics can be modeled when their substrate doping concentration vs. depth away from the surface is known, having been set through device engineering. An important aim of this chapter is to show that the equations we have developed so far for the case of uniform substrates can largely continue to be used with nonuniform ones, provided certain adjustments are made. We will illustrate several effects assuming the simple case of a single, threshold-adjust implant.

[†] Such effects can more effectively be addressed using lateral nonuniformity, as discussed in Sec. 9.3.2.

**FIGURE 9.3**

(a) A nMOS transistor with an ion-implanted channel; (b) three-terminal structure used to facilitate the study of (a).

Although the ideal threshold-adjust implant would be a sheet next to the surface as discussed previously, in practice this is not possible, as already mentioned near the beginning of this section; one ends up with a profile, or distribution, of ions vs. depth. This is indicated in Fig. 9.3a for a complete transistor. The corresponding three-terminal structure is shown in Fig. 9.3b; consistent with our practice in past chapters, we will be using the study of this simpler structure as an intermediate step toward developing a model for the complete transistor.

In general, if the doping profile does not vary significantly with depth into the silicon, then variations in behavior of real devices with respect to the uniform-substrate models we have developed so far are small, and experimental data can be fitted well by adjusting parameters such as N_A (which controls ϕ_0 and γ) and V_{FB} . This is because the predominant effect of a variation in transverse doping is to modify how the edge to the depletion region, formed below the gate, varies with bias, which in turn affects how the threshold voltage varies with bulk voltage. Because the threshold voltage varies weakly with doping, through the square root dependence of γ on N_A [see (9.3.5) and (9.3.6)],

using an effective N_A to characterize threshold voltage variation with V_{SB} can be rather effective in modeling experimental data. For devices with doping profiles that vary significantly with depth, simply adjusting N_A and V_{FB} of a model for a uniformly doped device may not be sufficient. We will demonstrate this in the following.

As is illustrated by the previous examples, one of two types of implant can be used. In one, the same doping type is used as the bulk, thus resulting in higher overall concentration near the surface, which gives a so-called high-low profile. In the other, one uses the opposite doping type, thus lowering the bulk doping density at the silicon surface compared with its value deep in the bulk; this results in the so-called low-high, or retrograde, profile. Initially, we analyze the case of a high-low profile; we will then see that the resulting equations can apply to a low-high profile if a simple sign change is made.

To illustrate certain important effects in a simple way, we will initially concentrate on the effect of the implant on the threshold voltage.

9.3.2 Threshold Voltage

High-Low Profile The case of high-low profile is illustrated in Fig. 9.4a. Let N_{AB} be the substrate doping before the implant and $N_i(y)$ be the concentration of implanted *p*-type ions vs. depth y below the surface. Then the effective substrate doping will be $N_{AB} + N_i(y)$, as shown. Because of the complicated shape of the implant profile, a detailed analysis is difficult.⁹⁻⁵⁶ However, useful results can be obtained by approximating this profile as shown in Fig. 9.4b,¹⁶ where N_i and d_i are appropriate constants chosen to make the resulting models as accurate as possible. As a starting point, d_i is sometimes chosen at the point where $N_i(y) = N_{AB}$ in Fig. 9.4a, or at a depth equal to the mean plus the standard deviation of the distribution $N_i(y)$, assuming an initially gaussian shape; d_i can be, for example, 50 nm. N_i is then chosen so that $N_i d_i$ is equal

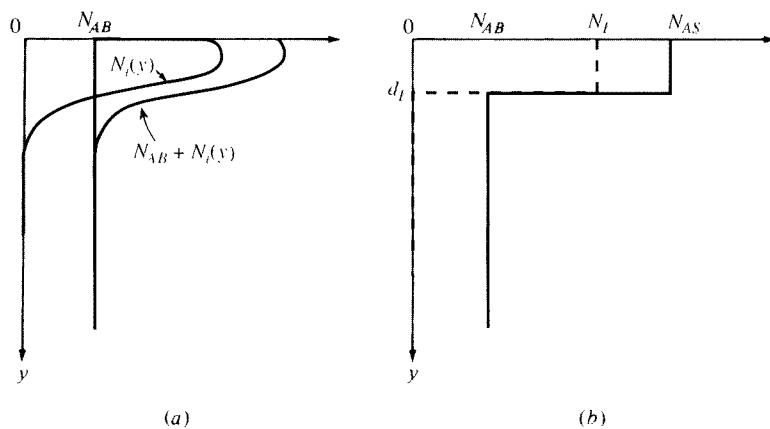


FIGURE 9.4
High-low implant profile: (a) substrate doping concentration before implant N_{AB} , implant concentration $N_i(y)$, and total doping concentration $N_{AB} + N_i(y)$ vs. depth from surface; (b) step approximation for (a).

556 OPERATION AND MODELING OF THE MOS TRANSISTOR

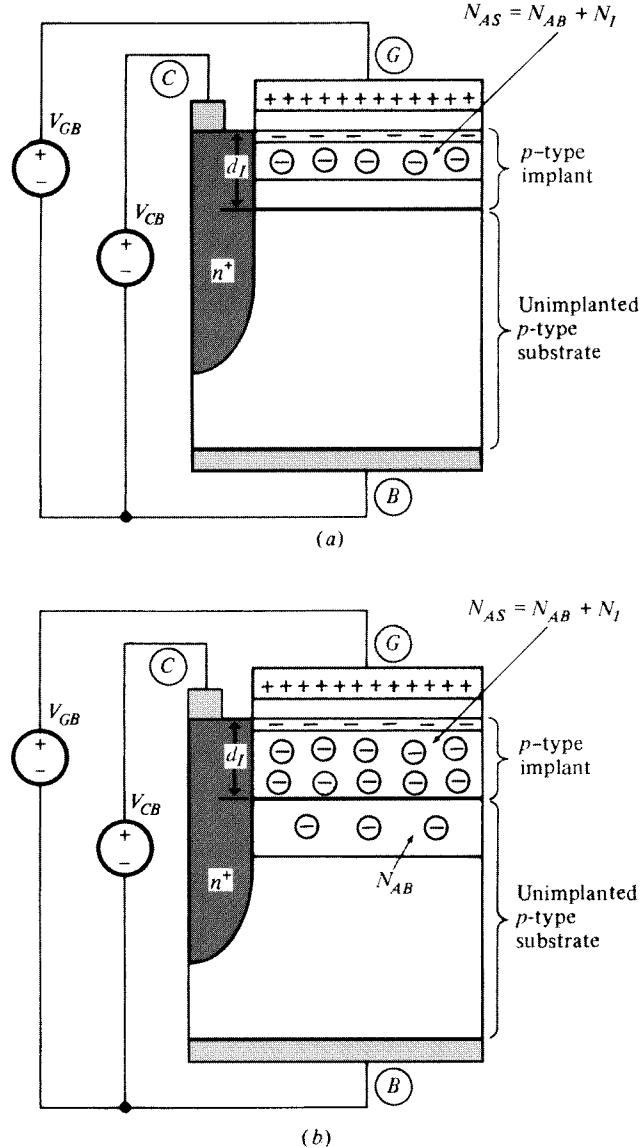


FIGURE 9.5

The structure of Fig. 9.3b with a high-low profile in the strong-inversion region: (a) small V_{CB} (depletion region totally within the implant); (b) large V_{CB} (depletion region extending outside the implant). The depletion region around the n^+ region is not shown for simplicity.

to the total effective dose [the integral under $N_i(y)$ in Fig. 9.4a]. We emphasize that these choices represent only a starting point. The best values for these parameters depend on what aspects of I - V characteristics need be modeled and the bias range over which reasonable accuracy is desired.

Consider now the corresponding three-terminal structure of Fig. 9.3b, assuming a very small V_{CB} . As V_{GB} is increased, a depletion region is formed at the surface. Its width increases with V_{GB} until a strong-inversion layer is created, at which point the depletion region width becomes practically pinned to a certain value d_{Bm} . Assume that this value is smaller than the effective implant depth d_I , as shown in Fig. 9.5a.[†] If now V_{CB} is increased while strong inversion is maintained, the depletion region will widen and its bottom will eventually reach the bottom of the simplified implant profile. The critical value of V_{CB} at which this happens will be denoted by V_I and will be evaluated shortly. For $V_{CB} \leq V_I$, we can think of the gate oxide/implanted region as a device by itself and consider the material below it as part of the substrate contact. The effective substrate doping of this device will be, from Fig. 9.4b,

$$N_{AS} = N_{AB} + N_I \quad (9.3.7)$$

We can use then the results of Chap. 3, with N_A replaced by N_{AS} . The width of the depletion region in strong inversion will be, from (3.4.7),

$$d_{Bm} = \sqrt{\frac{2\epsilon_s}{qN_{AS}}} \sqrt{\phi_{01} + V_{CB}}, \quad V_{CB} \leq V_I \quad (9.3.8)$$

where ϕ_{01} will be somewhat larger than that corresponding to a substrate doping of N_{AB} (Sec. 2.6). The depletion region charge per unit area can be found as in Sec. 3.4.2:

$$Q'_B = Q'_{BI} = -\gamma_1 C'_{ox} \sqrt{\phi_{01} + V_{CB}}, \quad V_{CB} \leq V_I \quad (9.3.9)$$

where

$$\gamma_1 = \frac{\sqrt{2q\epsilon_s N_{AS}}}{C'_{ox}} \quad (9.3.10)$$

The inversion layer charge per unit area will be, again from Sec. 3.4.2,

$$Q'_I = -C'_{ox}(V_{GC} - V_{T1}), \quad V_{CB} \leq V_I \quad (9.3.11)$$

where

$$V_{T1} = V_{FB1} + \phi_{01} - \frac{Q'_{BI}}{C'_{ox}} \quad (9.3.12a)$$

$$= V_{FB1} + \phi_{01} + \gamma_1 \sqrt{\phi_{01} + V_{CB}} \quad (9.3.12b)$$

[†] Note that the depletion region around the n^+ region is not shown for simplicity.

with

$$V_{FBI} \approx V_{FB} \quad (9.3.13)$$

where V_{FB} is the “flatband” voltage corresponding to the unimplanted substrate,[†] given by (9.3.4). The critical value V_I of V_{CB} at which the depletion region bottom reaches the bottom of the implant can be found by setting $V_{CB} = V_I$ and $d_{Bm} = d_I$ in (9.3.8), and solving for V_I . This gives[‡]

$$V_I = \frac{qN_{AS}d_I^2}{2\epsilon_s} - \phi_{01} \quad (9.3.14)$$

Increasing V_{CB} above V_I (assuming V_{GB} is large enough to maintain strong inversion) will move the depletion region bottom outside the implanted region, as shown in Fig. 9.5b. The results obtained earlier obviously do not apply in this case since they were obtained on the assumption that a single doping concentration characterizes the depletion region. In contrast, here we have a region with doping N_{AS} , followed by a region with doping N_{AB} . Let ψ_s be the surface potential, taken across *both* regions. A detailed analysis can be performed along the lines of Chaps. 2 and 3; in fact, (2.3.1), (2.3.4), and (2.4.22) hold unmodified. Such analysis leads to similar conclusions as in Sec. 3.2. In particular, in strong inversion we have again $\psi_s \approx \phi_0 + V_{CB}$, and we will use the value $\phi_0 = \phi_{01}$ as before for simplicity. The resulting *total* depletion region charge per unit area (as contributed by *both* regions in Fig. 9.5b) can be found using basic electrostatics (Appendix A), and is given by (Prob. 9.3)

$$Q'_B = Q'_{B2} = -qM - \gamma_2 C'_{ox} \sqrt{\phi_{01} - \frac{qMd_I}{2\epsilon_s} + V_{CB}}, \quad V_{CB} \geq V_I \quad (9.3.15)$$

where

$$\gamma_2 = \frac{\sqrt{2q\epsilon_s N_{AB}}}{C'_{ox}} \quad (9.3.16)$$

[†] We are neglecting here the small contact potential between the implanted region and the unimplanted substrate. This is consistent with the overall level of approximations in the present analysis. We should note here that even when V_{GB} is such that the surface is neutral, one cannot expect that the electrostatic potential will be constant in the vertical direction *throughout* the substrate, because the dopant and mobile carrier concentration varies in that direction [see (1.2.19)]. Thus the corresponding energy bands will not be “flat” throughout, in contrast to the case of uniform substrates (Sec. 2.2). In implanted device work, the name “flatband voltage” is a carryover from uniform substrate device discussions and does not imply such flatness.

[‡] If the depletion region is within the implant for $V_{CB} = 0$ (as we have assumed in the beginning of this subsection), V_I will be positive. If, instead, the depletion region is deeper than the implant even for $V_{CB} = 0$, V_I will come out negative. This would indicate that the results obtained above for $V_{CB} \leq V_I$ are irrelevant for common applications (in which $V_{CB} \geq 0$), and we should instead consider the results for $V_{CB} > V_I$, which follow.

and

$$M = N_I d_I \quad (9.3.17)$$

is the implant “dose” corresponding to the rectangular shape of the distribution in Fig. 9.4b, measured in ions per cm^2 (or in cm^{-2}). It is seen that (9.3.15) with $M = 0$ reduces to the corresponding equation for an unimplanted device (3.4.8b).

One can find the inversion layer charge per unit area from (9.3.15), (2.3.1), (2.3.4), and (2.4.22) by eliminating among these equations the quantities Q'_G , ψ_{ox} , and Q'_B and using $\psi_s \approx \phi_{01} + V_{CB}$ as before. This gives

$$Q'_I = -C'_{\text{ox}}(V_{GC} - V_{T2}), \quad V_{CB} \geq V_I \quad (9.3.18)$$

with

$$V_{T2} = V_{FB} + \phi_{01} - \frac{Q'_{B2}}{C'_{\text{ox}}} \quad (9.3.19a)$$

$$= V_{FB} + \phi_{01} + \frac{qM}{C'_{\text{ox}}} + \gamma_2 \sqrt{\phi_{01} - \frac{qMd_I}{2\epsilon_s} + V_{CB}} \quad (9.3.19b)$$

As a check, let us consider the case where d_I is negligibly small. Then (9.3.19b) reduces to (3.4.13), with the only difference being that V_{FB} is augmented by the quantity qM/C'_{ox} . This makes sense, since $d_I = 0$ implies that all of the implanted ions are at the surface. Thus they act like the effective interface charge Q'_o in (9.3.4). Since these ions have a negative charge, their charge per unit area is $(-q)M$; thus V_{FB} is effectively augmented by qM/C'_{ox} .

It is seen that (9.3.18) is of the same form as (9.3.11). Also, (9.3.19b) has the same form of functional dependence on V_{CB} as (9.3.12b). Equation 9.3.19b can be written as follows:

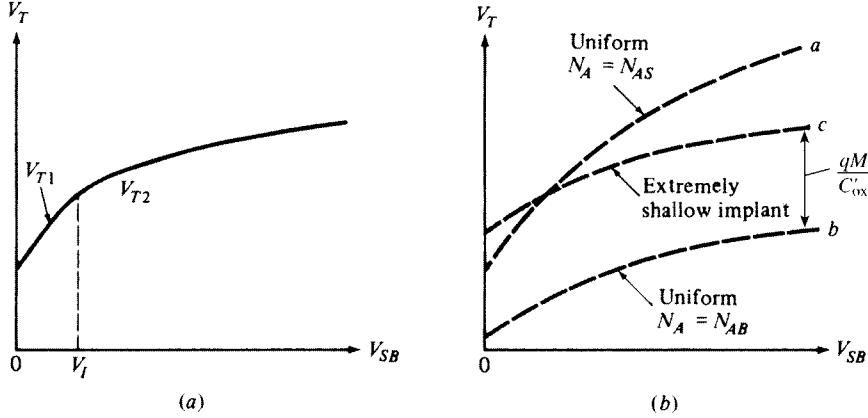
$$V_{T2} = \left[V_{FB} + qM \left(\frac{1}{C'_{\text{ox}}} + \frac{d_I}{2\epsilon_s} \right) \right] + \left[\phi_{01} - \frac{qMd_I}{2\epsilon_s} \right] + \gamma_2 \sqrt{\left[\phi_{01} - \frac{qMd_I}{2\epsilon_s} \right] + V_{CB}} \quad (9.3.20)$$

In fact, consider a *fictitious* unimplanted device with a threshold voltage given by

$$V_{T2} = V_{FB2} + \phi_{02} + \gamma_2 \sqrt{\phi_{02} + V_{CB}} \quad (9.3.21)$$

Then it is easy to see that such a device will have a threshold that behaves identically to (9.3.20) if we choose

$$\phi_{02} \equiv \phi_{01} - \frac{qMd_I}{2\epsilon_s} \quad (9.3.22)$$

**FIGURE 9.6**

Threshold voltage vs. source-substrate bias for an *n*MOS transistor with a high-low implant profile (Fig. 9.3a): (a) for a step-approximated implant; (b) for three limit cases discussed in the text.

and

$$V_{FB2} \equiv V_{FB} + qM \left(\frac{1}{C'_\text{ox}} + \frac{d_I}{2\epsilon_s} \right) \quad (9.3.23)$$

It is emphasized here that V_{FB2} is the flatband voltage of the *fictitious* unimplanted device and should not be thought of as a flatband voltage of the real device (similarly for the quantity ϕ_{02}). Nevertheless, we can take advantage of the convenient, familiar functional form of (9.3.21) and use it to describe the actual implanted device under consideration for $V_{CB} \geq V_I$, viewing ϕ_{02} and V_{FB2} as mere symbols defined by (9.3.22) and (9.3.23).[†] Note that $V_{FB2} > V_{FB1}$, $\phi_{02} < \phi_{01}$, and $\gamma_2 < \gamma_1$.

In a transistor (Fig. 9.3a), the gate-source threshold voltage can be found by using the fact that, at the source, the role of terminal *C* in Fig. 9.3b is played by the source terminal *S*; thus, replacing V_{CB} by V_{SB} in (9.3.12b) and (9.3.21) gives

$$V_T = \begin{cases} V_{FB1} + \phi_{01} + \gamma_1 \sqrt{\phi_{01} + V_{SB}}, & V_{SB} < V_I \\ V_{FB2} + \phi_{02} + \gamma_2 \sqrt{\phi_{02} + V_{SB}}, & V_{SB} \geq V_I \end{cases} \quad (9.3.24)$$

Since two regions are distinguished, and a number of approximations were made in deriving V_T for each region, it is important to check for possible anomalies at the critical point $V_{SB} = V_I$. A simple calculation shows that both $V_T(V_{SB})$ and dV_T/dV_{SB} are

[†] One should not, for that matter, be surprised if for some combination of process parameters ϕ_{02} is found to have a negative value. This simply means that the implant is such that $qMd_I/(2\epsilon_s) > \phi_{01}$, from (9.3.22). It is easy to check that the quantity under the square root in (9.3.21) never becomes negative in the domain of definition of that expression ($V_{CB} \geq V_I$).

continuous at this point. Thus, no troublesome “kinks” will be caused in the transistor I - V characteristics by using this model for V_T .

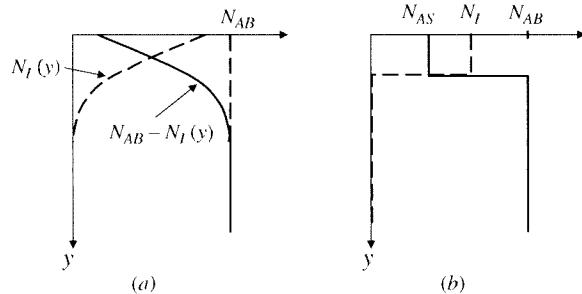
The quantity $V_T(V_{SB})$ is plotted in Fig. 9.6a. Two regions are clearly distinguishable.¹⁶ This behavior is verified by experiment. To obtain some intuition about this behavior,¹⁶ consider an unimplanted device with uniform doping N_{AS} throughout the substrate. Then $V_T = V_{T1}$ for all V_{SB} values, resulting in curve *a* in Fig. 9.6b, characterized by γ_1 . Next, consider a device with uniform substrate concentration N_{AB} (no implant). Then $V_T(V_{SB})$ will look as shown by curve *b* in Fig. 9.6b and will be characterized by a body effect coefficient γ_2 as given by (9.3.16). Finally, consider the substrate concentration unchanged from the value N_{AB} and assume that the device has been subjected to an extremely shallow implant such that all ions end up practically at the oxide-semiconductor interface; the dose is assumed to be $M = N_i d_i$, the same as for Fig. 9.4b. Now, the total charge per unit area of these ions, $(-q)M$, simply serves to modify the value of the effective interface charge Q'_o , as already discussed. From (9.3.4) it follows that this will only result in a positive shift in V_{FB} by an amount qM/C'_{ox} . Hence, for this case, the plot of V_T vs. V_{SB} will be as shown by curve *c* in Fig. 9.6b. This curve results from a vertical shift of curve *b* by an amount equal to the change in V_{FB} . The substrate doping is still N_{AB} and, thus, the body effect coefficient is still γ_2 . The plot of V_T in Fig. 9.6a follows curve *a* for $V_{CB} < V_I$ and is close to curve *c* if V_{SB} is large. If d_i is reduced, the “transition point” in the plot will move to the left. For extremely shallow implants, the curve in Fig. 9.6a will become practically the same as curve *c* in Fig. 9.6b. The device can then be characterized as discussed in previous chapters, with only a shift in V_{FB} . Comparing the curve in Fig. 9.6a to curve *b* in Fig. 9.6b (unimplanted device), it is seen that a *p*-type implant on a *p*-type substrate *increases* the extrapolated threshold voltage (for a given V_{SB}).[†]

The simple model we have presented is adequate for demonstrating the important threshold effects associated with the devices we are considering. (Other models are discussed elsewhere.²⁰) Since a number of approximations were made in the course of deriving the model, one should allow for the adjustment of parameter values, notably M and d_i , to attain satisfactory agreement with measured results. A related discussion was presented in Sec. 4.17.[‡] “Doping transformations” have been proposed for choosing the effective values of N_{AS} (or N_i) and d_i for good accuracy.^{27,34,42,51,54,57} In some of these techniques, these effective values are made functions of V_{SB} .

Low-High Profile We now consider the case in which the implant is of the opposite type from the substrate; i.e., it consists of donor ions for the case of a *p* substrate. If the implant concentration is again denoted by $N_i(y)$, because the donor

[†] This threshold-adjust implant is often in addition to implants used to limit short-channel effects mentioned before. Since the latter would extend below the former, to depths where punchthrough would normally occur, the doping below the threshold-control implant considered in this chapter may not be N_{AB} as assumed, and, in fact, the picture of the depletion region there may be quite complicated. As a first-order correction, then, one may have to modify the value of N_{AB} appearing in our expressions. If the effective N_{AB} is allowed to be adjusted by a “parameter extraction” system (Chap. 10) for best matching to measurements, this modification would be done automatically.

[‡] An exact analysis of threshold voltages due to an arbitrary profile of any shape or depth in terms of dose and centroid parameters is given elsewhere.²⁵

**FIGURE 9.7**

Low-high profile: (a) substrate doping concentration before implant, ΔL , donor implant concentration ΔI , and total doping concentration, γ , vs. depth from surface; (b) step approximation for (a).

charge subtracts from the substrate acceptor charge, the effective substrate doping will now be $N_{AB} - N_i(y)$. This case is shown in Fig. 9.7a; in drawing this figure, we have assumed that the peak of the implant distribution is above the oxide-semiconductor interface. We assume that the implant peak density is not too high, so that the combined concentration is of the same type as that of the original substrate,[†] i.e., $N_{AB} - N_i(y)$, is positive for all y . The resulting rectangular approximate profile will now be as in Fig. 9.7b, where we have

$$N_{AS} = N_{AB} - N_I \quad (9.3.25)$$

One can go through the same development as above, and the only change in the results will be a $-$ sign in front of N_i [and, due to (9.3.17), in front of M]. We again obtain (9.3.24), with the parameters in it given by the preceding equations with the sign changes just mentioned. This gives $V_{FB2} < V_{FB1}$, $\phi_{02} > \phi_{01}$, and $\gamma_2 > \gamma_1$. If V_I is positive, a two-part plot again results for V_T vs. V_{SB} , with the slope of the second part being higher than what it would have been if the substrate were uniform with concentration N_{AS} . The overall curve, and its slope, are again continuous.

Real Profiles If, instead of the step profiles used previously, one uses an actual profile, such as that shown in Fig. 9.4a, the general trends revealed earlier are still observed, although no separate regions are observed in the V_T vs. V_{SB} plot, as the depletion region edge moves over the gradual tail of the doping concentration. However, it is no longer possible to derive simple analytical results. An engineering approach is instead to add a correction to the simple body effect equation for uniform substrates, (9.3.6a), to account for the nonuniformity. Because the nonuniformity is gradual, a very simple, linear correction becomes possible, as follows:^{98–101}

$$V_T = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0 + V_{SB}} + K_2 V_{SB} \quad (9.3.26)$$

[†] The case where this does not hold is discussed in Sec. 9.3.4.

where K_2 is introduced as a parameter that is determined from fitting experimental data. To see how the last term in this equation helps, we will use the high-low profile case as an example. Consider first the main part of (9.3.26), excluding the last term. The plot of that part is as in curve *a* in Fig. 9.6*b*. Assume that K_2 has a slightly negative value. The negative contribution this imparts on the slope of V_I vs. V_{SB} can significantly affect the small positive slope of the rest of the equation when V_{SB} is large and tends to flatten the curve out for such V_{SB} values. However, at low values of V_{SB} , where the slope of the rest of the equation is large, the influence of the last term on the overall slope is small percentagewise. This then tends to give the overall equation (9.3.26) the trend seen in Fig. 9.6*a*, albeit without distinguishing two regions. For low-high profiles, a small positive value would instead be used for K_2 , the effect of which will be mostly apparent at large V_{SB} values, where the slope of the rest of the equation is small.

Equation (9.3.26) has proven to be remarkably effective at fitting data from real devices. Note that in general other parameters, such as the bulk doping level used in the expression for γ in (9.3.5), are considered to be “effective” values also to be determined by parameter extraction, to improve the fit of the previous equation to measured results. Thus, such parameters lose some of their physical meaning. This is not a problem in practice, because real transistors differ from the idealized structures that we are able to model analytically, so in fact most “parameters” are, in practice, fitting parameters, although clearly N_A has a more physical meaning than K_2 . More on parameter extraction is discussed in Chap. 10.

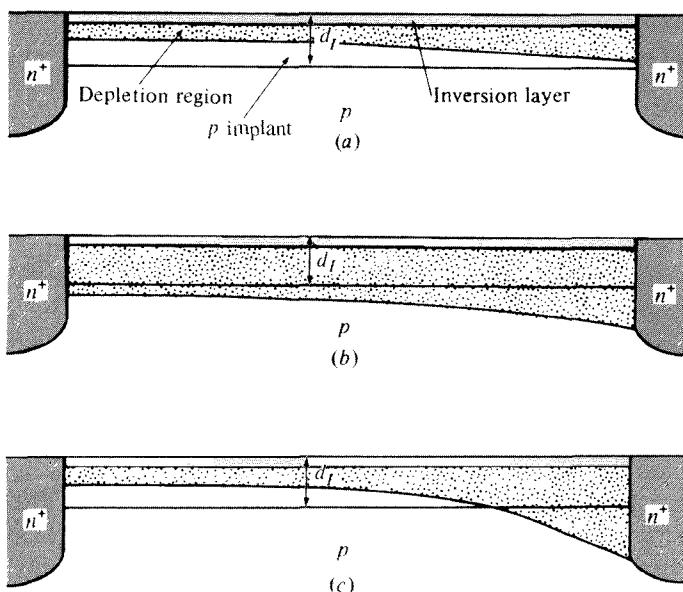


FIGURE 9.8
Depletion region (shaded area) in an *n*MOS transistor with a high-low implant profile. The depletion region around the n^+ region is not shown for simplicity. (a) $V_{SB} < V_{DB} < V_I$; (b) $V_I < V_{SB} < V_{DB}$; (c) $V_{SB} < V_I < V_{DB}$. Strong-inversion operation is assumed.

9.3.3 Drain Current

Let us now consider the complete transistor in Fig. 9.3a. Assume initially a step profile as before for simplicity. From the previous discussion it is evident that the depletion region under the inversion layer may be shallower or deeper than the implant, depending on bias voltages. As an example, Fig. 9.8 shows three possible cases in strong inversion. Explicit strong-inversion models can be developed for the drain current taking such cases into account, starting from (4.7.6) and using the right expression for Q'_B for each V_{CB} , as already determined.^{20,102} The following comments focus on general models valid in all regions of inversion.

As in Chap. 4, the inversion layer charge can be used to find the drain current, as in (4.3.7) and (4.3.11). Note that those equations were developed without assuming uniformity for the substrate and therefore are valid here. However, strictly speaking, one can no longer use the development following those equations, as that development is based on (4.3.12), which is based on the assumption of a uniform substrate. Unfortunately, using the actual depletion charge for a nonuniform substrate in (4.3.11) does not lead to tractable analytical results. Thus, invariably certain approximations are made. There are several reasons that approximations can lead to acceptable results in practice:

- The variation of doping with depth is gradual, unlike the case of the step profile in the previous section.
- The details of the variation of the depletion charge Q'_B tend to be obscured by the fact that only the integral of that charge appears in the drift equation, (4.3.11a).
- The main terms of the current equations in (4.3.11) (e.g., the linear terms and the square-law terms) appear in this equation independent of the form of Q'_B ; this means that the most important parts of these equations are the same as those for uniform substrates [see (4.3.14)].

For all these reasons, it is found that, for nonuniform substrates, one can continue using the form of the general models of Chap. 4 and only make certain adjustments to improve accuracy. A main form of adjustment is already obvious from our discussion in the previous section: one can adjust the threshold voltage, and use it in strong inversion, where drain current equations contain it explicitly. For general models, which do not contain the threshold voltage explicitly, one may adjust the values of V_{FB} and γ or adjust the effective doping concentration, which affects both. From our discussion in the previous section, we know that the doping level at the edge of the depletion region has a decisive effect on device characteristics, and that the location of this edge depends on the surface potential. Because of this, one of the adopted solutions is to introduce an effective doping in the uniform substrate current equations that depends on the surface potential.^{27,42,54} This approach does *not* mean that the bulk doping physically changes with the surface potential, but that for the purpose of modeling, one can conveniently adopt this view. For models based on threshold voltage (or pinchoff voltage), the introduction of this approach, which indirectly makes γ a function of V_{SB} for source-referenced models, has proven effective in modeling experimental data. The effective doping variation introduced must be continuous.

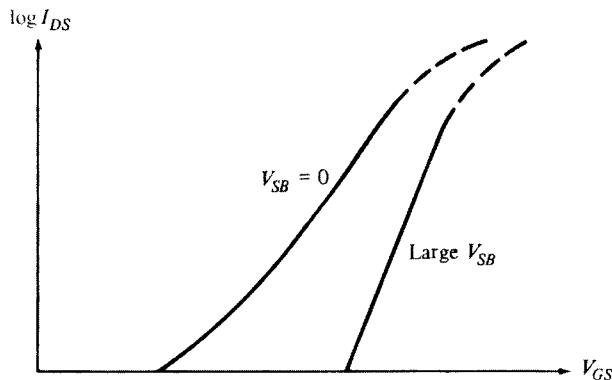


FIGURE 9.9
Weak inversion $\log I_{DS}$ vs. V_{GS} in saturation, with V_{SB} as a parameter, for an nMOS transistor with a high-low implant profile.

A problem with the ψ_s -dependent bulk doping approach for surface potential-based models is that it significantly alters the procedure that must be used to solve for ψ_s . Thus, (4.3.15) can no longer be solved for a given V_{GS} as before, as γ is now a function of ψ_s through the ψ_s -dependent doping in (9.3.5). Another approach used is to make the effective substrate doping a function of V_{GB} .¹⁰³ This works because γ has a fixed value for each V_{GB} , and thus (4.3.15) can be solved for a given V_{GB} . This allows standard solution techniques for ψ_s to be retained and leads to accurate modeling of the gate capacitance of low-high-doped MOSFETs over gate bias, from accumulation to strong inversion.¹⁰³

Qualitative Deviations from Uniform-Substrate Behavior The general shape of I - V characteristics of devices with moderately nonuniform substrates is often similar to that of ideal, uniform-substrate devices. However, in the presence of strong nonuniformity, differences can occur, as we have discussed. For example, in strong inversion, the case in Fig. 9.8c can result in some deviation from the square-law behavior expected from Sec. 4.7.3.¹⁰² Deviations may also be encountered in weak inversion. Depending on implant details, in some cases plots of $\log I_{DS}$ vs. V_{GS} in saturation may look as shown in Fig. 9.9, which assumes a high-low profile. For large V_{SB} , the gate voltage needed to bring the surface to weak inversion is large. This causes a deep depletion region, with its edge outside the implant. The device then behaves qualitatively as an unimplanted device. The slope of the curve is proportional to $1/n$ as before, with n given by (4.8.14), where $\gamma = \gamma_2$ as given by (9.3.16). However, for low V_{SB} , the required gate voltage for weak inversion is small; the corresponding depletion region is shallow, and its edge is inside the implant. Then $\gamma = \gamma_1$ as given by (9.3.10), and thus n and S (Sec. 4.8) are large and the slope is small. For low-voltage digital applications this can be undesirable, because it now takes a larger “swing” of V_{GS} to change the current, and it may be more difficult to turn the device off. Note, also, that as V_{GS} increases, the depletion region edge can be moving over a region with a widely varying concentration (from the “center” of the implant toward its “sides”). Then no simple value for n can be defined, and the slope of $\log I_{DS}$ with V_{GS} changes as shown in the figure; I_{DS} is no

longer exponential with V_{GS} . This can create problems in some analog circuits, which rely on the near-exponential behavior normally expected of weak inversion. For low-high profiles, the curvature of the plot would be the opposite of the one shown. Such problems can be solved by proper device engineering.

pMOS Devices In pMOS devices, an *n*-type substrate is used. Ion implantation is used in pMOS devices for the same reasons already discussed. For example, if it is desired to make a pMOS device more difficult to turn on, i.e., make the threshold voltage more *negative*, an *n*-type implant is used, resulting in a high-low profile. The resulting substrate nonuniformity can be analyzed as before, with some obvious sign changes, along the lines of Sec. 4.16.

9.3.4 Buried-Channel Devices[†]

In our discussion of low-high profiles we have assumed that the donor implant peak concentration is lower than the unimplanted substrate concentration, so that the implanted substrate remains *p* type throughout (Fig. 9.7). If, instead, the peak donor concentration is higher than that of the unimplanted substrate, we have the situation shown in Fig. 9.10a. We now end up with a region next to the surface that is *n* type. A rectangular profile will be used for simplicity as before; this is shown in Fig. 9.10b. With $N_I > N_{AB}$, the region between the surface and $y = d$ is viewed as a uniformly doped *n*-type region, with an effective *donor* concentration of

$$N_{DS} = N_I - N_{AB} \quad (9.3.27)$$

As may be expected from this development, this type of implant lowers the threshold even more than in the case of Fig. 9.7. In fact, this is the way *depletion* devices were made, when such devices were popular in early MOS fabrication

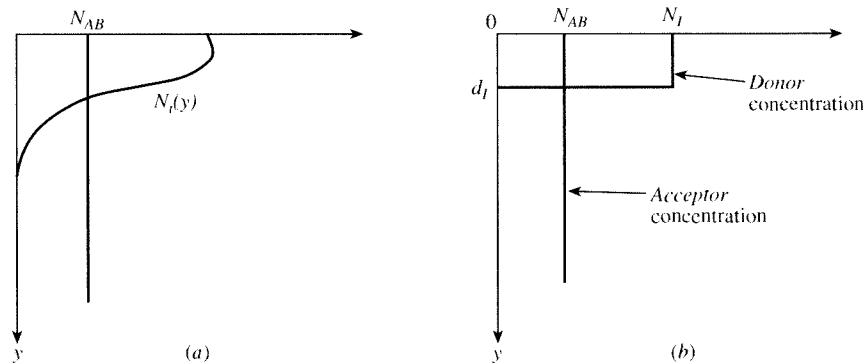
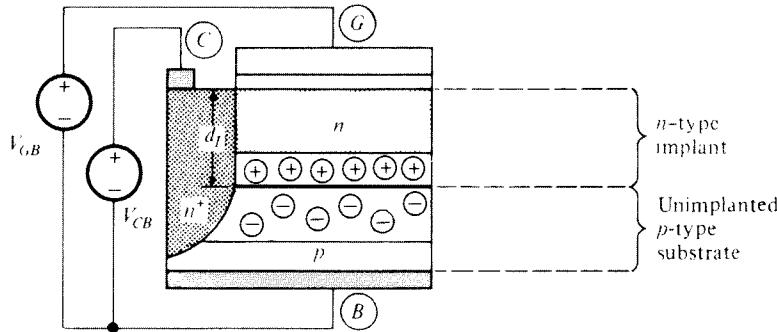


FIGURE 9.10

(a) Substrate acceptor doping concentration before implant (N_{AB}) and donor implant concentration $N_I(y)$ vs. depth from surface; (b) step approximation for the donor implant concentration of (a). Here it is assumed that $N_I > N_{AB}$.

[†] This section can be skipped without loss of continuity. Buried-channel operation is rarely used today, although it is being considered in device research.

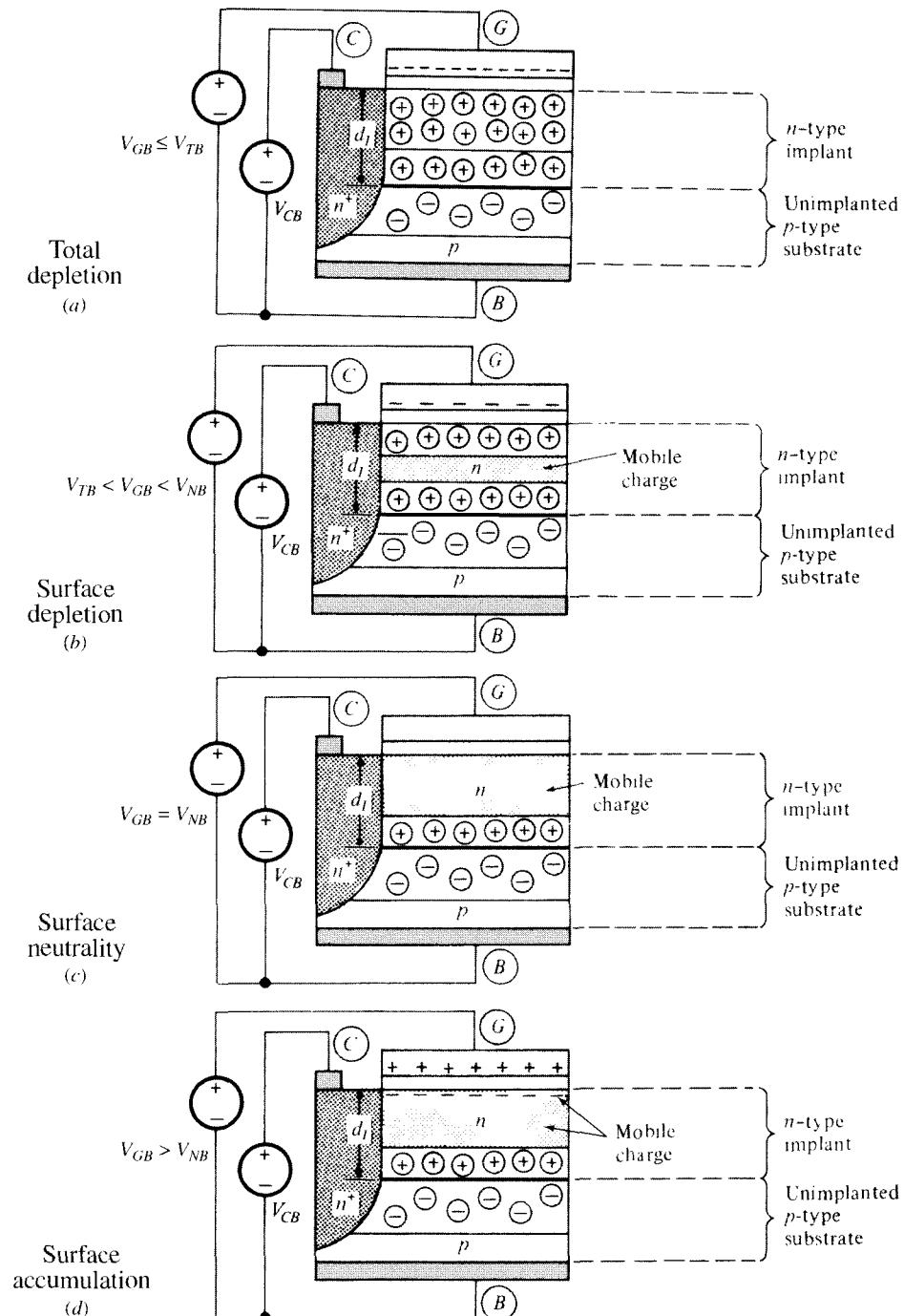
**FIGURE 9.11**

A three-terminal structure corresponding to a buried channel *n*MOS transistor. The depletion region within the n^+ region is not shown for simplicity.

processes. We remind the reader that such devices have a negative threshold and are “normally on”; see Fig. 4.40b. This can be confirmed by considering what happens if the implanted region in Fig. 9.3b has a net *n*-type doping; this region then forms a conductive bridge between the *n*-type source and the *n*-type drain. In this device, a mode of operation that we have not yet encountered can occur: if the gate is made sufficiently negative, some of the electrons near the surface are repelled, and the surface becomes depleted. The conducting channel is thus limited to a path in the *n* implant *below* the surface. Devices in which this is the main operating region are called *buried-channel* devices. By contrast, transistors with implants as in Fig. 9.4 or Fig. 9.7 are surface-channel devices. Although buried-channel devices are rarely used today, this mode of transistor operation is still being studied in research for future devices. We now briefly discuss the effects involved.^{20,21,45,49,51,59–96}

As usual, let us first consider a three-terminal structure, as shown in Fig. 9.11. For simplicity in drawing this figure, we have assumed that the interface charge Q'_o is zero. The *p* substrate and the *n* implant form a *pn* junction. The depletion region associated with this junction is shown in the figure. However, note that the depletion region around the n^+ region boundary is not shown in order to keep the figure simple. The *n* implant can be contacted through the n^+ region, which is contacted through terminal *C*. Thus, an external nonnegative voltage V_{CB} acts as a reverse bias for the implant-substrate junction. The value of V_{CB} determines the width of the depletion region, and the magnitude of the charge per unit area on each side of the junction.

The nondepleted part of the *n* implant contains mobile electrons. In the complete transistor, all these would be available for conduction, provided no other part of the implant were depleted. However, as already mentioned, if V_{GB} is made sufficiently negative, the resulting negative charges on the gate will repel some of the mobile electrons, and a depletion region (containing ionized donor atoms) will form at the top of the implant. If V_{GB} is made sufficiently negative, then all of the implant will be depleted (will be pinched off), and there will be no mobile electrons left in

**FIGURE 9.12**

Condition below the oxide of a buried-channel device for various gate-substrate voltage values. The depletion region around the n^+ region boundary is not shown, and Q'_o is assumed zero for simplicity.
 (a) Total depletion; (b) surface depletion; (c) surface neutrality; (d) surface accumulation.

it. The value of V_{GB} at which this is attained is defined as the gate-body threshold voltage[†] V_{TB} , and the resulting situation for $V_{GB} \leq V_{TB}$ is shown in Fig. 9.12a.

According to this simple picture, if the n implant were part of a complete transistor, at $V_{GB} = V_{TB}$ there would be no mobile electrons available for conduction. This simple point of view is analogous to considering only strong inversion in unimplanted devices and saying that the current reduces to zero when the gate voltage is equal to the threshold voltage. While this view is not precise, it is often used to describe first-order behavior. It is to be understood that, if very low current operation is of interest, this view should be refined.[‡]

Let us now consider what happens at other V_{GB} values. As V_{GB} is raised above V_{TB} , the charge on the gate becomes less negative, and only part of the implant needs to be depleted in order to balance it. This is shown in Fig. 9.12b. If the structure were part of a transistor, the mobile electrons in the undepleted part of the implant would contribute to conduction. This undepleted part of the implant would thus form the channel of the transistor. As already mentioned, since this channel is below the surface, it is referred to as a buried channel, and devices operating in this mode are called buried-channel devices.[§]

If V_{GB} is raised further, a point will be reached at which no gate charges will exist (or, more properly, if the interface charge Q'_o is not assumed zero as in the figure, the gate charges will just balance Q'_o). Then the top of the n region will not be depleted, as shown in Fig. 9.12c. The value of V_{GB} at which this situation occurs is denoted by V_{NB} in the figure. Note that the undepleted part of the implant is neutral, as each mobile electron in it is balanced by an ionized donor atom.

If V_{GB} is raised still further, positive charges will be placed on the gate (or if Q'_o is not assumed zero, $Q'_G + Q'_o$ will be positive). These charges must now be balanced by negative charges. Electrons will thus accumulate at the top of the implant, as shown in Fig. 9.12d. If the structure were part of a transistor, in this case the electrons in the accumulation layer would contribute to conduction, in addition to the mobile electrons in the neutral part of the implant.

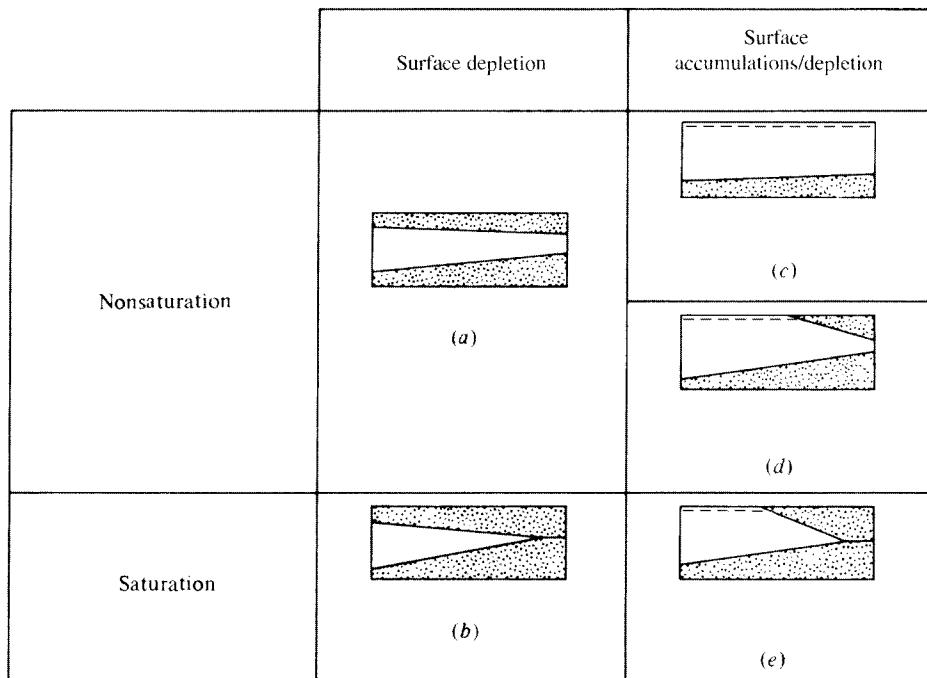
A quantitative analysis for the cases discussed previously can be found in Appendix I, where the threshold voltage is calculated. It is found there that buried channel devices exhibit a body effect, because increased V_{CB} for a given V_{GC} widens the depletion region between the n and p regions and reduces the cross section through which the current flows. The body effect coefficient turns out to be larger than what would correspond to an unimplanted device on the p -substrate used. Physically, this makes sense as, for a buried-channel device, the channel is further away from the gate, in comparison to a surface-channel device; thus, the gate releases some of its control to the body.

Using the preceding results for the complete transistor (Fig. 9.3a, with an n -type channel assumed), one can see that there are several cases that must be

[†] In some of the literature, this quantity is also referred to as the "pinchoff voltage" in this context; this term should not be confused with the pinchoff voltage V_p in this book.

[‡] Among other things, the assumption that depletion regions edges are "sharp" would have to be reconsidered.

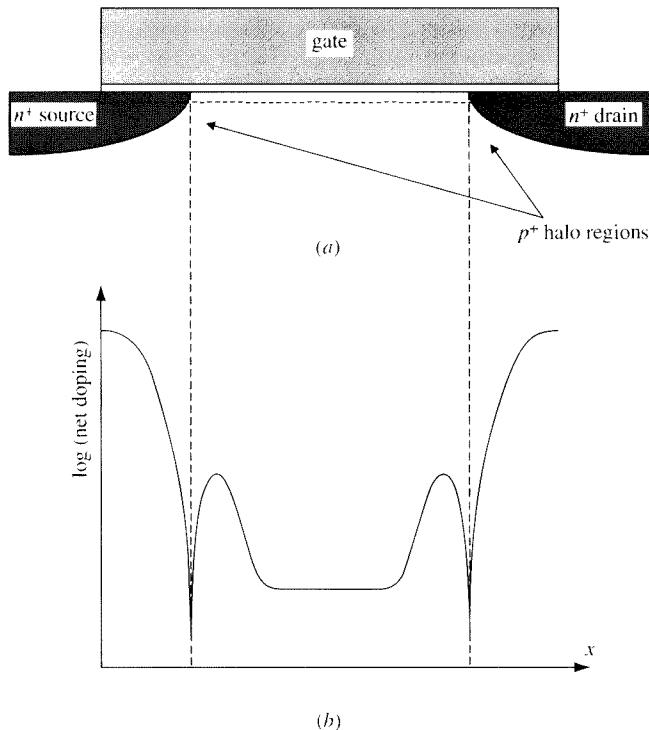
[§] If a low-dose implant is used, the upper part of the implant-substrate junction depletion region may reach all the way to the surface, even with $V_{CB} = 0$. In that case, increasing V_{GB} would eventually attract electrons to the surface, and one would have a surface-channel device. We are not considering this case here.

**FIGURE 9.13**

Modes of operation for an *n*MOS buried-channel transistor; shown is only the implant part. Depletion regions are shown shaded.

distinguished, as shown in Fig. 9.13.²¹ In drawing this figure, it is assumed that $V_{DB} > V_{SB}$; thus, the channel-to-substrate potential, V_{CB} , increases from source to drain. This has the effect of increasing the reverse bias across the *np* junction, thus widening the bottom depletion region toward the right, as shown. It also has the effect of decreasing $V_{GC} = V_{GB} - V_{CB}$ (for a given V_{GB}), thus reducing the field through the oxide and decreasing the mobile carriers available, from source to drain. With high enough drain voltages, it is possible for the upper and lower depletion regions to meet near the drain, and we then have channel pinchoff, resulting in drain current saturation.

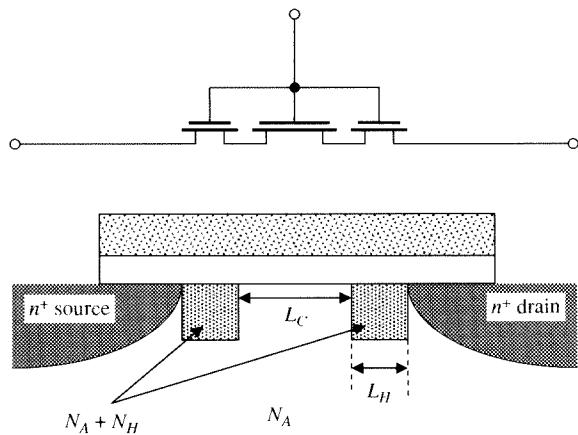
To determine the drain current, one cannot use the equations in Chap. 4, as they are based on the presence of an inversion layer charge, Q_f . Here, instead, conduction occurs because of *two* different types of charges: the charge due to the mobile electrons in the neutral part of the *n* region, and the charge due to the accumulation of electrons next to the surface. One or both of these charges can be zero in a given situation, as shown in Fig. 9.13. The two charges contribute parallel paths for the current; thus the total charge responsible for conduction is the sum of these two charges. An explicit strong-inversion model based on this approach can be developed,^{20,21,102} as discussed in Appendix J. Models valid in all regions of operation, analogous to the general models of Secs. 4.3 and 4.4, have been proposed.^{86,95}

**FIGURE 9.14**

(a) A transistor with halo implants. (b) Doping profile for (a); shown is the magnitude of the net doping concentration on a log scale, along the broken horizontal line in (a). The doping changes from n to p at the boundaries between the n^+ regions and the halo regions.

9.4 Substrate Lateral Nonuniformity

It was seen in Chap. 5 that a reason for the presence of undesirable short-channel effects is the fact that the depletion regions around the source and drain spread into the body region under the channel. To reduce the extent of these depletion regions, one can use higher substrate doping. However, this results in other undesirable characteristics, such as stronger body effect and larger junction capacitances. Thus, one ends up increasing the body doping selectively, only around the source and drain, using the so-called halo or pocket implants, as shown in Fig. 9.14a. The magnitude of the doping density $|N_A - N_D|$ along the horizontal broken line at a small distance from the Si-SiO₂ interface in this figure is shown in Fig. 9.14b. Note that at the source and drain edges the doping concentration changes from n to p , and thus has to go through zero. The higher doping partially “shields” the depletion region in the center of the transistor from being influenced by the drain voltage, and therefore reduces the amount of drain-induced barrier lowering, or DIBL (Sec. 5.5). The halo implants therefore give reduced currents in weak and moderate inversion, reducing the residual current when the transistor is turned off in digital operation. As we have discussed in Chap. 5, this is very important because the nonswitching, or “static,” power dissipation in large digital circuits such as

**FIGURE 9.15**

Idealized sectional representation of a halo-implanted transistor. N_H is the additional doping level in the halo regions.

microprocessors, has become an important problem in modern technologies. The introduction of the halo implants has an effect on the shape of the transistor characteristics, as will now be described with the help of a simplified model.

Figure 9.15 shows an idealized schematic representation of a cross section of the device, with the halo-implanted regions depicted as being distinct regions of additional uniform doping concentration N_H . As can be seen, the device can be approximated as a series connection of three separate MOS transistors, one for each of the source and drain halo regions, with doping density $N_H + N_A$ and length L_H , and one for the central region, with doping density N_A and length L_C . Note that the lateral extent of the halo regions is determined by the implant step that forms them, and it is fixed. Therefore if the channel length of the transistor is decreased, the length L_C of the central transistor in Fig. 9.15 decreases, but the length of the halo regions at the source and drain ends remains fixed. We have

$$L_C = L - 2L_H \quad (9.4.1)$$

Consider the operation of a halo-implanted transistor as the gate voltage is increased from weak inversion to strong inversion, for a small drain voltage so the device operates deep in nonsaturation. Both the Fermi potential ϕ_F and the body effect coefficient γ are greater for the halo regions than for the central region of the transistor. This means that the threshold voltage (4.7.19) for the outer transistors in the model of Fig. 9.15 is greater than for the central transistor. Consequently, for a given gate potential the inversion charge density will be greater in the central portion of the transistor than in the halo regions. Specifically, there will be some gate bias where the central region will be starting to invert strongly, yet the halo regions will still be in moderate inversion. As the gate bias is increased further, the halo

regions will start to enter strong inversion. However, at this point, the central region is fairly heavily inverted; thinking of the channel as three resistive regions connected in series, we can see that the resistance of the central region will be substantially smaller than that of the halo regions, because the inversion charge density there is higher (note that we can think of the channel as a resistor, thanks to our assumption of operation in deep nonsaturation). The complete transistor therefore behaves, for dc

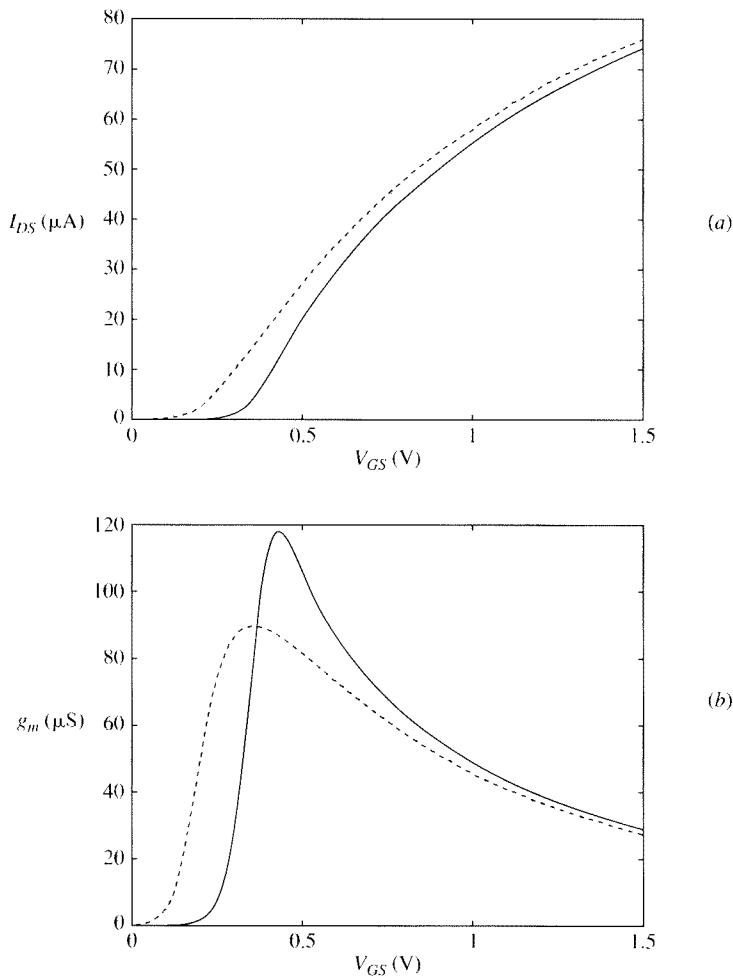


FIGURE 9.16

(a) Drain-source current and (b) transconductance vs. gate-source voltage for a halo-implanted device (solid line), and for a uniform-substrate device with doping identical to that in the center of the halo device. $t_{ox} = 2 \text{ nm}$, $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $N_H = 1.5 \times 10^{18} \text{ cm}^{-3}$, $W = 1 \mu\text{m}$, $L = 0.2 \mu\text{m}$, $V_{DS} = 0.05 \text{ V}$, and $V_{SB} = 0 \text{ V}$. The length of the halo regions is 20 nm.

current, as if its effective channel length were just that of the halo regions, i.e., $2L_H$. As the gate bias continues to increase, the level of inversion becomes strong everywhere along the channel, and so the composite transistor begins to behave more as a single, strongly inverted transistor of length L . This simple observation turns out to be very useful in interpreting the characteristics of halo-implanted devices qualitatively.

Figure 9.16 shows the drain current I_{DS} and the transconductance $g_m = \partial I_{DS} / \partial V_{GS}$ for a uniformly doped transistor (with $N_A = 5 \times 10^{17} \text{ cm}^{-3}$) and for a halo-doped transistor (with a peak doping in the halo region of $1.5 \times 10^{18} \text{ cm}^{-3}$), the latter simulated using the three-section model of Fig. 9.15. As seen, the uniformly doped transistor turns on first because its threshold is lower, as discussed earlier. The slope of the curve in Fig. 9.16a for this device, i.e., the small-signal transconductance, is proportional to W/L . The halo-doped transistor turns on “later”: its current is prevented from reaching a significant value until the gate voltage has exceeded the threshold of the halo regions, which is large due to the large doping there. For gate voltages somewhat higher than this, the device behaves as if it had an effective length of $2L_H$, as already explained. Thus the slope of the curve is proportional to $W/(2L_H)$ and is larger than that of the uniform-substrate device. For higher gate voltages, the entire halo-implanted device begins behaving as a device with channel length L , the conductance becomes proportional to W/L , and the curve becomes approximately parallel to that of the uniform-substrate device. The high-slope region leads to a peaked transconductance, as seen in Fig. 9.16b.

Figure 9.17 shows the gate capacitance to the rest of the structure, for $V_{DS} = V_{BS} = 0$. The effect of the nonuniform lateral doping on gate capacitance is noticeable: there are two distinct “turn-on” regions, corresponding to where the more lightly doped central region becomes moderately/strongly inverted and, at a higher gate bias, where the more heavily doped halo regions become moderately/strongly

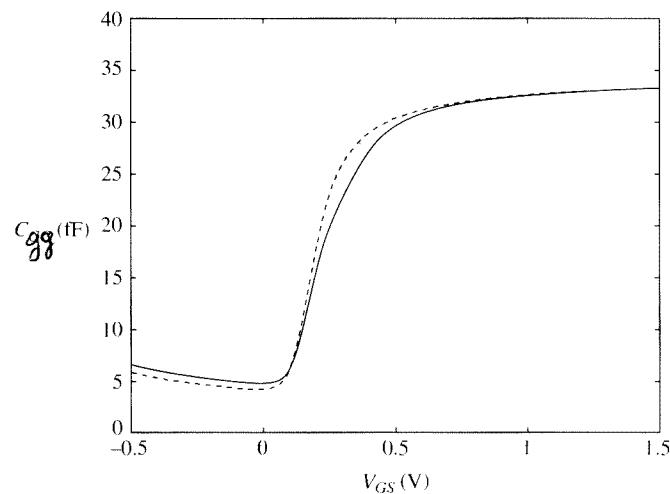


FIGURE 9.17

Total gate capacitance for a halo-implanted device (solid line) and for a device with uniform-substrate doping equal to that at the center of the halo device. Parameter values are the same as for Fig. 9.16.

inverted. Note that for the gate capacitance, the three sections appear in parallel, rather than in series as for the drain current.¹⁰⁴

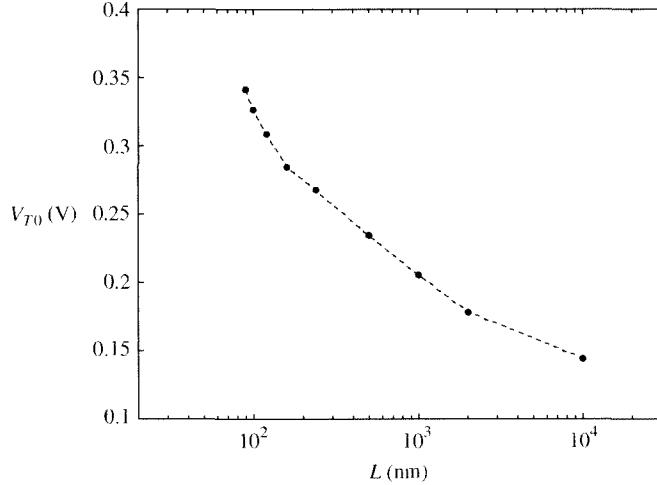
Although simulation of the previous effects is possible with a three-section model,^{104,105} in practice the transition from the halo-region doping to the main substrate doping is not sudden, as assumed earlier, but gradual. This makes the effects less pronounced. Also, for simulation of large circuits, this approach is computationally time-consuming (circuit simulation time is strongly tied to model computation time, and using a three-section model approximately triples circuit simulation time). Therefore various approximations have been developed using modifications of single-section models to account for the main effects of nonuniform lateral doping. A main thread in such approaches is to concentrate on how the total channel length influences the device characteristics. Some average substrate doping, appropriate for a given device length, is used, and the device is otherwise assumed to behave as one on a uniform substrate. While any single fixed substrate doping value cannot capture the effects illustrated in Figs. 9.16 and 9.17 above, this approach at least can provide reasonable overall agreement between the model and measurements, for various channel lengths. To see how this approach works, consider Fig. 9.15 in the case where the channel is extremely long; then the halo regions can practically be neglected and, once turned on, the device operates almost as if it were on a uniform substrate with doping concentration N_A .[†] Think now of the channel length decreasing, and the halo regions approaching each other. At $L = 2L_H$, the two regions meet, and the doping concentration becomes uniform again, at the value $N_H + N_A$ (always assuming the simplified profile in Fig. 9.15). For intermediate lengths, a uniform substrate concentration between the preceding two values, N_A and $N_H + N_A$, is used. To determine a reasonable value for this concentration,¹⁰¹ note that for the assumed step lateral doping profile of Fig. 9.15, the average doping in the channel region is

$$\begin{aligned} N_{\text{eff}}(L) &= \frac{N_A L_C + 2(N_H + N_A)L_H}{L_C + 2L_H} \\ &= N_A \left(1 + 2 \frac{N_H}{N_A} \frac{L_H}{L} \right), \quad L \geq 2L_H \end{aligned} \quad (9.4.2)$$

As L decreases, $N_{\text{eff}}(L)$ increases from N_A to $N_H + N_A$, as expected from the preceding discussion. It is now assumed that a device of channel length L behaves approximately as one on a uniform substrate with doping concentration $N_{\text{eff}}(L)$. This approach leads to corresponding expressions for parameters in which the doping concentration plays a role. For example, the body effect coefficient in (3.4.3) is replaced by an effective body effect coefficient:

$$\gamma_{\text{eff}}(L) = \frac{\sqrt{2q\epsilon_s} N_{\text{eff}}(L)}{C'_{\text{ox}}} \quad (9.4.3)$$

[†] Of course, the turn-on behavior of the device is still determined by the threshold of the halo regions, so errors can be expected in weak inversion if a uniform doping is assumed; however, this case of extremely long channels is of theoretical interest only, as in practice the channel length is not much larger than the halo region length.

**FIGURE 9.18**

Measured V_{T0} vs. L , using a log horizontal axis, for a halo-implanted 90-nm technology. $W = 5.0 \mu\text{m}$.

which, using (9.4.2), can be cast in the form

$$\gamma_{\text{eff}} = \gamma \sqrt{1 + L_{PE}/L} \quad (9.4.4)$$

where γ is the body effect coefficient for a uniform substrate doping concentration N_A , and $L_{PE} = 2(N_H/N_A)L_H$ is considered a fitting parameter. This is the basis of the “re-parameterization” used in the BSIM4 model.¹⁰¹ The above body effect coefficient can now be used in expressions such as that for the threshold voltage in (3.4.15):

$$V_{T0} = V_{FB} + \phi_0 + \gamma_{\text{eff}}(L)\sqrt{\phi_0} \quad (9.4.5)$$

Measurements of threshold voltage vs. channel length for a modern technology are shown in Fig. 9.18. From such data, one can determine the parameter L_{PE} (Prob. 9.4).

Another approach calculates an effective threshold as follows. In Fig. 9.15, assume $V_{SB} = 0$, and denote the corresponding threshold values for the center and halo regions by V_{T0C} and V_{T0H} , respectively. Define the threshold of the combination, V_{T0} , as the value needed for the combination to yield the same current as a uniform-substrate device, using common expressions such as the weak-inversion equation (4.8.11). Then for $L > 2L_H$, the zero-bias threshold voltage as a function of length can be approximated by¹⁰⁴ (Prob. 9.5)

$$V_{T0}(L) = V_{T0C} + n\phi_t \ln \left(1 + \frac{2L_H}{L} \left(e^{(V_{ionH} - V_{T0C})/(n\phi_t)} - 1 \right) \right) \quad (9.4.6)$$

The above approaches capture the increase in the effective threshold voltage with decreased channel length, already mentioned in Sec. 5.4.2. In modern technologies, this is the dominant geometry dependence of the effective threshold voltage. The increase in threshold with decreasing L can be expected intuitively from Fig. 9.14; as the channel is shortened, the implants tend to merge and the effective substrate doping increases; hence, the body effect and threshold voltage also increase, as dictated by (9.4.3) and (9.4.5). At even shorter channel lengths, the substrate is essentially uniform, and charge-sharing effects can in principle be observed (Sec. 5.4); then the threshold can decrease with further decreases of the channel length. However, sometimes the channel length where this begins to happen is below the minimum allowed channel length for a given fabrication technology.

For surface potential-based models, which do not have threshold voltage as a parameter, an equivalent doping can be calculated, to give results consistent with the threshold voltage (9.4.5) of the three-transistor equivalent.¹⁰⁴ It can be deduced from the analysis in Sec. 3.4.2 that the primary influence of doping concentration on the threshold voltage in (3.4.15) and the equation based on it earlier, (9.4.5), is through the body effect coefficient. This is because in the quantity $\phi_0 = 2\phi_F + \Delta\phi$ (Sec. 3.4), the Fermi potential ϕ_F varies only weakly (logarithmically) with doping [see (1.2.14)]; thus one can assume that ϕ_0 is independent of L and calculate the resulting body effect coefficient needed to make the threshold equal to (9.4.5). One can then calculate the effective substrate concentration using (9.4.3). Expressions for N_{eff} using this approach are derived in the references.^{104,106}

Improved compact modeling of laterally nonuniform bulk doping effects is an ongoing research topic; the compact MOSFET models used at present for IC design include some form of effective doping variation with length and can include different models for dc current and for charges. However, they do not capture the bias-dependent gate capacitance behavior seen in Fig. 9.17.

9.5 Well Proximity Effect

As already mentioned, well regions in a CMOS process are formed by ion implantation. During this process, ions are supposed to land only where the well is to be formed. To prevent them from landing elsewhere, parts of the surface are protected by a substance called “photoresist.” Ions that impinge on photoresist interact with

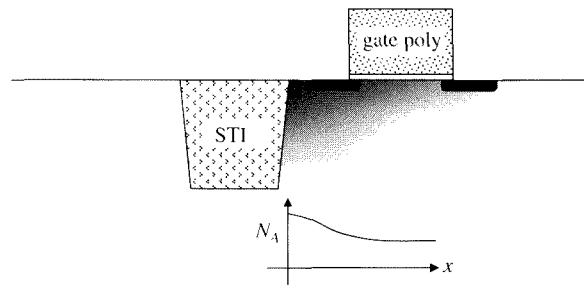
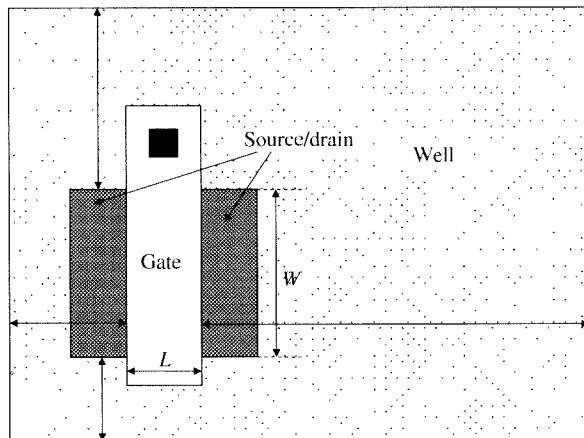


FIGURE 9.19

Well proximity effect. The well doping concentration varies with distance from the well edge.

**FIGURE 9.20**

Layout of a transistor with respect to well edges. Double arrows show the distances from the well edges, which determine the magnitude of the well proximity effect.

the photoresist atoms statistically and are “scattered” at random angles. Multiple scattering events drain the ions of their kinetic energy, so they stop within the photoresist (which is subsequently stripped from the silicon wafer surface). Ions that impinge near the edge of the photoresist, though, can scatter back out of the photoresist, and then become embedded in the silicon adjacent to the photoresist edge. Because of this phenomenon, the well doping one ends up with is not uniform, but rather is enhanced in regions adjacent to the edges of a well, to a distance of about $1 \mu\text{m}$ ^{107,108}. This is shown in Fig. 9.19.[†] This effect is known as the “well proximity effect” (WPE).

As we have discussed at length, the net bulk doping concentration is one of the key physical parameters than controls the operation of MOSFETs. Because the proximity of a transistor to a well edge changes the channel doping, it can affect transistor behavior significantly. For older technologies (approximately $0.25 \mu\text{m}$ and above) the overall size of transistors meant they could not be placed close enough to a well edge, and thus the enhanced doping level there did not significantly affect their behavior. For more modern technologies, WPE has become an important issue and can cause modeling inaccuracies and consequent circuit malfunction if not taken into account.

There are two aspects to understanding and modeling WPE: characterizing the distance from the relevant parts of a transistor to well edges and determining the effect of the resulting doping enhancement on transistor electrical behavior, manifested in its effect on model parameters. Figure 9.20 shows a MOSFET in a simple

[†] Ion implantation to a wafer is often done not perpendicularly to the surface, but rather at an angle, to prevent ions from avoiding scattering and penetrating too deeply by “channeling” along open paths in the silicon lattice. This may give the impression that the well doping enhancement is from ions that are “reflected” from the photoresist edge. This is incorrect; the scattering occurs within the photoresist and is observed even if implantation is perpendicular to the surface.

rectangular well. There are four different relevant spacings between the transistor and the well edges, as indicated by double arrows in the figure. To include the effect of the additional well dose from each adjacent edge on every model parameter is a daunting task, and it blurs the lines between layout extraction, modeling, and simulation, which are generally separate steps in an IC design flow. Therefore the approach taken for WPE modeling has been to separate the steps of extracting effective distance indicators from the device layout and of modeling the change in device behavior based on these indicators. This enables implementation of WPE across different types of models; for example, for a surface potential-based model the WPE can be handled by making the V_{FB} parameter a function of these indicators, and for a threshold voltage-based model, V_{T0} can be made a function of the indicators. Note that it is not only the distance from the well edge that is important in Fig. 9.20, but also the gate length of the transistor; this is because as the gate length increases, the portion of the transistor affected by the WPE decreases, and thus the effect of the transistor characteristics becomes less pronounced. It should also be noted that not all transistors have as simple layouts as that in Fig. 9.20, and in practice corner contributions also need to be included. Considerations that apply to complex layouts, including multifinger transistors, are provided in Ref. 109.

The second step in handling the WPE is to use the preceding information to make adjustments to a model. Although it would appear that the simplest way of doing this is to modify the parameter associated with N_A , this is not done in practice. Experimentally, the threshold voltage, body effect coefficient, and mobility are observed to vary with well proximity. Rather than tie the first two together through N_A , they are kept separate and have separate WPE variation models. This gives greater flexibility in fitting experimental data.¹¹⁰ For all-region models, instead of a threshold voltage shift, a shift in V_{FB} is used; this, from (3.4.15), is equivalent.

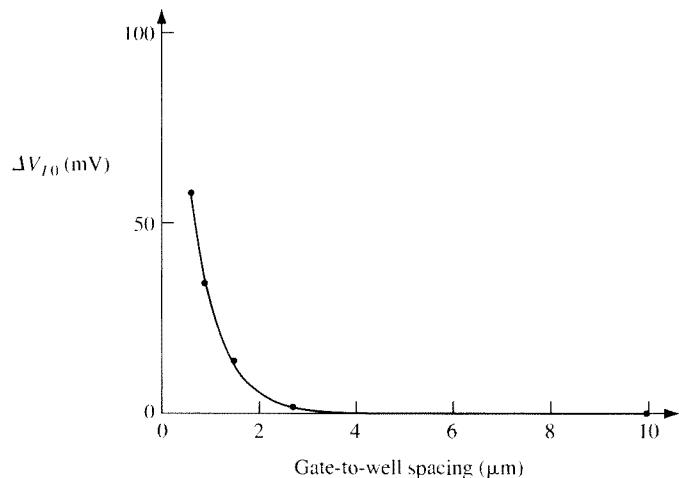
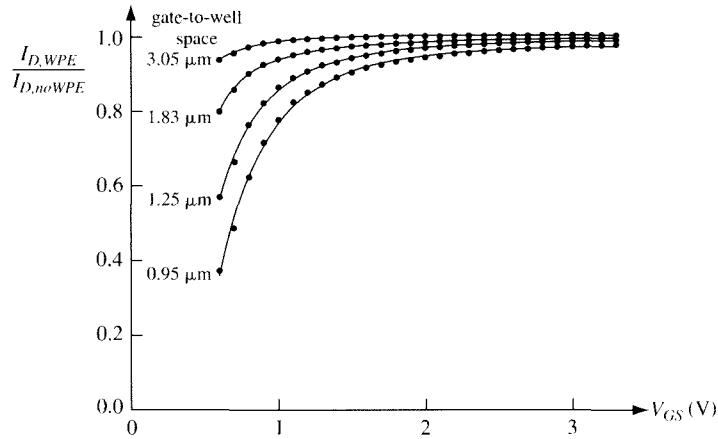


FIGURE 9.21
Measured threshold voltage shift vs. source side gate-to-well edge spacing; all other sides of the device are far from a well edge. These measurements are from a 130-nm technology. Adapted from Ref. 111 (Copyright © 2006 by IEEE).

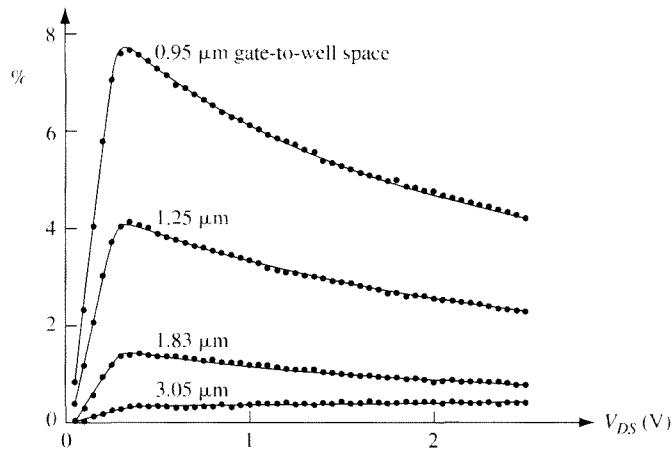
**FIGURE 9.22**

Well proximity effect on saturated drain current. The vertical axis shows the measured ratio of the saturation current of a device with the source end placed adjacent to the well edge and all other sides far from the well edge, to the saturation drain current of a transistor that has all sides far from well edges. $V_{DS} = 3.3$ V, $V_{SB} = 0$. The plot is based on measurements reported in Ref. 111 (Copyright © 2006 by IEEE).

Figure 9.21 shows the threshold voltage shift from the well proximity effect as a function of the well-edge to gate-edge spacing for a 130-nm technology.¹¹¹ The amount of the threshold shift reaches several tens of mV and is still considerable even when the well edge is 1 μm from the edge of the gate.

There are several subtle consequences of the WPE.¹¹¹ The alteration in the average bulk dopant concentration does significantly alter device characteristics. However, as Fig. 9.19 shows, the effect also induces a longitudinal variation in the bulk doping concentration. In some respects, this makes the transistor similar to a halo-doped device, with enhanced *p*-type doping at the source and drain ends of the channel, assuming the source and drain are both adjacent to well edges and at equal distances from them. As we have seen, this affects both the dc and small-signal (capacitance) characteristics of the transistor. In addition, if only one end, either source or drain, of a transistor is adjacent to a well edge, then it is only that end that undergoes significant enhancement of the bulk doping. Figure 9.22 shows the ratio of saturation drain currents of two devices. The first device has its source end placed adjacent to the well edge, and all other sides far from the well edge; the second device has all sides far from well edges. Significant current mismatch is seen.

Due to the WPE, the electrical behavior of a device can depend upon which end is connected in a circuit as the “drain” and which as the “source.” Figure 9.23 shows this effect. Note that the asymmetry is smallest at low drain bias, which unfortunately is where many measurements for MOSFET characterization are made and is large in saturation, where many MOSFETs are biased in analog circuits. Needless to say, “matched” transistors that are supposed to behave identically (an important design consideration for analog circuits) need to be identically placed and biased with respect to well edges.

**FIGURE 9.23**

Well proximity effect on device symmetry. Vertical axis is the measured percentage difference in current obtained when the source and drain terminals are interchanged. $V_{GS} = 1$ V, $V_{SB} = 0$ V. Adapted from Ref. 111 (Copyright © 2006 by IEEE).

Modeling of the WPE is difficult, because of its complex dependence on layout, and is based largely on empirical functions. The interested reader can find more information on the book's website.¹¹²

9.6 Stress Effects

Essentially of no importance before shallow trench isolation was introduced at around the 180-nm technology node, mechanical stress is now one of the key “knobs” that device technologists use to optimize transistor performance. Specifically, to give the maximum drain current for a given transistor W/L ratio, stressors are deliberately introduced into CMOS technologies. The first-order effect of stress is to alter the band structure of a semiconductor, which changes the effective mass of the mobile electrons or holes and therefore alters their mobility; it has been experimentally observed that stress also alters the saturation velocity. Finally, because stress affects the spacing of atoms in the lattice structure, it changes the solid-state diffusion rates of dopants through the silicon lattice, which causes changes in the final doping profiles in the body after thermal annealing steps. The threshold voltage and body effect coefficient, which depend on effective body doping, can therefore also be indirectly altered by stress effects.

Stress can be, unintentionally or deliberately, introduced into devices in CMOS technologies via different mechanisms.¹¹³ There are two key characteristics of stress that affect device performance. The first is whether the stress is uniaxial, i.e., is applied in a single direction, or whether the stress is biaxial, i.e., is applied in two directions. The second is whether the stress is compressive or tensile (the latter term meaning “stretching”). Thin films, called stress liners, are formed above transistors and provide uniaxial stress along the channel length direction; they are visible in

Fig. 1.40. The mechanism through which this is accomplished during fabrication is quite complex. The liner material is deposited at high temperature; as the wafer is cooled, the difference in thermal contraction between the Si and the liner material results in tensile or compressive stress, depending on the material used. The presence of the gate breaks the geometric symmetry of the induced stress, thus resulting in uniaxial strain along the length of the channel. *n*MOS mobility is enhanced by the use of a liner that generates tensile stress, i.e., pulls outward with respect to the center of the device; *p*MOS mobility is enhanced by a liner that generates compressive stress, i.e. that pushes inward toward the center of the device. Uniaxial stress on the channel of *p*MOS devices can be further enhanced by introducing germanium into the source and drain regions;¹¹⁴ the lattice spacing of the resulting SiGe regions is larger than that of silicon and so induces additional compressive stress in the channel. Biaxial stress can be induced by forming the channel in silicon that is grown at high temperature on top of a layer that is made of different crystalline material, e.g.,^{113–115} SiGe. The Si lattice spacing is smaller than that of SiGe; thus, the Si lattice is stretched to match up with the underlying SiGe lattice. Even though inducing stress effects adds complexity into manufacturing processes, it is worth the effort; *p*MOS saturation drain current in deliberately strained devices can be more than double that in unstrained devices, other things being equal.

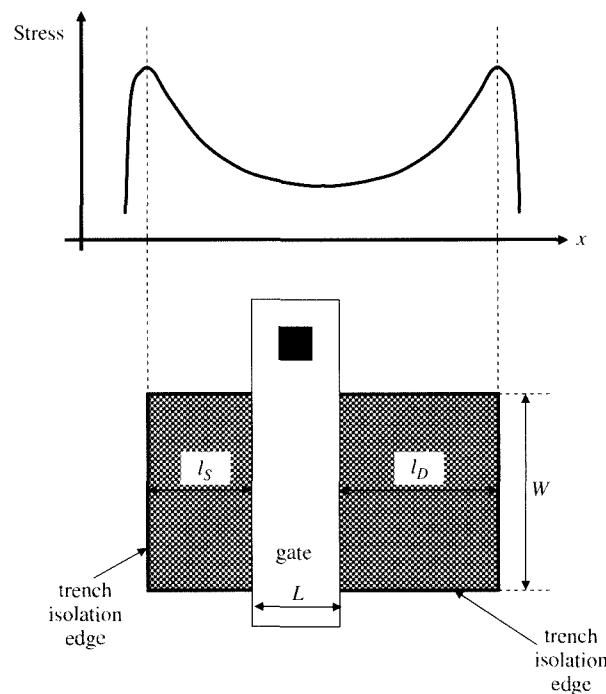
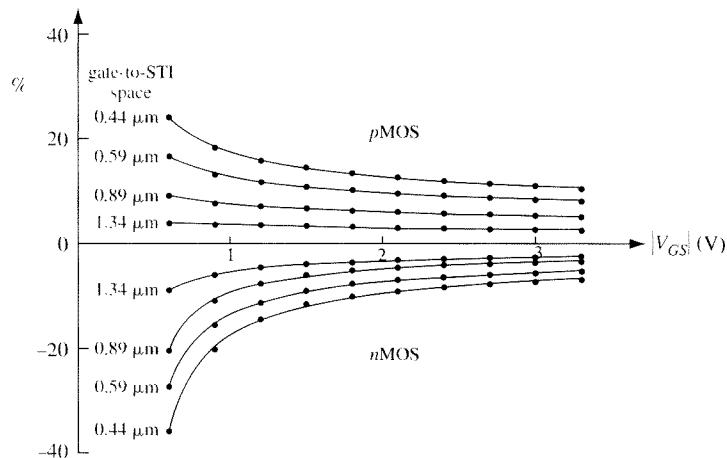


FIGURE 9.24

Stress at various points along the length of a transistor channel due to neighboring STI regions. The heavy line in the lower half indicates the trench isolation edge.

**FIGURE 9.25**

The STI proximity effect vs. V_{GS} with gate-to-STI spacing as a parameter. The vertical axis is the measured percentage difference in saturation current in comparison to a reference device with the STI edges far from the gate. Adapted from Ref. 111 (Copyright © 2006 by IEEE).

Originally, shallow trench isolation (STI) was the first source of stress observed to affect transistor performance.¹¹⁸ This source varies with layout and so has been incorporated into models. Figure 9.24 shows a top view of a MOS transistor; each edge of the STI that defines the boundary of the region in which the transistor is formed is a source of stress. The stress is observed to decrease approximately inversely with distance from the trench edge.¹¹⁶ Unlike the WPE, which can be reduced to practically zero by moving the well edges sufficiently far from a device, STI defines the boundaries of a device in the width direction and so cannot be completely prevented from affecting a device. One can only change the spacings l_S and l_D shown in the figure.

Figure 9.25 shows the STI proximity effect vs. V_{GS} , with the gate-to-STI spacing as a parameter. The vertical axis is the percentage difference in the saturation drain current of this device, compared with a reference device with the STI edges far from the gate. Shallow trench isolation induces compressive stress, which increases hole mobility and decreases electron mobility, hence the current increase for *p*MOS and decrease for *n*MOS, as seen.

The effect of stress on mobility is often modeled by assuming that (1) the mobility varies linearly with stress and (2) the stress varies inversely with distance from the STI edge to the center of the transistor's gate. This is then further refined to include the experimentally observed fact that, for a fixed STI region size, the change in mobility decreases as the channel length decreases. Due to the largely empirical nature of the resulting equations, fitting parameters are introduced, whose values are determined by comparing the model to experimental results.¹¹⁶ Additional geometry offsets and empirical temperature dependence are added in practice, and a similar stress dependence model is applied to the saturated carrier velocity. Further details,

and information on how to compute STI effects for multifingered devices and more complex layouts than shown in Fig. 9.24, are available elsewhere.^{101,116}

In contrast to the STI effect on mobility, which for a given $l_S + L + l_D$ decreases as L decreases, the STI effect on threshold voltage is observed to increase as L decreases. This is attributed to the increased sensitivity of short devices to the halo doping profiles, which are changed by stress via the impact of the latter on dopant diffusion.¹¹¹ An additive correction to the threshold is used to model this effect. For all-region models, a shift in V_{FB} is used instead. If it is desired that devices be well matched to each other, one should make sure that the stress on them is matched, too, by making sure that corresponding edges are at the same distance from stress sources.

At present all modeling of stress effects is empirical and is not based on first-principles theoretical analysis. It is unlikely that this situation will change any time soon. The innovation in technologies to enhance device performance by deliberately introducing stress continues to advance rapidly; theoretical modeling of the effects of stress on device behavior lags the technological innovation rate, both because of the increasing complexity of the geometric relationships between stress sources and device behavior, and because of the immaturity of direct modeling of how stress sources alter physical parameters that control device performance.

9.7 Statistical Variability

During semiconductor manufacturing, it is desirable for the physical and geometrical properties that define a transistor, e.g., t_{ov} , N_A , V_{FB} , ΔW in (1.6.1) and ΔL in (1.6.2), to have the exact values intended. However, this is not possible; each parameter is characterized by a statistical distribution. Thus, such parameters vary between devices within one chip (called a “die”[†]), vary between dice within a wafer, vary between wafers within a lot (wafers are manufactured in groups, called “lots”), and vary between different manufacturing lots and between different manufacturing factories (called “fabs”) for the same process. These variations in physical properties cause variations in the electrical properties of transistors, such as current and charges for specific values of applied terminal voltages. In turn, these variations in device electrical properties lead to variations in the electrical performance of the circuits they form. An understanding of statistical variability in MOS transistors is therefore important for MOSFET modeling and circuit design.

Sources of Variability There are many physical sources of variability. Semiconductor manufacturing processes consist of a sequence of steps,¹¹⁷ including photolithographic patterning, etching, ion implantation, and, at higher temperatures, annealing steps (to repair damage to the silicon lattice from implantation and to redistribute the implanted dopants through solid-state diffusion), and oxidation. The results of performing one particular step can vary between different, yet supposedly identical, pieces of equipment (called “tools”) and can change over time even for the same tool. This leads to variations between wafers from different fabs and between different wafer lots from the same fab.

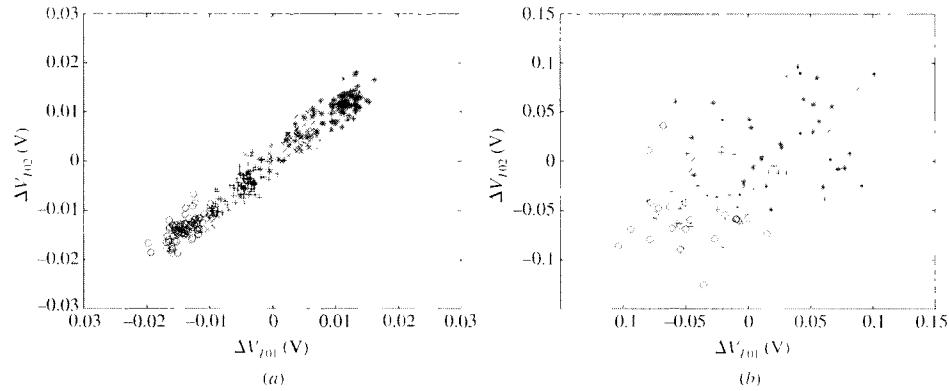
[†] The plural is “dice,” although often “die” is used for the plural, too.

These types of variations will generally cause a difference in specific physical properties from one wafer to another, but not from one chip to another on the same wafer.

Oxidation and some types of annealing are done in furnaces, and there can be nonuniformities in the temperature distributions in the furnaces. Also, oxidation is done with H_2O introduced into the furnace and its concentration can be nonuniform throughout the furnace. These steps can therefore lead to variations between individual wafers in a lot and to nonuniformities across a wafer and therefore across a die. Imperfections in the optical systems used during photolithographic patterning lead to variations across a die in the on-wafer patterns that are formed to define transistor lengths and other dimensional features. In processing steps that involve chemicals such as etchants, which selectively remove materials from the surface of a wafer, the local density of the material being etched affects the rate of etching (a phenomenon called “loading”), which introduces additional within-die variations.

Besides these macroscopic physical variations, there are also variations at the atomic level. The physical boundaries of the various structural regions that make up a transistor, such as the gate, cannot be manufactured to be perfectly smooth at an atomic level, but rather have microscopic variations. This leads to line edge roughness^{119,120} and makes the gate shape nonrectangular. For many purposes a gate can still be assumed to be rectangular and to be characterized by its average length and width; however, because the variations in the edges that define the gate are statistical, the average L and W will be different for different devices within a die, between dice and between wafers, even if they are designed with the same desired on-wafer L and W . In addition, the ion beams used to implant dopants into the silicon are composed of discrete particles; therefore, in a specific region implanted for a specific time duration, there will be statistical fluctuations in the total number of implanted ions at different sites in a die that are intended to be identical. This leads to random dopant fluctuations, which affect N_A and V_{T0} among other parameters.^{121–125} Uncontrollable microscopic fluctuations in t_{ox} also contribute to variations in device behavior.¹²⁶ As may be expected, these atomic-level variations can be reduced by increasing the physical size of a device, so the variations average over a large length or area or volume; however, the technology scaling path is in the direction to shrink (Sec. 5.15) size; so the importance of microscopic level variability has significantly increased today.

Global and Local Variations Figure 9.26 shows measured V_{T0} variations, with respect to mean value, from four different wafer lots (represented by the different symbols) from pairs of identical n MOS transistors placed next to each other, for large (wide/long) devices (part *a*) and small (narrow/short) devices (part *b*). Note the different scales used for the axes in each part. If there were no variability, we would have $\Delta V_{T02} = \Delta V_{T01} = 0$, and the plots would be just a dot at the center. Instead, the variations shown are observed. The overall amount of variation is much larger for the small device than for the big device (about five times greater here). This trend, of the variation in electrical performance increasing as device size decreases, is almost always observed. From the previous discussion, part of this is because the amount of variability from atomic-level fluctuations increases as device dimensions get smaller. In addition, the impact of variations in ΔL on the channel length $L = L_m - \Delta L$ is relatively greater for devices of smaller mask length L_m . (This directly affects I_{DS} and the

**FIGURE 9.26**

Measured V_{T0} scatter plots of pairs of identical, adjacent nMOS devices with (a) $W = L = 24.9 \mu\text{m}$ and (b) $W = 0.3 \mu\text{m}$, $L = 0.15 \mu\text{m}$. Different symbols represent different wafer lots; the variation is with respect to the mean value for each geometry. Note the different scales used for the axes in (a) and (b).

charges, and indirectly affects ΔW through the dependence of the latter on length). Similarly, variations in W have a relatively greater effect on narrower devices.

As seen in the figure, the correlation between the threshold voltages of the two large devices is much greater than that of the two small devices, and the systematic effect of lot-to-lot variations is clear in the “clumping” of data for the different wafer lots (shown by different symbols) for the big devices; however, this is difficult to discern in the data for the small devices. This is again a consequence of the atomic-level variations, which are uncorrelated between devices, and which have a much greater effect on the small devices. The correlation seen for the large devices is because the variations are primarily due to *global* variation in the average value of physical parameters between dice, wafers, and lots. The lack of correlation for the small devices is because the variations are primary due to *local* variation, or *mismatch*, which comes from random atomic-level variations that exist even between adjacent transistors.

It is important to distinguish whether statistical variations are correlated or uncorrelated between devices on the same die. The overall variability of electrical characteristics of transistors, e.g., the drain current, depends on geometry and bias, but is at best a few percent. Even though overall variability cannot be reduced, the *difference* in performance between identically sized devices placed adjacent to one another on a die can be small, much smaller than the overall variation, and can be controlled by appropriate choice of device size (the bigger the device, the smaller the standard deviation of the differences, assuming the size is not too large, in which case global variations must be considered) and bias. For this reason, precision MOS circuits, especially analog ones, are often designed to leverage this high degree of matching between devices, so analysis and design of these circuits needs to distinguish between correlated global statistical variations (which have little effect on differential performance) and uncorrelated mismatch variations (which determine differential performance).¹²⁷

Modeling Variability To represent global statistical variability, it is possible to extract model parameter values from measurements of many devices and use sophisticated numerical techniques to generate statistical models based on these values.¹²⁸ However, the root cause of statistical fluctuations in the electrical behavior of MOSFETs is variations in *physical* parameters such as t_{ox} , N_A , V_{FB} , ΔL , and ΔW statistical modeling can also be done through these parameters rather than (possibly nonphysical) model parameters.¹²⁹ Using physical parameters as the basis for statistical modeling enables predictive estimation of variation, naturally captures correlations between electrical performances (for example, between measured V_{T0} and g_m through their common dependence on t_{ox}), and can provide guidance on how to minimize the effect of statistical variations on circuit performance. It also provides a common basis for modeling global and local variations. We will thus start by using physical parameters for statistical modeling. In practice, although the mobility μ is a material property and not strictly a definable or controllable physical property, such as t_{ox} or N_A , it is convenient to treat it as an independent statistical modeling parameter as this enables accurate modeling of the g_m of wide, long transistors.

Each parameter used for statistical modeling comprises a nominal value to which global and local variations are applied. For parameters such as t_{ox} , N_A , and μ , it is natural to consider variations to be relative to their nominal value, so we have for t_{ox} , for example,

$$t_{ox} = t_{ox,nom} \left(1 + \frac{\delta t_{ox,g}}{t_{ox,nom}} + \frac{\delta t_{ox,l}(W, L)}{t_{ox,nom}} \right) \quad (9.7.1)$$

where $t_{ox,nom}$ is the nominal value of t_{ox} , $\delta t_{ox,g}/t_{ox,nom}$ is the relative global variation, and $\delta t_{ox,l}/t_{ox,nom}$ is the relative local variation. The standard deviation of the relative global variation is independent of the size of a transistor. Because the local variation comes from microscopic fluctuations in t_{ox} , it could be expected that these will average out more for a larger gate area than for a smaller gate area. This is indeed the case, and it can be shown that the variance (the square of the standard deviation) of $\delta t_{ox,l}/t_{ox,nom}$ is proportional to $1/(WL)$.¹³⁰ Similar global and local variations are applied to N_A and μ , and the variances of the local variation of N_A and μ also scale as $1/(WL)$.

For V_{FB} , ΔL and ΔW , it is better to consider the variations to be absolute quantities, not relative quantities. For example, the nominal value of ΔW could be zero, so a relative, or percentage, variation does not make much sense; the absolute variation in effective channel width is to first order independent of the nominal value of that quantity. The same is true for ΔL and V_{FB} . For the flatband voltage we therefore have

$$V_{FB} = V_{FB,nom} + \delta V_{FB,g} + \delta V_{FB,l}(W, L) \quad (9.7.2)$$

where $V_{FB,nom}$ is the nominal value of V_{FB} and $\delta V_{FB,g}$ and $\delta V_{FB,l}$ are the global and local variations. Averaging of microscopic fluctuations in V_{FB} again is over the area

of a device, so the variance of the local variation $\delta V_{FB,l}(W,L)$ is proportional to $1/(WL)$. For line edge roughness, variations in ΔL are not influenced by length and average over the width of a device; therefore

$$\Delta L = \Delta L_{\text{nom}} + \delta \Delta L_g + \delta \Delta L_l(W) \quad (9.7.3)$$

where $\delta \Delta L_g$ is the global variation, and the variance in the local variation $\delta \Delta L_l$ is proportional to $1/W$.¹³⁰ Similarly width variations average over the length of a device, so

$$\Delta W = \Delta W_{\text{nom}} + \delta \Delta W_g + \delta \Delta W_l(L) \quad (9.7.4)$$

where $\delta \Delta W_g$ is the global variation, and the variance of the local fluctuation $\delta \Delta W_l$ is proportional to $1/L$.¹³⁰

Simulations of statistical variability for devices within a circuit are often done using the Monte Carlo technique, in which sets of random values for each statistical variable are generated. The values for the variations of the global component of each parameter are common to all devices (which is why they are called “global” variables), and separate statistical values are generated for the variations of each local parameter for each individual instance of a transistor in a circuit. This is how mismatch between devices is simulated using these models. The total variance of a parameter is the sum of the variances of its global and local statistical components (which are independent), so, for example, for the flatband voltage we have

$$\sigma_{V_{FB}}^2 = \sigma_{\delta V_{FB,g}}^2 + \sigma_{\delta V_{FB,l}}^2(W,L) \quad (9.7.5)$$

where σ denotes the standard deviation (the square root of the variance). Notice that this naturally embodies the observed increase in total variation as geometry decreases, because $\sigma_{\delta V_{FB,l}}^2$ is proportional to $1/WL$. The correlation coefficient[†] between the flatband voltages of two devices, of lengths L_1 and L_2 and widths W_1 and W_2 , is (Prob. 9.7)

$$\rho_{V_{FB,12}} = \frac{\sigma_{\delta V_{FB,g}}^2}{\sqrt{(\sigma_{\delta V_{FB,g}}^2 + \sigma_{\delta V_{FB,l}}^2(W_1, L_1))(\sigma_{\delta V_{FB,g}}^2 + \sigma_{\delta V_{FB,l}}^2(W_2, L_2))}} \quad (9.7.6)$$

In the limit for large devices, $\sigma_{\delta V_{FB,l}} \ll \sigma_{\delta V_{FB,g}}$ and the correlation coefficient approaches 1. For very small devices, $\sigma_{\delta V_{FB,l}} \gg \sigma_{\delta V_{FB,g}}$ and the correlation coefficient approaches zero. This is precisely the behavior observed in the data in Fig. 9.26. (V_{F0})

[†] The correlation coefficient for two random variables x and y is defined as $E[(x - \mu_x)(y - \mu_y)] / (\sigma_x \sigma_y)$ where E denotes expected value. $\mu_x = E(x)$ and $\mu_y = E(y)$ are the mean values of the random variables x and y , and σ_x and σ_y are their standard deviations. The derivation of (9.7.6) from this definition is dealt with in Prob. 9.7.

is not a physical parameter, but is a function of V_{FB} , N_A , and t_{ox} , and for short and narrow devices, through the geometry dependence of threshold voltage, it is also a function of ΔL and ΔW ; the cumulative influence of the global and local variations in the physical parameters that control V_{T0} leads to the decrease in correlation as size decreases, seen in the data of Fig. 9.26.)

Measurement of Variability A commonly used approach is to measure the global and local variances of each parameter directly. For the total variance, i.e., the sum of the global and local variances in (9.7.5), this is done using special test structures that are placed in the scribe lines (the vertical and horizontal spaces between dice on a wafer, which is where the wafer is cut when finished to yield the final die for packaging). These are measured electrically as part of the manufacturing process control procedures. Although it is not possible to measure all parameters directly, there are techniques that provide good estimates. For example, the dielectric capacitance per unit area is often measured from a large gate capacitance structure (although dc gate leakage currents are making this procedure more complex), and this allows t_{ox} to be computed from (2.2.4). N_A cannot be measured electrically, but its nominal value can be found from detailed device characterization as follows. The threshold voltage is measured both at zero substrate bias, and a nonzero substrate bias V_{SB} ; these two values must satisfy, from (4.7.20a), (3.4.3), and (2.2.4):

$$V_T(V_{SB}) = V_{T0} + \frac{\sqrt{2q\epsilon_s N_A t_{ox}}}{\epsilon_{ox}} \left(\sqrt{\phi_0 + V_{SB}} - \sqrt{\phi_0} \right) \quad (9.7.7)$$

The dependence of ϕ_0 on N_A is weak, and an approximate value can be assumed for it. Thus, N_A can be computed from (9.7.7) if t_{ox} is known. However, it is often simpler to directly use the quantities that are measured for statistical modeling, rather than trying to use these to infer the underlying values of more physical parameters. Consequently V_{T0} , rather than V_{FB} , and γ , rather than N_A , are often used for statistical modeling, along with t_{ox} , μ , ΔL , and ΔW . From measurements of these parameters from a sufficiently large number of wafer lots, the sum of the variances of the global and local components of variation can be calculated; both contribute to the observed total variance from (9.7.5).

Given the approximate nature of the measurements of some physical parameters, and the fact that for characterizing the local variation high precision is needed (the amount of mismatch between large devices can be well below 1 percent), a common approach for local variation modeling is to use different parameters, which can be measured more reliably, as a basis for statistical MOSFET modeling. In this approach V_{T0} is still used, and the so-called gain factor

$$\beta = \frac{W}{L} \mu C'_{ox} \quad (9.7.8)$$

is introduced as a basic statistical parameter; both are taken to have a $1/(WL)$ dependence for their variance;^{130,131} in some treatments, the body effect coefficient γ and

additional parameters can also be included.^{132,133} The gain factor for a transistor is computed from the slope of the characteristic in Fig. 4.33 at the point where this slope is maximum. The slope at that point is actually the peak transconductance as determined from numerical differentiation of measured $I_{DS} - V_{GS}$ data in nonsaturation, at low V_{DS} , typically 50 or 100 mV, for $V_{SB} = 0$. The basis for this is the source-referenced strong-inversion model (4.7.24a), which in nonsaturation gives

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \frac{W}{L} \mu C'_{ox} V_{DS} = \beta V_{DS}^{\dagger} \quad (9.7.9)$$

With careful attention to test structure design and measurement procedures, very precise measurements of V_{T0} and β mismatch can be made.¹³⁴⁻¹³⁶ Measurement over multiple geometries enables characterization of the variances of the mismatch as a function of geometry, and usually these are expressed through the quantities A_{VT} and A_{β} in the following formulas:¹³⁰

$$\sigma_{\delta V_{T0}}^2 = \frac{A_{VT}}{WL} \quad (9.7.10)$$

$$\sigma_{\delta \beta / \beta}^2 = \frac{A_{\beta}}{WL} \quad (9.7.11)$$

where a relative variation is used for the gain factor mismatch. The A_{VT} and A_{β} coefficients can be used to compare the degree of matching of different device types and different technologies, and they are used for mismatch simulation by defining the variances of V_{T0} and β to use for generating statistical samples for each size device to use in Monte Carlo simulation.

Care needs to be exercised when using parameters such as V_{T0} , β , and A_{β} for statistical modeling. The primary problem is that variations in these parameters can be correlated through the underlying parameters t_{ox} , N_A , V_{FB} , ΔL , and ΔW . If these correlations are significant and are not taken into account, then simulations will *underpredict* the overall statistical variation (Prob. 9.8). In addition, some of the second-order effects that are included in detailed circuit simulation models, such as polysilicon depletion, quantum mechanical effects, and gate tunneling current (Chap. 5), depend on the values of parameters such as t_{ox} and N_A , and if β and γ are used as statistical parameters rather than t_{ox} and N_A , then statistical variations in these effects will be ignored. More important, in modern device structures, V_{T0} mismatch can be significantly affected by the halo implants (Sec. 9.4) at the ends of the channel.¹³⁷ The fluctuations in the doping of the halo regions does not depend on channel length, but averages over channel width only, therefore the $1/(WL)$ dependence assumed

[†] Note that this simple formula does not include the effect of effective mobility dependence on the transverse field, discussed in Sec. 4.11.

in (9.7.10) needs to be modified to be a $1/W$ dependence for the halo region.¹³⁷ Nevertheless, the inverse-area-dependent mismatch model based on V_{T0} and β is widely used and has proven effective for a long time and for a wide variety of devices and technologies.

The practical requirement of being able to measure statistical variations accurately has led us away from using physical parameters, such as N_A and V_{FB} , as a basis for statistical modeling. However, it is still possible to use these parameters for the purposes of statistical modeling and to infer, rather than directly measure, their statistical variations. The procedure is called *backward propagation of variation* and involves some mathematical complexity.^{138,139} We will now provide a brief outline of this procedure. If any parameter of a model is perturbed, this will lead to perturbations in the predictions of that model for drain current and charges, for derived quantities such as V_{T0} , β , and γ , or even for circuit performance parameters, such as the switching delay of digital circuits. By appropriate adjustment of the perturbations of the physical model parameters, the simulated variations in the electrical performance of devices or circuits can be made to match the measured values. In this way, rather than measuring the parameter variations directly, they are inferred from their effect on the variations of other measured quantities. For global statistical modeling, these are the variances observed during the fab process control measurements made on scribe grid structures.¹³⁸ For mismatch modeling, these are not V_{T0} and β mismatch, but rather are the variances of the difference in drain current between adjacent, identical devices measured over a wide range of device geometries and biases.¹³⁹ The parameter variations are implemented as separate global and local perturbations, as in (9.7.1) through (9.7.4), and additional statistical parameters can be included if measurements that are affected by them are available, for example, the length and doping concentration of a halo implant region.¹³⁹

Direct measurement of parameters can involve device structures or applied biases that are not representative of those used in real circuits to try to make the effect of a parameter most observable (for example, using a wide and long transistor to minimize narrow and short channel effects, or biasing at low V_{DS} , to minimize the influence of velocity saturation and CLM). Models are not exact. They are approximations, and thus determining parameter values from one size or layout of device at one bias level, and using it to predict device behavior for a different size or layout or bias level, can introduce error. Where possible, statistical variations should be determined from device sizes, layouts, and biases that are most commonly used in circuits.

Variations in measurements do not just include the variations in the devices being measured. Measurements include noise from the measurement equipment, variations from temperature drifts over time (even for a temperature-controlled measurement environment), and variations in parasitic series resistance from the probe-to-pad contacts and other sources. Care needs to be exercised to ensure that the measured statistical data represent the actual device variations and are not corrupted by variability from these other sources.

Historically, characterization of global and local variations has been done as two separate steps. Some time ago, the local component of variation was a lot smaller than the global component, so the total variation from fab measurements was assumed to be only from the global component, and the local variation was measured

separately and then added to the global variation. This is still reasonable if the fab measurements are made on large structures. However, for small devices, the local component now dominates, as Fig. 9.26 illustrates, and so characterization of the local and global variations needs to be done self-consistently, otherwise the local component will be “double-counted” for small devices.

With statistical parameters like those in (9.7.1) through (9.7.4), Monte Carlo analysis is done by including, as statistical sampling variables, a global variation for each parameter and a separate local variation for each device. Because many simulation runs, typically tens to hundreds, are used for a Monte Carlo analysis, a lot of computer simulation time can be required. More efficient statistical simulation can be done by generating a few specific values for each parameter, which are designed to span the range expected in manufacturing. These are called *corner*, *case*, or *skew* models. The variations can be specified for the parameters themselves, or for specific measures of electric performance of devices, in which case the parameters are adjusted so the specified variations in the electrical performances are met.¹³⁸ A limited number of corner models are usually generated, often based on high and low values of drain current and leakage as these are key for digital circuits. Corner models can efficiently model variability, but they have limitations: they are based on targeting specific performances for specific types of circuits. If they are used for analysis of different measures of performance for different types of circuits, which do not correlate with those used to generate the corner models, then they will not accurately represent the variability. This is why corner models developed for digital circuit design are in general inadequate for analog and RF circuit design.¹⁴⁰

Besides parameter correlations within one type of device, correlations exist between different types of devices, and these should also be factored in statistical models. The gate dielectric growth step is common to *n*MOS and *p*MOS devices, and so the global component of variation is highly correlated between both types of devices; it usually is assumed to be perfectly correlated. The patterning and etching of the gates is also common to *n*MOS and *p*MOS devices; therefore a significant component of the global variation in ΔL is common. The correlation is not perfect, as the implants and anneals that define the source and drain regions, which also contribute to ΔL , are different. This partial correlation can be modeled by including a variation component related to the polysilicon dimension variation, which is common to both *n*MOS and *p*MOS devices, and separate variations in the source and drain encroachments under the gate, which are different for the separate devices. The gate polysilicon can also be used as a separate layer for resistors in some manufacturing processes. In that case, the width variation of such a polysilicon resistor should be linked to the global variation in length of the MOS transistor gates. Simulation models are not always set up to that level of detail.

References

1. K. G. Aubuchon, *Proceedings of the International Conference on Properties and Use of M.I.S. Structures*, Grenoble, France, 1969, p. 575.
2. R. W. Bower and H. G. Dill, *IEEE International Electron Devices Meeting*, paper 16.6 (unpublished), 1966.

3. M. R. MacPherson, "The adjustment of MOS transistor threshold voltage by ion implantation," *Applied Physics Letters*, vol. 18, pp. 502–504, 1971.
4. T. Masuhara, M. Nagata, and N. Hashimoto, "A high-performance n-channel MOS LSI using depletion type load elements," *IEEE Journal of Solid-State Circuits*, vol. SC-7, pp. 224–231, 1972.
5. W. S. Ruska, *Microelectronics Fabrication*, McGraw-Hill, New York, 1987.
6. S. M. Sze, *VLSI Technology*, 2nd ed., McGraw-Hill, New York, 1988.
7. J. Y. Chen, *CMOS Devices and Technology for VLSI*, Prentice Hall, Englewood Cliffs, N.J., 1990.
8. S. K. Ghandhi, *VLSI Fabrication Principles—Silicon and Gallium Arsenide*, 2nd ed., Wiley, New York, 1994.
9. M. R. MacPherson, "Threshold shift calculation for ion implanted MOS devices," *Solid-State Electronics*, vol. 15, pp. 1319–1326, 1972.
10. M. Kamoshida, "Threshold voltage and gain term β of ion implanted n-channel MOS transistors," *Applied Physics Letters*, vol. 22, pp. 404–405, 1973.
11. G. Doucet and F. Van de Wiele, "Threshold voltage of nonuniformly doped structures," *Solid-State Electronics*, vol. 16, pp. 417–423, 1973.
12. P. P. Peressini and W. S. Johnson, "Threshold adjustment of n-channel enhancement mode FETs by ion implantation," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., December 1973, pp. 467–486.
13. R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassons, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. SC-9, pp. 256–268, 1974.
14. E. C. Douglas and A. G. F. Dingwall, "Ion implantation for threshold control in COSMOS circuits," *IEEE Transactions on Electron Devices*, vol. ED-21, pp. 324–331, 1974.
15. M. Kamoshida, "Electrical characteristics of boron-implanted n-channel MOS transistors," *Solid-State Electronics*, vol. 17, pp. 621–626, 1974.
16. V. L. Rideout, F. H. Gaensslen, and A. LeBlanc, "Device design considerations for ion-implanted n-channel MOSFETs," *IBM Journal of Research and Development*, January 1975, p. 50.
17. G. Doucet, F. Van de Wiele, and P. Jespers, "Theoretical and experimental study of MOS transistors nonuniformly doped by SILOX technique," *Solid-State Electronics*, vol. 19, pp. 191–199, 1976.
18. R. R. Troutman, "Ion-implanted threshold tailoring for insulated gate field-effect transistors," *IEEE Transactions on Electron Devices*, ED-24, pp. 182–192, 1977.
19. H. Feltl, "Onset of heavy inversion in MOS devices doped nonuniformly near the surface," *IEEE Transactions on Electron Devices*, vol. ED-24, pp. 288–289, 1977.
20. E. Demoulin and F. Van de Wiele, "Ion implanted MOS transistors," pp. 617–676, in *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
21. G. Merckel, "Ion implanted MOS transistors—Depletion mode devices," pp. 677–688, in *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
22. J. R. Brews, "Threshold shifts due to nonuniform doping profiles in surface channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-26, p. 1696, 1979.
23. M. Nishida and M. Aoyane, "An improved definition for the onset of heavy inversion in a MOS structure with nonuniformly doped semiconductors," *IEEE Transactions on Electron Devices*, vol. ED-27, pp. 1222–1230, 1980.
24. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
25. J. R. Brews, "Physics of the MOS transistor," chap. 1 in *Silicon Integrated Circuits, Part A*, D. Kahng (editor), Applied Solid-State Science Series, Academic Press, New York, 1981.
26. L. M. Dang and H. Iwai, "Modeling the impurity profile of an ion-implanted IGFET for the calculation of threshold voltages," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 116–117, 1981.
27. P. K. Chatterjee, J. E. Leiss, and G. W. Taylor, "A dynamic average model for the body effect in ion implanted short-channel ($L = 1 \mu\text{m}$) MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 606–607, 1981.
28. K.-Y. Fu, "A new analysis of the threshold voltage for non-uniform ion-implanted MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 1810–1813, 1982.

29. K. Shenai, "Analytical solutions for threshold voltage calculations in ion-implanted IGFETs," *Solid-State Electronics*, vol. 26, pp. 761–766, 1983.
30. A. H. Marshak and R. Shrivastava, "On threshold and flat-band voltages for MOS devices with polysilicon gate and nonuniformly doped substrate," *Solid-State Electronics*, vol. 26, pp. 361–364, 1983.
31. F. Van de Wiele, "On the flat-band voltage of MOS structures on nonuniformly doped substrate," *Solid-State Electronics*, vol. 27, pp. 824–826, 1984.
32. D. A. Divekar and R. I. Dowell, "A depletion-mode MOSFET model for circuit simulation," *IEEE Transactions on Computer-Aided Design*, vol. CAD-3, p. 80, 1984.
33. D. A. Antoniadis, "Calculation of threshold voltage in nonuniformly doped MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 303–307, 1984.
34. P. Ratnam and C. A. T. Salama, "A new approach to the modeling of nonuniformly doped short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1289–1298, 1984.
35. C. R. Viswanathan, B. C. Burkey, G. Lubberts, and T. J. Tredwell, "Threshold voltage in short channel MOS devices," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 932–940, 1985.
36. H. J. Park and C.-K. Kim, "An empirical model for the threshold voltage of enhancement nMOSFETs," *IEEE Transactions on Computer-Aided Design*, vol. CAD-4, p. 629, 1985.
37. C.-Y. Wu and Y.-W. Daih, "An accurate mobility model for the I-V characteristics of n-channel enhancement-mode MOSFETs with single-channel boron implantation," *Solid-State Electronics*, vol. 28, pp. 1271–1278, 1985.
38. D. M. Rogers, J. D. Hayden, and D. D. Rinerson, "Model for the channel-implanted enhancement-mode IGFET," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 955–964, 1986.
39. G. T. Wright, "Physical and CAD models for the implanted-channel VLSI MOSFET," *IEEE Transactions on Electron Devices*, vol. ED-34, pp. 823–833, 1987.
40. A. Dasgupta and S. K. Lahiri, "An analytical threshold voltage model of short-channel MOSFETs with implanted channels," *IEEE Transactions on Electron Devices*, vol. ED-34, p. 1177, 1987.
41. R. V. Booth, M. H. White, H.-S. Wong, and T. J. Krutsick, "The effect of channel implants on MOS transistor characterization," *IEEE Transactions on Electron Devices*, vol. ED-34, p. 2501, 1987.
42. N. D. Arora, "Semi-empirical model for the threshold voltage of a double implanted MOSFET and its temperature dependence," *Solid-State Electronics*, vol. 30, pp. 559–569, 1987.
43. T. J. Krutsick and M. H. White, "Consideration of doping profiles in MOSFET mobility modeling," *IEEE Transactions on Electron Devices*, vol. 35, pp. 1153–1155, 1988.
44. S. Karmalkar and K. N. Bhat, "A process-parameter-based circuit simulation model for ion-implanted MOSFETs and MESFETs," *IEEE Journal of Solid-State Circuits*, vol. 24, pp. 139–145, 1989.
45. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
46. P.-S. Lin and C.-Y. Wu, "A new simplified two-dimensional model for the threshold voltage of MOSFETs with nonuniformly doped substrate," *IEEE Transactions on Electron Devices*, vol. 38, p. 1376, 1991.
47. S. Karmalkar and K. N. Bhat, "The shifted-rectangle approximation for simplifying the analysis of ion-implanted MOSFETs and MESFETs," *Solid-State Electronics*, vol. 34, pp. 681–692, 1991.
48. L. Lukasiak and A. Jakubowski, "The influence of nonuniform doping profile on I-V characteristics of MOS transistors," *IEEE Transactions on Electron Devices*, vol. 40, pp. 453–455, 1993.
49. K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, N.J., 1993.
50. U. Cilingiroğlu, *Systematic Analysis of Bipolar and MOS Transistors*, Artech House, Boston, 1993.
51. N. Arora, *MOSFET Models for VLSI Circuit Simulation—Theory and Practice*, Springer-Verlag, Vienna, 1993.
52. T. Skotnicki, C. Denat, P. Senn, G. Merckel, and B. Hennion, "A new analog/digital CAD model for sub-halfmicron MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, San Francisco, December 1994, pp. 165–168.
53. J.-J. Maa and C.-Y. Wu, "A new simplified threshold-voltage model for n-MOSFETs with nonuniformly doped substrate and its application to MOSFETs miniaturization," *IEEE Transactions on Electron Devices*, vol. 42, p. 1487, 1995.

54. C. Lallement, M. Bucher, and C. Enz, "Modeling and characterization of non-uniform substrate doping," *Solid-State Electronics*, vol. 41, pp. 1857–1861, 1997.
55. F. Gámiz, J. A. Lopez-Villanueva, J. B. Roldán, and J. E. Carceller, "Influence of the doping profile on electron mobility in a MOSFET," *IEEE Transactions on Electron Devices*, vol. 43, pp. 2023–2025, 1996.
56. J. R. Brews, "Sensitivity of subthreshold current to profile variations in long-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 43, p. 2614, 1996.
57. C. McAndrew, "Useful numerical techniques for compact modeling," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 121–126, April 2002.
58. N. Rao, private communication.
59. J. R. Edwards and G. Mar, "Depletion-mode IGFET made by deep ion implantation," *IEEE Transactions on Electron Devices*, vol. ED-20, pp. 283–289, 1973.
60. J. S. T. Huang, "Characteristics of a depletion-mode IGFET," *IEEE Transactions on Electron Devices*, vol. ED-20, pp. 513–515, 1973.
61. T. W. Sigmon and R. Swanson, "MOS threshold shifting by ion implantation," *Solid-State Electronics*, vol. 16, pp. 1217–1232, 1973.
62. A. M. Mohsen and F. J. Morris, "Measurements on depletion-mode field-effect transistors and buried channel MOS capacitors for the characterization of bulk transfer charge-coupled devices," *Solid-State Electronics*, vol. 18, pp. 407–416, 1975.
63. J. R. Verjans and R. J. Van Overstraeten, "Electrical characteristics of boron-implanted n-channel MOS transistors for use in logic circuits," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 862–868, 1975.
64. J. S. T. Huang and G. W. Taylor, "Modeling of an ion-implanted silicon-gate depletion mode IGFET," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 995–1001, 1975. See also W. Marcinia and H. Madura, "Comments on the Huang and Taylor model of ion-implanted silicon-gate depletion-mode IGFET," *Solid-State Electronics*, vol. 28, pp. 313–315, 1985.
65. P. E. Schmidt and M. B. Das, "D.C. and high-frequency characteristics of built-in channel MOS-FETs," *Solid-State Electronics*, vol. 21, pp. 495–505, 1978.
66. G. R. Mohan Rao, "An accurate model for a depletion mode IGFET used as a load device," *Solid-State Electronics*, vol. 21, pp. 711–714, 1978.
67. R. A. Haken, "Analysis of the deep depletion MOSFET and the use of the dc characteristics for determining bulk-channel charge coupled device parameters," *Solid-State Electronics*, vol. 21, pp. 753–761, 1978.
68. T. E. Hendrikson, "A simplified model for subpinchoff condition in depletion mode IGFETs," *IEEE Transactions on Electron Devices*, vol. ED-25, pp. 435–441, 1978.
69. Y. A. El-Mansy, "A nonlinear CAD model for the depletion-mode IGFET," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., 1978, pp. 20–25.
70. K. Nishiuchi, H. Oka, T. Nakamura, H. Ishikawa, and M. Shinoda, "A normally-off type buried channel MOSFET for VLSI circuits," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., December 1978, p. 26.
71. F. H. Gaensslen and K. C. Jaeger, "Temperature dependent threshold behavior of depletion mode MOSFETs," *Solid-State Electronics*, vol. 22, pp. 423–430, 1979.
72. M. Wordeman, "Characterization of depletion-mode MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., December 1979, pp. 26–29.
73. G. Baccarani, F. Landini, and B. Ricco, "Depletion-mode MOSFET model including a field-dependent surface mobility," *IEE Proceedings*, vol. 127, part I, pp. 62–66, 1980.
74. Y. A. El-Mansy, "Analysis and characterization of the depletion-mode IGFET," *IEEE Journal of Solid-State Circuits*, vol. SC-15, pp. 331–340, 1980.
75. N. Ballay and B. Baylac, "Analytical modelling of depletion-mode MOSFET with short- and narrow-channel effects," *IEE Proceedings*, vol. 127, part I, pp. 225–230, December 1981.
76. M. R. Wordeman and R. H. Dennard, "Threshold voltage characteristics of depletion-mode MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 1025–1030, 1981.
77. U. Ohno and Y. Okuto, "Electron mobility in n-channel depletion type MOS transistors," *IEEE Transactions on Electron Devices*, vol. ED-29, pp. 190–194, 1982.

78. R. Ratnam and A. B. Bhattacharyya, "Accumulation-punchthrough model of operation of buried-channel MOSFETs," *IEEE Electron Device Letters*, EDL-3, pp. 203–204, 1982.
79. T. Yamaguchi and S. Morimoto, "Analytical model and characteristics of small geometry buried-channel depletion MOSFETs," *IEEE Journal of Solid-State Circuits*, vol. SC-18, pp. 784–793, 1983.
80. S. Haque-Ahmed and C. A. T. Salama, "Depletion mode MOSFET modelling for CAD," *IEE Proceedings*, vol. 130, part I, pp. 281–286, 1983.
81. J. S. T. Huang, J. W. Schrankler, and J. S. Kueng, "Short-channel threshold model for buried-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1889–1895, 1984.
82. F. M. Klaassen and W. Hes, "Compensated MOSFET devices," *Solid-State Electronics*, vol. 28, pp. 359–373, 1985.
83. D. Ma, "A physical and SPICE-compatible model for the MOS depletion device," *IEEE Transactions on Computer-Aided Design*, vol. CAD-4, pp. 349–356, 1985.
84. M.-W. Chiang, J. Choma, Jr., and C. Kao, "A simulation method to completely model the various transistor I-V operational modes of long channel depletion MOSFETs," *IEEE Transactions on Computer-Aided Design*, vol. CAD-4, pp. 322–328, 1985.
85. G. J. Hu and R. H. Bruce, "Design tradeoffs between surface and buried channel FETs," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 584–588, 1985.
86. C. Turchetti and G. Masetti, "Analysis of the depletion-mode MOSFET including diffusion and drift currents," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 773–782, 1985.
87. C.-Y. Wu and K. C. Hsu, "Mobility models for the I-V characteristics of buried-channel MOSFETs," *Solid-State Electronics*, vol. 28, pp. 917–923, 1985.
88. K. C.-K. Weng, P. Yang, and J.-H. Chern, "A predictor/CAD model for buried-channel MOS transistors," *IEEE Transactions on Computer-Aided Design*, vol. CAD-6, pp. 4–16, 1987.
89. S. Karmalkar and K. N. Bhat, "The correct equivalent box representation for the buried layer of BC MOSFETs in terms of the implantation parameters," *IEEE Electron Device Letters*, vol. EDL-8, pp. 457–459, 1987.
90. Z. Yu and X. Zhao, "A semi-analytical approach to the evaluation of threshold voltage in depletion MOS's with nonuniformly doped substrates," *IEEE Transactions on Electron Devices*, vol. 35, p. 993, 1988.
91. S. W. Tarasewicz and C. A. T. Salama, "Threshold voltage characteristics of ion-implanted depletion MOSFETs," *Solid-State Electronics*, vol. 31, pp. 1441–1446, 1988.
92. C. D. Parikh and J. Vasi, "Modeling of a depletion-mode MOSFET," *Solid-State Electronics*, vol. 30, pp. 699–703, 1987. See also comments by C. Turchetti and G. Masetti, *Solid-State Electronics*, vol. 31, pp. 1747–1748, 1988, and authors' response, *Solid-State Electronics*, vol. 31, p. 1749.
93. M. J. Van de Tol and S. G. Chamberlain, "Buried-channel MOSFET model for SPICE," *IEEE Transactions on Computer-Aided Design*, vol. CAD-10, pp. 1015–1035, 1991.
94. Y. Yin and J. A. Cooper, Jr., "Simple equations for the electrostatic potential in buried-channel MOS devices," *IEEE Transactions on Electron Devices*, vol. 39, pp. 1770–1772, 1992.
95. B.-J. Moon, C.-K. Park, K.-M. Rho, K. Lee, M. Shur, and T. A. Fjeldly, "Analytical model for p-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 38, pp. 2632–2646, 1991.
96. C. Bulucea and D. Kerr, "Threshold voltage control in buried-channel MOSFETs," *Solid-State Electronics*, vol. 41, pp. 1345–1354, 1997.
97. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge University Press, Cambridge, 1998.
98. A. S. Silburt, R. C. Foss, and W. F. Petrie, "An efficient MOS transistor model for computer-aided design," *IEEE Transactions on Computer-Aided Design*, vol. CAD-3, no. 1, pp. 104–111, 1984.
99. S.-W. Lee and R. C. Rennick, "A compact IGFET model-ASIM," *IEEE Transactions on Computer-Aided Design*, vol. 7, no. 9, pp. 952–975, 1988.
100. Y. Cheng, M. Chan, K. Hui, M.-C. Jeng, Z. Liu, J. Huang, K. Chen, J. Chen, R. Tu, P.-K. Ko and C. Hu, *BSIM 3v3 Manual* (Final Version), University of California, Berkeley, Electronics Research Laboratory Memorandum No. UCB/ERL M97/2, 1997.
101. BSIM website. Available at <http://www-device.eecs.berkeley.edu/~bsim3/>. Accessed Oct. 1, 2009.
102. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed., Oxford University Press, Oxford, 1999.

103. G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, 2006.
104. R. Rios, W.-K. Shih, A. Shah, S. Mudanai, P. Packan, T. Sandford, and K. Mistry, "A three-transistor threshold voltage model for halo processes," *Proceedings of the IEEE International Electron Devices Meeting*, December 2002, pp. 113–116.
105. S. Mudanai, R. Rios, W.-K. Shih, P. Packan, and S.-W. Lee, "Halo doping: Physical effects and compact modeling," *Proceedings of the Nanotechnology Conference*, pp. 644–647, May 2006.
106. PSP website. Available at <http://pspmodel.asu.edu/>. Accessed Oct. 1, 2009.
107. T. B. Hook, J. Brown, P. Cottrell, E. Adler, D. Hoyniak, J. Johnson, and R. Mann, "Lateral ion implant straggle and mask proximity effect," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1946–1951, 2003.
108. Y.-M. Sheu, K.-W. Su, S.-J. Yang, H.-T. Chen, C.-C. Wang, M.-J. Chen, and S. Liu, "Modeling well edge proximity effect on highly-scaled MOSFETs," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 831–834, September 2005.
109. J. Watts, K.-W. Su, and M. Basel, "Netlisting and modeling well-proximity effects," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2179–2186, 2006.
110. M. Basel, J. Xi, J. Watts, P. Humphries, K. Su, S. Moinian, R. Thoma, and K. Onozawa, "Guidelines for extracting well proximity effect instance parameters," CMC website. Available at <http://www.geia.org/index.asp?bid=1333>.
111. P. G. Drennan, M. L. Kniffin, and D. R. Locascio, "Implications of proximity effects for analog design," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 169–176, September 2006.
112. Book's website: www.oup.com/tsividis_mcandrew
113. V. Chan, K. Rim, M. Jeong, S. Yang, R. Malik, Y. W. Teh, M. Yang, and Q. Ouyang, "Strain for CMOS performance improvement," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 667–674, September 2005.
114. P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, et al., "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 657–660, December 2004.
115. G. Scott, J. Lutze, M. Rubin, F. Nouri, and M. Manley, "NMOS drive current reduction caused by transistor layout and trench isolation induced stress," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 827–830, December 1999.
116. K.-W. Su, Y.-M. Sheu, C.-K. Lin, S.-J. Yang, W.-J. Liang, X. Xi, C.-S. Chiang, J.-K. Her, Y.-T. Chia, C. H. Diaz, and C. Hu, "A scaleable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 245–248, September 2003.
117. J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology: Fundamentals, Practice, and Modeling*, Prentice Hall, Englewood Cliffs, N. J., 2001.
118. C.-L. Huang, H. Soleimani, G. Grula, N. D. Arora, and D. Antoniadis, "Isolation process dependence of channel mobility in thin-film SOI devices," *IEEE Electron Device Letters*, vol. 17, pp. 291–293, 1996.
119. P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, M. leong, and M. Hargrove, "Modeling line edge roughness effects in sub-100 nanometer gate length devices," *Proceedings of the IEEE International Conference on Simulation of Semiconductor Processes and Devices*, pp. 131–134, September 2000.
120. A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decanometer MOSFETs introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1254–1260, 2003.
121. K. Nishinohara, N. Shigyo, and T. Wada, "Effects of microscopic fluctuations in dopant distributions on MOSFET threshold voltage," *IEEE Transactions on Electron Devices*, vol. 39, no. 3, pp. 634–639, 1992.
122. H.-S. Wong and Y. Taur, "Three-dimensional "atomistic" simulation of discrete random dopant distribution effects in sub-0.1 μm MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 705–708, December 1993.

123. P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 9, pp. 1960–1971, 1998.
124. A. Asenov, "Every atom counts," *IEE Electronics Systems and Software*, vol. 2, no. 6, pp. 26–31, 2004–2005.
125. Y. Ye, F. Liu, S. Nassif, and Y. Cao, "Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness," *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 900–905, June 2008.
126. A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Transactions on Electron Devices*, vol. 49, no. 1, pp. 112–119, 2002.
127. P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, 2005.
128. J. A. Power, B. Donellan, A. Mathewson, and W. A. Lane, "Relating statistical MOSFET model parameter variabilities to IC manufacturing process fluctuations enabling realist worst-case design," *IEEE Transactions on Semiconductor Manufacturing*, vol. 7, no. 3, pp. 306–318, 1994.
129. P. Cox, P. Yang, S. S. Mahant-Shetti, and P. Chatterjee, "Statistical modeling for efficient parametric yield estimation of MOS VLSI circuits," *IEEE Transactions on Electron Devices*, vol. 32, no. 2, pp. 471–478, 1985.
130. M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 5, pp. 1433–1440, 1989.
131. K. R. Lakshminarayana, R. A. Hadaway, and M. A. Copeland, "Characterization and modeling of mismatch in MOS transistors for precision analog design," *IEEE Journal of Solid-State Circuits*, vol. CS-21, no. 6, pp. 1057–1066, 1986.
132. M. Bucher, *Analytical MOS Transistor Modelling for Analog Circuit Simulation*, Ph.D. thesis, no. 2114, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 1999.
133. T. Serrano-Gotarredona and B. Linares-Barranco, "A new five-parameter MOS transistor mismatch model," *IEEE Electron Device Letters*, vol. 21, no. 1, pp. 37–39, 2000.
134. H. Tuinhout, "Characterisation of systematic MOSFET transconductance mismatch," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 131–136, March 2000.
135. J. A. Croon, H. P. Tuinhout, R. Difrenza, J. Knol, A. J. Moonen, S. Decoutre, H. E. Maes, and W. Sansen, "A comparison of extraction techniques for threshold voltage mismatch," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 235–240, April 2002.
136. T. Ewert, H. Tuinhout, N. Wils, and J. Olsson, "Design and implementation of an ultra high precision parametric mismatch measurement system," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 149–154, April 2005.
137. J. B. Johnson, T. B. Hook, and Y.-M. Lee, "Analysis and modeling of threshold voltage mismatch for CMOS at 65nm and beyond," *IEEE Electron Device Letters*, vol. 29, no. 7, pp. 802–804, 2008.
138. C. C. McAndrew, "Efficient statistical modeling for circuit simulation," in *Design of Systems on a Chip: Devices and Components*, R. Reis and J. Jess (editors), Kluwer Academic, New York, pp. 97–122, 2004.
139. P. Drennan and C. C. McAndrew, "Understanding MOSFET mismatch for analog design," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 3, pp. 450–456, 2003.
140. J. Krick, "Statistical transistor SPICE modeling in advanced CMOS technologies," paper presented at Compact Modeling of Variability Workshop, IEEE/ACM International Conference on Computer-Aided Design, November 2008.

Problems

- 9.1.** Show that, with $V_I(V_{SB})$ as given by (9.3.24), both $V_I(V_{SB})$ and dV_I/dV_{SB} are continuous at $V_{SB} = V_I$, and⁵⁸

$$\left. \frac{dV_I}{dV_{SB}} \right|_{V_{SB}=V_I} = \frac{\epsilon_s}{\epsilon_{ox}} \frac{t_{ox}}{d_I} \approx 3 \frac{t_{ox}}{d_I}$$

- 9.2.** The channel of an *n*MOS transistor with $t_{ox} = 2.5$ nm and a *p* substrate with $N_A = 1 \times 10^{16} \text{ cm}^{-3}$ is implanted with acceptors using an effective dose of $2.5 \times 10^{12} \text{ cm}^{-2}$. The value of d_I is 75 nm. $V_{FB1} = -0.7$ V and $\phi_{01} = 0.75$ V. Plot $V_T(V_{SB})$ vs. V_{SB} for V_{SB} between 0 and 2 V.
- 9.3.** Prove (9.3.15) using basic electrostatics (Appendix A). Plot the charge density, the electric field, and the potential as a function of y . Assume that the potential is zero deep in the substrate, and that the total potential drop across the entire depletion region is $\phi_{01} + V_{CB}$. Find the depth of the depletion region in the substrate, and from that find Q'_{B2} .
- 9.4.** Propose values for the parameters γ , L_{PE} , V_{FB} , and ϕ_0 needed in (9.4.4) and (9.4.5) to model the measured data in Fig. 9.18.
- 9.5.** Consider a halo-doped MOSFET as consisting of three transistors connected in series, as in Fig. 9.15. Assume that the weak-inversion current in a MOS transistor is proportional to $(e^{(V_{GS}-V_T)(n\phi_b)})/(1 - e^{V_{DS}/\phi_b})/L$ (Sec. 4.8). Consider a single-section equivalent model, which is supposed to provide the same current if the terminal voltages are the same. Use appropriate biases, length, and threshold voltage for each section and also for the single-section equivalent model. Equating the current in each section to each other and to the current of the single-section model, show that the threshold voltage of the composite transistor is given by (9.4.6).
- 9.6.** Why is it better to consider statistical variations in V_{T0} , W , and L to be in an absolute rather than relative sense? (Consider what a percentage variation would imply).
- 9.7.** Prove (9.7.6). Start with (9.7.2) for two devices of different widths and lengths, and use the definition in the footnote to (9.7.6). You may assume that both the global and local variations are normally distributed, and use the fact that $E(xy) = 0$ if x and y are uncorrelated and have zero means. The global variations are, by definition, common to both devices, and the local variations are independent between devices and independent of the global variation.
- 9.8.** Consider an arbitrary electrical performance E . Assume that variations in E are caused by variations in threshold voltage V_{T0} and variations in the transistor gain factor $\beta = \mu C'_{ox} W/L$, and that these are statistically independent quantities. Linearize E with respect to V_{T0} and β , and use propagation of variance (if $y = a_1x_1 + a_2x_2$ then $\sigma_y^2 = a_1^2 \sigma_{x_1}^2 + a_2^2 \sigma_{x_2}^2$) to determine the variance of E in terms of the variance of V_{T0} and β . Now assume that all of the statistical variation in both V_{T0} and β is correlated and comes from an underlying variation in t_{ox} . Linearize E with respect to t_{ox} , through the intermediate variables V_{T0} and β , and calculate the variance in E . How does this differ from the result obtained if it is assumed that V_{T0} and β are statistically independent? What does this say about the predictions of statistical variability when correlations between parameters are ignored?

CHAPTER**10**

Modeling for Circuit Simulation



10.1 Introduction

A major application of MOSFET models is in circuit simulation for integrated circuit (IC) design. Integrated circuit fabrication is a slow and extremely expensive process. It is therefore very important to verify that a design is functional before it is sent to manufacturing. Increasingly complex computer-aided design (CAD) systems are used for the implementation and verification of IC designs, and a significant component of these systems is circuit simulation. Many different types of simulators are available. Among these are standard SPICE-like simulators¹ that perform dc, small-signal (often called “ac”), transient, and noise analysis of all types of circuits, and the so-called fast MOS simulators that can do accurate transient analysis of very large digital circuits more efficiently than conventional SPICE-like simulators. Also, there exist RF simulators that do harmonic balance analysis of nonlinear circuits. Other specialty simulators and analysis methods exist, and over the past decades there has been a large effort by researchers and CAD engineers, at both universities and in the industry, toward the development of computer simulation algorithms and programs. But these simulators are not useful without accurate representations of the basic components (MOSFETs, bipolar transistors, resistors, capacitors, etc.) that are assembled to create the ICs. A key use of MOSFET models is thus for circuit simulation.

Development of a practical, industrial-strength model is an enormous undertaking; it involves much more than the analysis of the physics that underlies device behavior and the derivation of a set of equations that embody this behavior. Models for CAD are software programs, with complex models comprising over 20,000 lines of code, and they require proper software development practices such as source code and revision control, quality assurance and regression testing (a way to extensively

test the software code), documentation, release procedures, and ongoing enhancement and maintenance. It is beyond the scope of this book to go into details of “hard-core” model development. The number of people actively involved in model development is small, and their needs can be addressed by professional publications and conferences. However, a basic understanding of MOSFET modeling for CAD is necessary for one to be an effective user of MOSFET models. This chapter describes some practical aspects of device modeling for circuit simulation with emphasis on MOSFETs.

10.2 Types of Models

Models for the simulation of semiconductor devices come in several different forms. The most commonly encountered forms are listed in the next subsection.

10.2.1 Models for Device Analysis and Design

These are also known as TCAD (technology computer-aided design) models, where “technology” refers to fabrication technology. They comprise three main classes, listed in descending order of physical content:

- *Monte Carlo*. This is considered to be the “first principles” approach to semiconductor device modeling.^{2,3} Monte Carlo models accept as inputs information about the semiconductor band structure, scattering mechanisms, and other physical effects, and they perform probabilistic particle simulations to model carrier transport. We have seen an example of such a simulation in Sec. 5.9. Due to the detailed nature of these techniques, they are computationally slow and are used for semiconductor device studies rather than circuit simulation.
- *Moment methods*. These refer to a way to solve the Boltzmann transport equation (BTE), which describes the statistical distribution of carrier concentration, momentum, and energy in a semiconductor device.⁴ These methods involve partial differential equations that are discretized in space (and time for transient simulation). The physics enters via momentum and energy relaxation time models, and the device structure enters via doping profiles. The resulting computationally intensive models are used for device simulation.
- *Drift-diffusion*. This is essentially a simplified form of the BTE moment method, with the physics entering primarily through mobility models.⁵ The solution involves spatial discretization of partial differential equations, of Poisson’s equation, and of the carrier continuity equations for electrons and holes. Although primarily used for device simulation, modern drift-diffusion simulation programs run sufficiently fast to be used for simulation of small circuits, with up to several transistors.

Scattering (or “one flux”) models are also being investigated (Sec. 5.9) but are not in common use at the time of writing.

Although TCAD models are the most accurate, their large memory usage and long computation times make them impractical for use in circuit design. Nevertheless,

TCAD simulations are invaluable for observing and understanding the details of device operation, for helping optimize device designs and performance, and as a mechanism for verifying modeling of a specific effect. In a real device, all nonideal effects occur simultaneously and can be difficult to infer separately from measured data. In TCAD simulations, many individual effects can be enabled or disabled individually, for example, velocity saturation; this can help the development of models for each individual effect. The TCAD models are also very useful for determining parameter values for models used for circuit design prior to first-time manufacture in a new technology. These early models, often called “Rev 0” models,⁶ can be used for initial circuit design and exploration in the new technology.

10.2.2 Device Models for Circuit Simulation

For the simulation of circuits containing many transistors, one needs models that are computationally much faster than the preceding TCAD techniques. These faster models are discussed here, again in descending order of physical content.

Physical Compact Models These are the models of the type presented in this book; they are based on device physics and parameters that have a physical significance, with appropriate simplification to yield modeling equations. The key to physical compact modeling is to derive relations that are both accurate and computationally fast to evaluate. For analog and RF circuit design, physical compact models are the preferred choice. There are several reasons for this, the most important being accuracy. There is significant cost, in terms of design effort and speed, if the simulations predict that a circuit will function properly, yet the manufactured chip fails to function properly or to meet specifications. Hence simulation accuracy, which is predicated on the accuracy of the compact models used in it, is paramount.

Physical models take a long time to develop. Extra development time may be needed each time a fabrication technology changes significantly; for example, many of the small-dimension phenomena discussed in Chap. 5 were revealed as technology matured, and this necessitated new research and development to make adequate models possible. Once a good physical model is developed, though, its advantages, in comparison to other models for circuit CAD (see below) are very significant. To begin with, the parameters in a physical model have a physical significance. It is thus possible to relate the results provided by the model to the physical details of the transistor, a capability that is very important in circuit design. Notably, the effects of the geometry of the device (W and L values), and the effects of temperature changes, can be predicted for a given process a designer is working with. Within limits, it is even possible to predict what would happen if the process parameters were changed. This latter feature is of particular significance in statistical modeling. If the basic parameters of a model are the parameters of the process (e.g., substrate doping concentration, oxide thickness), one can use information about the systematic and random errors of these parameters, as well as such errors in the geometrical dimensions, to predict the statistics of the I - V characteristics or other device behavior (e.g., the transistor f_T). Thus, physical models can be used for statistical analysis,^{7–16} including correlations between parameters, mismatch, and global statistical variation. One can

use this information to predict ranges of expected performance variation and therefore yield for given specifications. Statistical modeling can then be used as a guide for redesign in order to improve yield. Statistical modeling can also be used to study the mismatch between identically laid out devices (Sec. 9.7); such studies are very important for analog circuit design.^{17–22}

It should be stressed that some of the advantages of physical models just mentioned might not be valid if the model parameters have not been chosen correctly. This issue will be discussed later in this chapter.

Good models are usually complex because they have to carefully take into account several phenomena that the simpler models ignore. By the time all regions of operation are included, one can easily end up with models with hundreds of parameters. These can provide accuracy at the expense of complexity. Note however that satisfactory accuracy and a large number of parameters do not always mean that the physics of the device has been modeled correctly. There exist models that are based on faulty premises but which, just because they contain many parameters, can give adequate accuracy after extensive empirical adjustment of their values. Such models, though, do not normally have predictive power.

Physical compact models also enable other types of simulations and analyses that are difficult or impossible with other fast approaches (discussion follows). Modeling of behavior over temperature is important and is included in physical compact models. Modeling of electrothermal effects is becoming increasingly important, and this is conceptually straightforward to implement in a compact model by self-consistently solving coupled electrical and thermal model equations.²³

Broadly, physical compact models for MOSFETs can be further separated into three types,²⁴ listed in the order in which they became generally available for circuit simulation:

- Threshold voltage-based models (in which the threshold voltage is generally referenced to the source), e.g., BSIM3²⁵ and BSIM4.²⁶ These models are characterized by an inversion charge that is proportional to $V_{GS} - V_T(V_{SB})$ and by a strong-inversion nonsaturation drain current that varies roughly as $[V_{GS} - V_T(V_{SB}) - \alpha V_{DS}/2]V_{DS}$, as in (4.7.24a). Such relations are then modified and/or expanded to cover other ranges.
- Models based on inversion charge as the primary quantity, e.g., ACM²⁷ and EKV.²⁸ These models are based on an inversion charge relation with a form similar to (4.4.21) and drain current that includes a drift component proportional to $Q'_{i0}^2 - Q'_{iL}^2$ and a diffusion component proportional to $Q'_{iL} - Q'_{i0}$, as in (4.4.17).
- Surface potential-based models, e.g., HiSIM2²⁹ and PSP.³⁰ These are characterized by both inversion charge and drain current being computed directly from the surface potential ψ_s (Sec. 4.3).

We will see examples of each type of model in Sec. 10.9.

As already mentioned, it can take years to develop an appropriate and accurate physical compact model for a specific device. Not all proposed device structures end up in commercial manufacturing technologies, and investment in an effort to develop

a compact model is not warranted until there is some expectation that the device will be used in real designs. Thus, other classes of circuit simulation models also find some use. We discuss these classes briefly before devoting the rest of the chapter to physical compact models.

Empirical Compact Models Rather than being based on physical analysis, these models are based on functions (e.g., hyperbolic tangent), selected empirically so that the shape of the function matches experimental data. Parameters of empirical models are therefore often fitting coefficients; however, some empirical models involve a first-order physical significance for some of their parameters.

There is a fuzzy boundary between physical and empirical compact models, as most physical models have some empirical content, such as linear or quadratic polynomials for temperature and some voltage dependencies, and most empirical models have some physical content, such as first-order scaling with width and length. Despite our emphasis on physical models, empirical models can be useful for some applications. For example, reasonably accurate simulation of the propagation delay of digital gates can be done with quite crude models. One of these is the “alpha-power law”[†] model³¹

$$I_{DS} = \begin{cases} 0, & V_{GS} \leq V_T \\ \frac{P_C W}{P_V L} (V_{GS} - V_T)^{\alpha'/2} V_{DS}, & V_{DS} < V'_{DS}, \quad V'_{DS} = P_V (V_{GS} - V_T)^{\alpha'/2} \\ \frac{P_C W}{L} (V_{GS} - V_T)^\alpha, & V_{DS} \geq V'_{DS} \end{cases} \quad (10.2.1)$$

where P_C and P_V are empirical parameters. The equation for the current is linear, rather than quadratic, in V_{DS} for strong-inversion nonsaturation operation and is not proportional to $(V_{GS} - V_T)^2$ in saturation. The model was derived purely by inspection of data from the observation that real devices deviated from the ideal quadratic dependence on gate overdrive in saturation³² (a fact that, as we know, comes from effective mobility dependence on the gate field, as well as from velocity saturation). It turns out that this model enables analytical calculation of reasonably good approximations to the gate delays of digital circuits and so has been widely used for this purpose. Interestingly, later evaluation gave this model a firmer physical basis.³³ On the other hand, this model would be of little use for most analog circuit design purposes. Empirical modeling should be used only as a last resort, when attempts to efficiently model physical effects correctly fail.

[†] The name is due to the fact that, in the original reference, the letter “alpha” is used instead of our α in the equations. We use α in order to avoid confusion with an unrelated parameter that we have denoted by “alpha” in conjunction with the source-referenced simplified strong-inversion model in Sec. 4.7.

“Black Box” Models There are many sophisticated numerical packages available today that will take as inputs arbitrary sets of data, analyze them, and then automatically generate some form of mathematical model for them. For example, one can feed such a program with resistance values measured over length L , width W , temperature T , and bias V ; the program will then investigate many model forms and return some model $R(L,W,T,V)$ that best fits the supplied data. Such black box models are occasionally used for device modeling. A reason they are not used more widely is that the equation the software comes up with in such models can be expected to be accurate only within the range of variations of bias, geometry, and temperature used to generate the model; outside this range, this equation may produce very wrong behavior. There is continued research on automated generation of black box models, and it is expected that the usefulness and accuracy of these algorithms will continue to improve. However, these approaches have so far been used primarily for complex interconnect structures and for system or block-level modeling, but not for MOSFET modeling; it is unlikely that they will be used for MOSFET modeling any time soon.

Table Lookup Models Rather than try to model data, table lookup models directly take data, over regular sweeps of bias voltages, and use some form of interpolation to calculate the device behavior for other bias values.³⁴ Data tables and interpolation algorithms therefore replace “models” per se. Table lookup models are useful for early evaluation of the circuit performance of experimental device structures. The data can come from TCAD simulations or measurements; some mixed device-circuit simulators have the capability to automatically generate table lookup models from numerical device simulations and to then use these table models for circuit evaluation.³⁵ Table lookup models for MOSFETs are also used in fast MOS simulators to improve simulation speed (for many circuit simulations, compact model evaluation time is the dominant component of overall simulation time). The data tables are generated internally to the simulator, based on evaluation of a compact model.

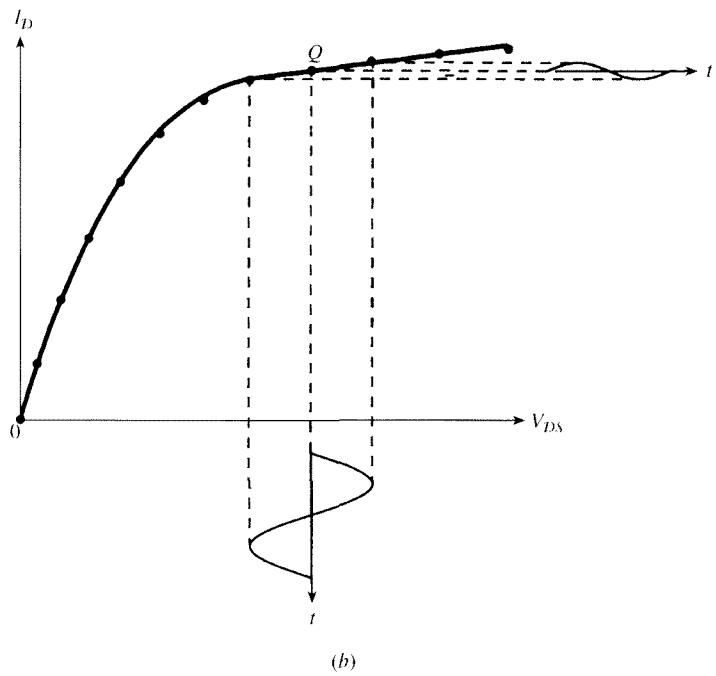
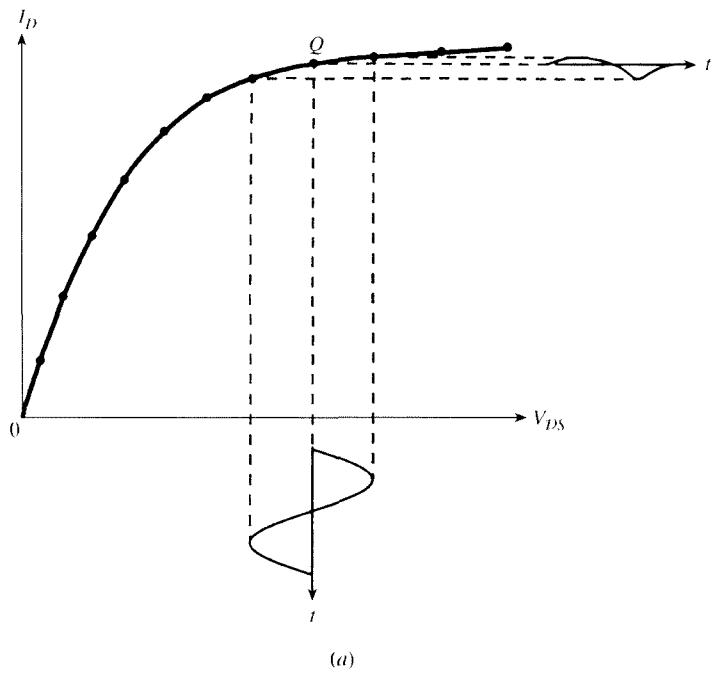
Table lookup models have some drawbacks, which is why they are not used universally. For multiterminal devices, large and complex multidimensional data arrays are needed. These are usually available for a limited range of terminal bias values, so robust extrapolation techniques are required outside of this range. Often there are artifacts, such as wiggles in I - V characteristics and unphysical negative conductances, resulting from the numerical interpolation procedures. Because these procedures are generally based on some type of low-order polynomial interpolation, they can be inaccurate for modeling regions where device characteristics vary exponentially with bias, such as in weak inversion for MOSFETs. Improved interpolation procedures for MOS transistors can help overcome these issues,^{35,36} but physical compact models still prevail over table models. Charge tables can be generated from measured or simulated terminal capacitances,^{37,38} but the procedure is not simple and robust. More important, the tables are generated based on low-frequency characterization and cannot capture complexities of behavior at high frequencies. In addition, although some types of scalability for table lookup models have been proposed, in general a data table has to be populated for each device geometry used in a circuit,

for each temperature for which data are available. Also, table lookup models provide little or no capability for modeling statistical variations, for modeling noise, for being easily retargeted when a manufacturing process shifts, or for providing extrapolations for predictions of future fabrication technologies.

10.3 Attributes of Good Compact Models

Here is a summary of requirements that a CAD model must ideally meet to be suitable for analog and mixed analog/digital circuit design:³⁹

1. The model should meet common requirements for digital circuits, such as reasonable accuracy of I - V characteristics, charges, extrinsic parasitics, and leakage currents.
2. The model should provide continuous results. Not only should currents and charges be continuous with respect to each terminal voltage, but also their *derivatives* should be continuous as well. This is important for numerical convergence; circuit simulators commonly solve the nonlinear Kirchhoff current law (KCL) equations using some form of the Newton-Raphson algorithm, which is based on linearization and requires “smooth” models to work effectively. Also, smooth variation of small-signal conductances requires continuity in the second-order derivatives of the current. However, *even higher-order derivatives must be continuous* for predicting accurately the nuances of the I - V characteristics; this is important, for example, in the evaluation of distortion.⁴⁰⁻⁴⁴ A graphical illustration³⁹ of how an otherwise accurate model can fail in predicting distortion is shown in Fig. 10.1. Distortion modeling places the most stringent requirements on a model.⁴⁰⁻⁴⁴
3. The model formulation and the model parameters should be as physical as possible, and the results should make physical sense. The model should not exhibit unphysical behavior, such as generating power, or predict conductances or capacitances with wrong algebraic signs.
4. The model should give good results even when the device operates non-quasi-statically, or at least it should degrade gracefully for such operation, as frequency is increased. If the model cannot be non-quasi-static, it should at least be a complete quasi-static model (Chap. 6), which includes the effect of each terminal voltage on the charge passing through each of the terminals.
5. It should give accurate predictions for white and flicker noise in all regions of operation, and for the correlation between gate and drain noise.
6. It should include all intrinsic effects and extrinsic parasitic elements that affect circuit performance.
7. It should meet all requirements over large bias ranges, including $V_{SB} \neq 0$, and encompassing the weak-, moderate-, and strong-inversion regions.
8. It should meet all requirements over the temperature range of interest.
9. It should meet all the preceding criteria for any combination of channel width and length values, from the minimum specified upward.

**FIGURE 10.1**

The I_D - V_{DS} curve as (a) measured and (b) modeled.³⁹ For a sinusoidal voltage variation, the shape of the corresponding current waveform in (b) is severely in error, although the error in modeling I_D is very small throughout (© 1994 by IEEE).

10. The user should have to specify only the layout parameters for each device, the temperature, and one set of model parameters valid for all devices of the same type independent of dimensions.
11. The model should not cause numerical problems in simulators for bias values outside of the normal operating range (these are encountered during the numerical iteration procedures used in simulators).
12. The model should provide a flag when it is used outside its limits of validity. For example, if the model is quasi-static and one attempts to use it, say, above the unity-gain frequency of the device, a warning should be given to the user that the result may be inaccurate.
13. The model should have as few parameters as possible (but just enough), and those parameters should be linked as strongly as possible to ones related to the device structure and fabrication process (e.g., oxide thickness, substrate doping, junction depth). This allows meaningful statistical and worst-case simulations and yield predictions, as well as studies of mismatches between devices (which are important in analog design). An emphasis on process and geometrical parameters is also necessary to make the model predictive and to make it easily adaptable in the face of process and layout rule changes. Empirical parameters without physical meaning should be avoided as much as possible.
14. The model should encompass operation in depletion and accumulation in a smooth and physical manner; accuracy in these regions is important in modeling MOS varactors.
15. The model should be computationally efficient and numerically robust.
16. It should be symmetric for symmetric devices. For example, interchanging the drain and source potential for such devices should reverse the direction of the current, but should leave its magnitude unchanged.
17. It should allow for the modeling of asymmetric devices; for example, it should allow the freedom of specifying junction capacitance parameters separately for the source and drain regions.⁴⁵ This is needed not only because a variety of layouts are used by designers, but also because in some technologies (notably high-voltage ones) different fabrication steps are used for the source and drain regions.
18. The model should be linked to an efficient parameter extraction method. The number of required test structures and measurements for parameter extraction should be as small as possible.
19. It should pass the benchmark tests described in Appendix K.

This is a large list of requirements, some of which can be quite difficult to meet. Modern surface potential-based MOSFET models are closer to meeting these requirements than are the source-referenced MOSFET models based on strong inversion that have been the most common in the past decades.

10.4 Model Formulation

The criteria listed in the previous section cover what a user should expect of a good model, and what a model developer should strive for. We now discuss several

issues involved in creating a model that satisfies these criteria. This section should be of value not only to professional model developers, but also to students doing modeling projects. It should even be of value to model users, who can benefit from the background provided to interpret the behavior of the models they use and to be able to talk intelligently with modeling teams responsible for model creation and maintenance.

A physical compact model is more than just a set of equations that define some $I(V)$ and $Q(V)$ relations for a particular device. The fundamental structure of a model needs to work properly with the analysis algorithms implemented in a circuit simulator, which can impose significant restrictions on how a model should and should not behave. For example, there should be no discontinuities in the basic modeling equations for a model. There are also first-order physical behaviors that a model should be constrained to follow. For example, with zero bias applied to all terminals of a device there should be no current flowing in the device. As straightforward as these requirements sound, it is easy to get lost in the details when defining a complex model and to inadvertently introduce some undesirable or unphysical characteristic. This section discusses some basic principles that should be followed when developing a practical compact model.

10.4.1 General Considerations and Choices

Several considerations and choices confront a model developer. Some of these are now discussed.

Type of Model The only type of model that can meet the previous set of requirements is a physical model. The advantages of physical models have already been discussed.

Choice of Parameters Careful thought must be given to the choice of primary parameters for the model. For example, V_{T0} should not be such a parameter, because this would hide the dependence of threshold on oxide thickness, which is important in statistical simulation (Sec. 9.7);⁴⁵ if, instead ϕ_{MS} and Q'_o are chosen as primary parameters, the dependence on oxide thickness will be there, as seen from (2.2.6) and (4.7.19). Ideally, only process, temperature, and geometry information should have to be supplied to the model, and every effort should be extended to get as close to this goal as possible. The parameters should not be redundant; each should be responsible for one or more effects, which could not have been predicted without it. The parameter set should be chosen with parameter extraction in mind. In fact, parameter extraction should be kept in mind from the very beginning of model development.

Choice of Reference It should be considered, right at the outset, whether the model will be source-referenced or body-referenced. The pros and cons of each approach have been discussed in Sec. 4.10. In principle, the two approaches should be equivalent, because a body-referenced model can be converted to a source-referenced model by using $V_{GB} = V_{GS} + V_{SB}$ and $V_{DB} = V_{DS} + V_{SB}$. However, the practices adopted in

most source-referenced models are responsible for a number of problems at $V_{DS} = 0$.⁴⁵ Among these are asymmetries of the source and drain capacitances at that point, a discontinuous slope in g_{sd} vs. V_{DS} plots (see Appendix K) and failure to predict the distortion of transistors biased at $V_{DS} = 0$.⁴⁶ Given this history, if a source-referenced approach is chosen, the model developer must make sure that such problems do not occur. Body-referenced models generally do not exhibit such problems. On the other hand, the implementation of certain effects, such as velocity saturation, is easier in source-referenced models, and can be difficult for a body-referenced model.

Choice of Modeling Expressions The ideal physical model should be based on general expressions, derived from physics (not interpolation), that are valid in all regions of operation and are explicit functions of the terminal voltages.[†] Such a model remains an elusive goal at the time of this writing. General models (Sec. 4.3) do make general expressions possible; however, these are in terms of surface potentials (not terminal voltages). Very accurate evaluation of surface potential is needed, and this involves an implicit equation in principle (Sec. 4.3). Special attention must be paid to developing appropriate numerical techniques for such calculations. By using good approximations as initial “guesses,” surface potentials can be evaluated in only a few iterations. Several examples of CAD models based on general expressions can be found in Refs. 49–52. The use of general expressions, combined with appropriate evaluation of surface potentials, ensures continuity of derivatives, which is important for several reasons, as already explained in Sec. 10.3. Real devices show smooth and continuous behavior and do not exhibit discontinuities.

The Use of Smoothing Functions Because of difficulties in implementing a model based on general expressions, a popular alternative has been to base the model on separate expressions for different regions of operation. For example, this is sometimes preferred because expressions for individual regions of inversion can be simple (at least for weak and strong inversion) and give the current explicitly in terms of the terminal voltages; the evaluation of surface potentials is not required in this case. Also, often separate expressions are used for nonsaturation and saturation. However, this approach has been responsible for numerous discontinuities in the derivatives of the current with respect to the bias voltages, which create problems with numerical convergence and cause serious errors in the prediction of small-signal parameters and distortion. To avoid such problems, if separate expressions must be used, they should be combined into one single expression with the help of appropriate “smoothing functions,” also sometimes called “limiting functions.” The resulting single expression can be used in all regions of operation, but the smoothing functions are such that they “reveal” particular expressions in particular regions, while ensuring continuity at the boundaries between regions. Historically, such approaches base a model on a physical derivation that is applicable in one region of operation, generally strong-inversion nonsaturation, and then introduce mathematical manipulations that make

[†] Current-control alternatives have also been proposed.^{28,47,48}

the model approximately correct in other regions of operation.^{53–57} The resulting models are known as *interpolation* models.

Consider as an example the simplified source-referenced strong-inversion model of Sec. 4.7.3. Equations (4.7.24) describe an abrupt transition from nonsaturation to saturation operation. Notice that the model can be equivalently expressed as

$$I_{DS} = \frac{W}{L} \mu C'_ox \left[(V_{GS} - V_T) - \frac{\alpha}{2} V_{DS,\text{eff}} \right] V_{DS,\text{eff}} \quad (10.4.1)$$

where

$$V_{DS,\text{eff}} = \begin{cases} V_{DS}, & V_{DS} \leq V'_{DS} \\ V'_{DS}, & V_{DS} > V'_{DS} \end{cases} \quad (10.4.2)$$

Rather than having an abrupt transition at $V_{DS} = V'_{DS}$ as in (10.4.2), it is physically more appropriate to have $V_{DS,\text{eff}}$ smoothly and asymptotically approach V'_{DS} as V_{DS} goes from below to above V'_{DS} . This can be done using a limiting function. Two commonly used forms of limiting functions for $V_{DS,\text{eff}}$ are^{25,58}

$$V_{DS,\text{eff}} = V'_{DS} - 0.5 \left[V'_{DS} - V_{DS} - \delta + \sqrt{(V'_{DS} - V_{DS} - \delta)^2 + 4\delta V_{DS}} \right] \quad (10.4.3)$$

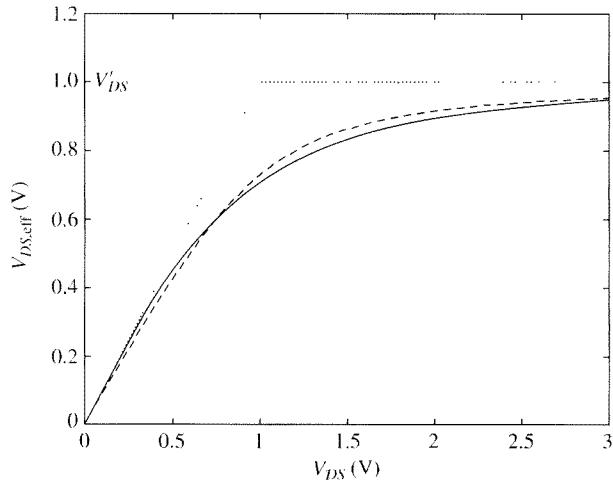
and⁵⁸

$$V_{DS,\text{eff}} = \frac{V_{DS}}{\left[1 + \left(\frac{|V_{DS}|}{V'_{DS}} \right)^A \right]^{1/A}} \quad (10.4.4)$$

where δ and A are parameters that control how tightly $V_{DS,\text{eff}}$ fits the asymptotes in (10.4.2) at $V_{DS} = V'_{DS}$. Figure 10.2 shows the behavior of both of these models.

Similarly, limiting functions are used for connecting strong- and weak-inversion behavior.^{25,45,55,57,59} Typical forms involve interpolation functions with a combination of logarithms and exponentials.^{25,55,57} (We have seen an example of this approach in Sec. 4.9.) In such an approach, the behavior predicted in moderate inversion is purely determined by the selection of which mathematical functions to use. Despite its lack of physical basis, the interpolation approach gives smooth, well-behaved characteristics in moderate inversion. Interpolation models have proven to be effective, and sufficiently accurate, in practice.

The “interpolation” approach can be a useful tool but is not a panacea; as with empirical modeling in general, *it should not be abused*, but rather should be used as a last resort, if direct smooth expressions are not possible. This tool has been a cornerstone of MOS transistor modeling. However, since techniques have been developed to solve some long-standing problems with properly incorporating effects such as velocity saturation into surface potential-based models, the importance of interpolation models is declining. Nevertheless, the concept of using smooth mathematical

**FIGURE 10.2**

Functions for $V_{DS,\text{eff}}$. Dashed line is (10.4.3) with $\delta = 0.1$. Solid line is (10.4.4) with $A = 2$. Dotted lines are asymptotes.

limiting functions, rather than hard “if” conditions, to transition from one mode of operation to another has become a standard tool for model developers.

The Use of Conditionals One constraint for “perfectly smooth” models is that they should have no “if” conditions in the model formulation (e.g., based on region of operation); but there are important reasons why conditionals are useful in model code. For efficiency, conditionals can be used to include or exclude blocks of model code that are turned on or off by various model parameters or switch settings, for example, for calculation of gate current or non-quasi-static effect modeling. Conditionals can also be used to avoid numerical calculation problems. For example, consider the use of the following common limiting function:

$$y = \frac{1}{A} \ln[1 + \exp(Ax)] \quad (10.4.5)$$

This function asymptotically approaches zero for large negative values of x , asymptotically approaches x for large values of x , has the value $\ln(2)/A$ at $x = 0$, and changes from this value toward the asymptotes as x moves away from zero. It is used to implement a relatively “hard” lower limiting behavior. However, for large positive x , the argument of the exponential can be large and cause a numeric overflow. One can avoid this problem by using instead

$$y = \begin{cases} \frac{1}{A} \ln[1 + \exp(Ax)], & x \leq 0 \\ x + \frac{1}{A} \ln[1 + \exp(-Ax)], & x > 0 \end{cases} \quad (10.4.6)$$

It is easy to show that this expression is identical to (10.4.5) for all x ; yet the exponential overflow for large positive x is avoided. Using different expressions for positive and negative x here does not introduce any discontinuity but enables well-behaved numerical evaluation.

Out-of-Range Behavior Although the model expressions are only intended to be accurate within certain ranges of voltages and currents, they must be well-behaved even far outside those ranges.⁴⁵ In the course of numerical iterations, a simulator may try values well outside the intended ranges; if an exponential or a polynomial blows up, or a denominator becomes zero during such excursions, attempts to achieve numerical convergence can terminate unsuccessfully. A robust model should be able to handle extreme bias values, well outside the range over which it was characterized, and still return numerically well-behaved values for currents, charges, and their derivatives with respect to terminal voltages. Reasonable behavior should also be ensured vs. other parameters, such as geometry or temperature. In general, as bias, geometry, or temperature approaches some limiting value or range, the behavior of a model should asymptotically approach the expected physical behavior. This is referred to as *asymptotic correctness*.⁴⁵ Though this seems imprecise, it is fairly easy to apply in practice. One can inspect the model equations and see if anything “blows up” unexpectedly.

An example of a case where inappropriate out-of-range behavior may occur is now given. It is common practice in modeling to include empirical variations of several quantities with respect to bias, temperature, or geometry by using low-order (linear or quadratic) polynomials in those quantities. This can seem to improve model accuracy when verified over a small number of data points but can lead to incorrect modeling over all required geometries and temperatures. For example, polynomials in L or W can yield very large values for large geometries. One way to avoid this problem is to add corrections not via polynomials in L or W , but via terms in $1/L$ or $1/W$. The contribution of these terms asymptotically goes to zero for large geometries, so they are in effect corrections for small geometries, which is generally where such corrections are needed anyway.

Charge vs. Capacitance Formulation In Chap. 8, we have shown that nine capacitance coefficients [see (8.2.12)] are needed to describe the small-signal charging currents of a MOS transistor. These can be calculated as the derivatives of three independent charges (although the four charges Q_S , Q_D , Q_G , and Q_B characterize the charging terminal currents of a MOS transistor, charge neutrality dictates that their sum is zero, so one can be calculated from knowledge of the other three). There is an important reason for modeling directly in terms of charges, rather than capacitances, which is now briefly discussed.

During transient simulations, all quantities involved are calculated at discrete time points t_k , $k = 1, 2, \dots$. In early circuit simulators, MOS transistor models were generally implemented based on capacitances.¹ Between any two points, say t_k and t_{k+1} , capacitances were considered constant and equal to their value at the first point, or the mean between the values at the two points. The approximate currents calculated in this way were then used to calculate charges through numerical integration.

This led to errors in calculating the total charge transferred between two such time points, and this error could in certain cases accumulate, leading to absurd results; for example, one could have a capacitor with the voltage across it starting from a given value, changing, and then returning to the same value, and yet the charge on its plates did not return to the value it had started with. This problem is known by the generic name of “charge non-conservation”^{60,61} and led to serious problems in the design of certain circuits. These notably include circuits with nodes that have no or only very small dc conductive paths to other nodes, such as certain digital and switched-capacitor circuits. The accumulation of charge errors led to the erroneous appearance of voltage buildup on such nodes. Although the size of the error could be reduced by reducing the time steps $t_{k+1} - t_k$, this only reduced rather than eliminated the problem and led to increased simulation time.

Modeling via charge relations of the form $Q = Q[V(t_k)]$ clearly does not have the preceding problem, because as long as the voltage is the same, the charge is the same, too. Compact models should therefore be formulated in terms of charge rather than in terms of capacitance. Of course, the model one implements must be correct to begin with; a complete quasi-static formulation is one of the requirements for ensuring charge conservation.^{60–63}

Non-Quasi-Static Analysis Formulations An introduction to the non-quasi-static operation of MOS transistors has been given in Sec. 6.7.3. Although the full non-quasi-static equations used there are useful for describing the principles, the numerical solution of such equations is too complex and computationally inefficient to be used for the purpose of circuit simulation. An approach that is efficient computationally is to introduce a model based on a relaxation time, or delay.^{64,65} Consider a device for which the terminal voltages have been held at some constant value, for a period of time sufficient for the charges to be static, and are then instantaneously changed to different values. The inversion layer charge $q_I(t)$ will now adjust over time to an eventual value Q_f , dictated by dc considerations, as electrons move through the transistor. In effect, $q_I(t)$ “relaxes” toward Q_f . This is modeled as a first-order system, with a time constant τ_r , called the “relaxation time” by an equation of the form^{29,64,65}

$$\frac{\partial q_I}{\partial t} = -\frac{q_I - Q_f}{\tau_r} \quad (10.4.7)$$

which is of the same form as for other first-order systems, e.g., a first-order RC circuit. The relaxation time is related to the transit time (Sec. 6.5), but in general additional empirical parameters are included that can incorporate velocity saturation effects, which are determined to best fit experimental data. This approach is not a detailed and self-consistent model, but for computer simulation it is simple and efficient to implement and addresses the main deficiencies of quasi-static models for simulation of fast transient responses. Introducing a relaxation time model means that during analysis a circuit simulator must introduce an additional internal state variable to keep track of the time-dependent q_I .

A more precise approach is based on the “spline collocation” technique. Introducing the drift-plus-diffusion current equation (4.3.4) into the continuity equation (6.7.5) gives^{66,67}

$$\frac{\partial q'_I}{\partial t} + \frac{\partial}{\partial x} \left[\mu \left(\frac{q'_I}{dq'_I / d\psi_s} - \phi_t \right) \frac{\partial q'_I}{\partial x} \right] = 0 \quad (10.4.8)$$

This defines a constraint that $q'_I(x,t)$ and its derivatives must satisfy along the channel. A given form for $q'_I(x,t)$ as a function of position along the channel is assumed, by using, for example, quadratic^{68,69} or cubic⁷⁰ spline polynomial basis functions. A number of points are specified along the channel, called the “collocation points,” and (10.4.8) is required to be satisfied at each of these points. This gives a set of ordinary differential equations that, along with conditions for continuity of $q'_I(x,t)$ and its derivatives at each of these points, can be solved for the coefficients of the basis functions. The inversion charge density at each of the collocation points can then be used to compute individual source and drain charges using the correct x/L partitioning of (6.3.9). This is a more complex approach than the relaxation time approximation, but it allows selectable accuracy vs. complexity trade-off through specification of the number of collocation points. Effects such as velocity saturation can also be taken into account using this technique.⁷⁰ It is important to note that although physical collocation points along the channel are introduced, this approach is not equivalent to the channel segmentation approach outlined with the help of Fig. 6.12. In channel segmentation, complete and independent MOSFET model evaluations are done for each segment. The spline collocation approach does introduce additional internal state variables into a model, but computations related to these variables are done to solve (10.4.8) and not to evaluate a complete MOSFET model for each segment. Thus, the computational complexity lies between that of the relaxation time approach and the channel segmentation approach.

10.5 Model Implementation in Circuit Simulators

User Interface When a model is used for circuit simulation, a “netlist” (a text file that describes the components of the circuit and how they are interconnected) is generated from the CAD system being used to define the circuit, and this netlist is read in by a circuit simulator, which then numerically evaluates the electrical behavior of the circuit. The MOSFETs in the circuit are represented by “instance lines” in the netlist, for example

```
m137 (d_node g_node s_node b_node) nmos_lp l=100n w=1u ...
```

Here “m137” is an identifier for the instance; the next items in parentheses indicate the names of the nodes in the circuit to which the drain, gate, source, and bulk terminals of the MOSFET instance are connected; “nmos_lp” is an identifier of the specific type of MOSFET in the technology in which the circuit is being designed;[†] and

[†] Semiconductor manufacturing processes generally include several different types of *n*MOS and *p*MOS transistors, e.g., thin oxide “core” devices and thicker oxide “input/output” transistors, perhaps with two or more options for threshold voltage to trade-off switching speed vs. leakage and power.

the remaining entries are “instance parameters” that specify the geometric layout of the particular MOSFET. In this example, just the length and width are shown.

The “nmos_lp” identifier links the MOSFET instance to a particular type of device available in the technology in which the design is being done. The model parameter file provided for this technology will indicate which type (formulation, equations) of model is to be used to model this type of device, and the values of the parameters that characterize that model for the specific device and technology, for example

```
.model nmos_lp modelName type=n tox=2.0e-9 vfb=-0.937 na=5.3e17 mu0=437.2 ...
```

The process that determines these parameter values is generally referred to as “characterization” or “parameter extraction.” The generation of these model parameter files is the job of modeling (or “characterization”) engineers. Note that the same core model equations are used to characterize many different types of MOSFETs in many semiconductor manufacturing technologies; therefore most “modeling” tasks are not the development of model equations but the extraction of model parameters. This task is the subject of Sec. 10.7.

MOS transistors are not stand-alone, isolated devices; their operation and behavior depends on adjacent structures such as metal interconnects and contacts. The details of these structures are not known prior to final layout, so they need to be estimated based on assumptions of typical layout, for schematic-based (i.e., prelayout) simulation. Later, once the physical IC layout is done, a “layout extractor” program computes the quantities involved more accurately; they are then included in the circuit netlist and the prelayout estimates are turned off. “Switch” parameters are included in models specifically for this purpose.

Circuit Simulation Circuit simulators are, at heart, sophisticated continuity equation and ordinary differential equation (ODE) solvers.^{1,71} The basic goal of a simulator is to solve the Kirchhoff current law equations at each node of a circuit: the sum of all currents impinging on a node should be zero. For dc analysis, this is just the dc currents in the devices that are connected to form the circuit being analyzed. For small-signal analysis, this is both the real and imaginary (conductive and susceptive) components of the phasor currents flowing in a linearized version of the network at its quiescent dc bias point. For transient analysis, it is the sum of the transport currents and the $\partial Q/\partial t$ charging currents. For harmonic balance analysis, it is the Fourier components of currents at each frequency that must sum to zero.⁷² There is an ever-growing number of sophisticated analysis techniques being added to circuit simulators, for example, for jitter or oscillator analysis, and these can require special capabilities in models. Simulation is in large part an exercise in modeling; the most sophisticated circuit analysis technique will give incorrect results if the device models that underlie that analysis are incomplete or incorrect. The requirements of the basic dc, small-signal, and transient analyses (harmonic balance analysis leverages both small-signal and transient analysis, for linear and nonlinear components, respectively) dictate how a model should be formulated. We have already seen that this is best done through $I(V)$ and $Q(V)$ relations. Even resistors are better modeled as $I = GV$, as this form is more

suitable for modified nodal analysis,⁷¹ in which the unknowns being solved for are voltages.[†]

Model Code Implementation Compact MOSFET models are complex; a hand-coded C program for implementation of a complete model in circuit simulators can exceed 20,000 lines of code, as already mentioned. Historically, this has meant that coding a full model, which includes all nonidealities and second-order effects important for practical IC design, was the domain of a small number of experienced people. The model codes used to include manually generated derivatives. Derivatives are required for the dc convergence algorithms used in simulators and also for small-signal analysis (Chaps. 7 and 8). Thus, derivatives are tightly integrated into the core algorithms and code of a simulator. Manually coding derivatives made model implementation and maintenance time-consuming, tedious, and error-prone. Also, details of model implementation can vary between simulators, and this has magnified the overall implementation effort and led to unnecessary rework when a model was updated.

Simulators have evolved to have model interfaces that separate models from the specifics of numerical methods, such as details of the integration methods used for transient analysis. A compact model can be abstracted to a small number of constituent parts: a set of terminals; a set of model parameters; a set of instance (layout geometry) parameters; an equivalent network topology; and a set of branch constitutive relations that define currents and charges as functions of voltage, temperature, and parameter values. Recognizing this allows a model to be defined at a high level, independent of the specifics of implementation in a particular simulator. Also, symbolic differentiation packages have become widely available, so derivatives can be automatically, rather than manually, generated. This significantly reduces model coding time and, at the same time, removes the main source of errors in model codes. The industry has therefore been moving toward defining compact models in high-level languages and using compilers to generate low-level model code. Models defined this way have only about 10–20 percent of the number of lines of code as a model with explicit hand-coded derivatives. One barrier to adopting this approach was the lack of a standard language for model definition, but in the end, Verilog-A emerged as an almost ideal language for this purpose. Key additions to support compact modeling in this language have been made.⁷³

There are two main ways to use a model defined in Verilog-A for circuit simulation. The first is via a run-time “native” compiler in a simulator (the latter being software that, while a simulation is running, compiles model code and then dynamically links it to the simulator). This is the easiest approach to follow, but it incurs an overhead in compilation time when a model is first used and, at present, generally results in computation times that are much longer than if a model was hand-coded. The second approach involves a separate compilation to make the model appear to be “built-in” to the simulator (although many simulators dynamically link such models

[†] For very small-valued resistances, i.e., very large conductances, the $V = IR$ form is necessary, for numerical stability.

at run time). Models generated this way can be nearly as efficient as hand-coded models.⁷⁴

We should stress that, although this book emphasizes physical compact modeling, in the end, a model used for circuit simulation is software. The model should be implemented based on an understanding of the requirements of the algorithms used in circuit simulators. Most model developers have a background in device physics and modeling, some have a background in circuits and circuit design as well; but few have the core background in software required to develop and maintain industrial-strength code for a widely used, complex model. This background includes proper source coding, the generation of model documentation, and extensive testing.

10.6 Model Testing

A model software implementation needs intensive testing before release. Some of this testing focuses on the software coding itself and involves revision control, effective code reviews, and, most important, quality assurance and regression testing to exercise a model thoroughly and to catch and correct problems before a model is released. Even if the software code performs well, the ultimate test is how well a model predicts the quantities it was designed to predict. Fortunately, there are several benchmark tests³⁹ that have been developed over the years to help in checking model predictions; these can be used by model developers (professional ones or students doing projects) as well as model users. Most of these tests are passed by professional models today, but not all are passed by all popular models! Benchmark tests are summarized in Appendix K.

10.7 Parameter Extraction

No matter how exact a physical model is, it cannot give accurate results unless appropriate values are used for its parameters. These values should be determined to best represent the measured behavior of the type of device being modeled. Exactly what “best” is can be difficult to define and can be different for different types of circuits. Ideally a model should fit measured dc currents (drain-source current, gate leakage, junction current, and impact ionization current), small-signal conductances and capacitances, high-frequency and large-signal behavior (including second- and third-order derivatives, for accurate modeling of distortion), and noise and should do so over a range of biases, geometries, and temperatures. In practice this is difficult, and there are inevitable differences between modeled and measured behavior, even for dc I_D over bias for a single geometry and temperature. There are several reasons for this. First, the value of some of the model parameters may not be known accurately; for example, the flatband voltage may not have been measured independently. Second, some of the parameters in the models are, as we have seen, basically empirical in nature, which means by definition that they should be chosen for best matching to measurement. Third, even if the value of a physical parameter is known accurately, this value may not be the best one to use in the model expressions. This is because

physical compact models are based on several assumptions and approximations. This is the case, for example, for the all-region, surface potential-based model; the core model derivation is based on an assumed uniform doping profile. Thus, using models with the “correct” values for their parameters results in a certain error. By slightly modifying the value of these parameters, the error can often be decreased; we have already discussed this fact in Sec. 4.17. In other words, a slight modification of the parameter values can in some instances partially make up for model deficiencies. Therefore, techniques are required to determine model parameters from experimental data.^{55,75–81}

The process of determining values for the parameters in a model is called *parameter extraction* or *characterization* and is combined with measurement. A large number of different bias voltages and signal excitations are applied to the transistor under test, by programmable test sources under the control of a computer. The response of the device (e.g., the corresponding drain current values) is measured by programmable equipment, also under computer control. An *optimizer* program^{82–85} then determines the best parameter values to use for a given model, so that a certain criterion is met; for example, it is usually desirable to choose the parameter values so that the I_D values predicted by the model match the measured values with as little overall error as possible, where “error” is defined in some appropriate way to be discussed.

In some cases, the data to be matched may come not from a measurement system, but from a device simulator. This would be necessary if it is desired to produce an analytical model before actual fabricated devices are available for measurements.

Error Criterion Let us assume that the quantity to be optimized is the drain current over a range of bias values. A common error criterion is the simple relative error measure

$$E_I = \sum_{j=1}^K w_{Ij} \left(\frac{\hat{I}_{Dj} - I_{Dj}}{I_{Dj}} \right)^2 \quad (10.7.1)$$

where I_{Dj} , $j = 1, \dots, K$ are measured I_D values for K different combinations of bias voltages, \hat{I}_{Dj} are the corresponding calculated values (the difference $\hat{I}_{Dj} - I_{Dj}$ is often called the “residual”), and w_{Ij} are weighting coefficients for the drain currents, which are often defaulted to equal 1. If desired, these coefficients can be chosen to be larger for some of the bias voltages, in which case the error at those biases will count more in the overall error, and the parameter extraction routine will thus be forced to “pay more attention” to those biases. The squares are used so that both positive and negative errors count equally and so that such errors do not cancel each other out; sometimes, absolute values are used instead of squares. Note that, for very small currents, the denominators in (10.7.1) become very small, whereas, in general, the errors represented by the numerators do not. This tends to assign more importance to the error at small current values and can be corrected by choosing the weighting coefficients accordingly. If no weighting coefficients are used, it may be desirable not to allow the denominator to decrease below a certain value, say I_{\min} ; in that case, the denominators in (10.7.1) can be replaced by $\max(I_{Dj}, I_{\min})$. A more sophisticated relative

error measure that is symmetric with respect to percentage errors and helps automatically reject “outliers” in data, especially noisy data at low bias levels, is⁸⁶

$$E_I = \sum_{j=1}^K w_{IDj} \left(\frac{\hat{I}_{Dj} - I_{Dj}}{|\hat{I}_{Dj}| + |I_{Dj}| + \epsilon} \right)^2 \quad (10.7.2)$$

where ϵ is small number, for example, of the order of I_{min} , below which errors are progressively scaled to zero.

Minimization of the error criterion (10.7.1) results in what is known as a “least mean square error fit,” or a “least squares fit” for short. If the function that is being fitted to experimental data is linear in the parameters of the model, the resulting optimization is called “linear regression.”^{87,88}

The root mean square (rms) error corresponding to (10.7.1) is:

$$E_{rms} = \sqrt{\frac{E_I}{K}} \quad (10.7.3)$$

For analog applications, minimizing (10.7.1) does not necessarily produce adequate results. Let us revisit a point we made in Chap. 7. Assume that extraction of a parameter for a certain model has been done in order to minimize the error measure (10.7.1), resulting in the satisfactory fit to data shown in Fig. 10.3a. When the same model is used to predict the small-signal output conductance g_o (the derivative $\partial I_D / \partial V_{DS}$), the result is as in Fig. 10.3b. The (relative) error involved in fitting this quantity can be defined in a similar way as before:

$$E_{go} = \sum_{j=1}^K w_{goj} \left(\frac{\hat{g}_{oj} - g_{oj}}{g_{oj}} \right)^2 \quad (10.7.4)$$

where w_{goj} are weighting coefficients for output conductance at each bias. As seen in Fig. 10.3b, the error in g_o can be very large, even though the fitting of I_D is very good. This is because the slope in saturation is so small that, even if it is in error by, say, a factor of 2, the corresponding effect on the I_D curve (Fig. 10.3a) will be small and will hardly be noticeable, if minimizing E_I is the objective. To rectify this situation, the optimization goal should be to minimize a weighted combination of both current and output conductance errors.⁸⁹ This can provide a significant improvement in g_o modeling accuracy, with negligible loss in overall I_D modeling accuracy.⁸⁹⁻⁹¹ g_o is usually computed from numerically differencing of I_D - V_D data, or from direct measurement (which can be difficult⁹²). Errors in g_m and g_{mb} can also be included. However, these parameters are usually predicted relatively accurately if the current is predicted accurately; modeling of g_o generally has significantly greater errors.

The previous errors involve only dc measurements. In a comprehensive parameter extraction procedure, the error in other quantities (capacitances, rise times, noise, etc.) should be taken into account.

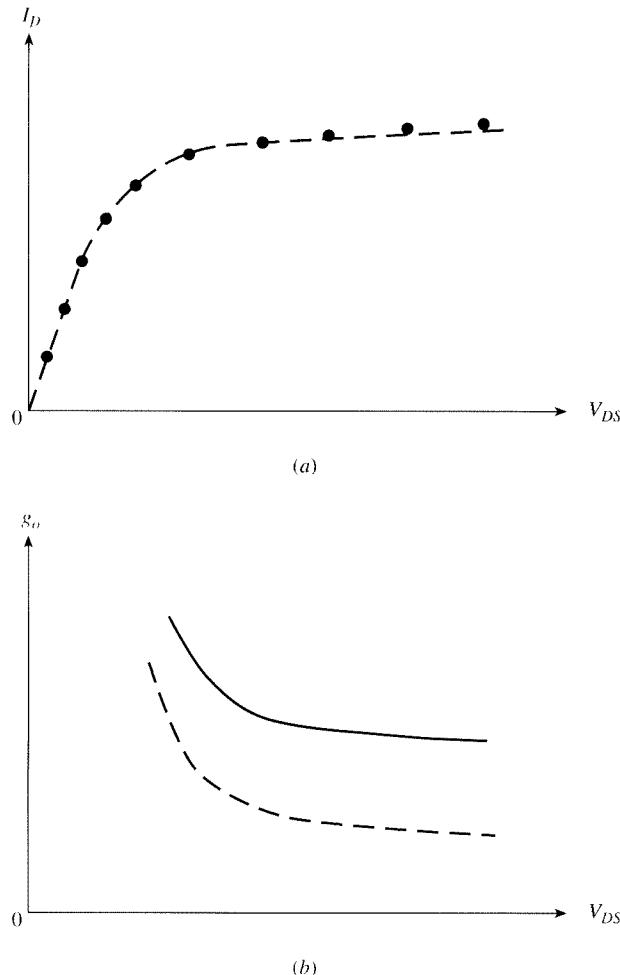


FIGURE 10.3
(a) I_D - V_{DS} characteristic. Points: measurements; broken line: model. (b) Slope g_o corresponding to (a). Solid line: measurement; broken line: model.

Extraction Strategies The basic strategy for parameter extraction is “divide and conquer”; i.e., separate parameters into groups that affect modeling of specific device behaviors and then characterize these parameter groups from measurements of specific geometries and biases that accentuate the influence of these parameters. Modern MOSFET models have hundreds of parameters; these, somewhat ironically thanks to the push for single-piece rather than regional models (a good thing in itself), affect modeling in all regions of operation. This would seem to undermine the “divide and conquer” approach. However, in general there are specific test structures, measurements, and analytic techniques that can separate data-parameter groupings and help make parameter extraction a stepwise, sequential process. For example, we have seen that wide/long

transistors tend to display the most “ideal” behavior, compared with our theoretical models, and that deviations from this behavior occur as channel length decreases and as channel width decreases. So basic parameters, such as V_{FB} , N_A , and μ_0 are determined based on measurements from a wide/long device, or at least initial estimates of them are computed from such a device. Parameters that affect short-channel behavior are then determined from a set of devices with different channel lengths, which are all wide so that there are minimal narrow channel effects in them. Similarly, parameters that affect narrow-channel behavior are determined from a set of devices with different channel widths, which are all long so that they are minimally affected by short-channel affects.

If possible, model parameters should be determined from device geometries and biases that are typical for circuits, based on measured quantities that are relevant to circuit performance, to avoid additional errors from extrapolation from atypical to normal conditions. However, this is not always possible; some steps can involve device geometries or biases that are not directly relevant to most circuits, but which greatly simplify determination of some parameter values, for example, the use of a wide/long transistor to characterize some basic parameters. The extraction sequence may involve small dedicated calculations and may also involve manipulations and approximations that lead to relations between various quantities that should be linear; this allows parameter values to be easily estimated from the slope and/or intercept of a plot using least squares fitting^{87,88} and may involve numerical optimizations that use a computer program to adjust parameter values so that some measure of error, e.g., a combination of (10.7.1) and (10.7.4) between modeled values and measurements is minimized. We will discuss each of these. In general, after a sequence of such steps has been followed to arrive at some initial estimates of parameter values, one or more “global” optimization steps, which involve large sets of parameters and large sets of data, are performed to fine-tune the model fit to the measured data.

A critical concept in parameter extraction is to understand exactly what a “parameter” is and what it is not. There are certain quantities, such as V_T and effective electrical channel length L , that are sometimes considered to be basic “physical” MOSFET parameters. Therefore, they should be able to be determined directly, and unambiguously, from measurements. Physical analysis techniques, such as scanning emission microscopy (SEM) and secondary ion mass spectrometry (SIMS), enable direct measurement of physical dimensions and dopant distributions in a MOSFET. But these do not translate directly into quantities such as V_T and L (the latter is a physical dimension but is not directly measured by an SEM). Even if we have some “physical” procedure to determine V_T and L , if these are used in two different MOSFET models, *they will give different results* because, as we have discussed, models are not perfect representations of real devices but have differences with respect to measured data, and these differences will be different for different model formulations. Therefore, as far as parameter extraction is concerned, a “parameter” is a *model* parameter, which is not a physical quantity. Some model parameters are strongly physically based and would be expected to have values close to those determined from physical measurement, for example, substrate doping N_A and saturation velocity

$v_{d,\max}$. Other model parameters are not as strongly physically based, or are empirical, and are in essence fitting parameters. There have been a large number of techniques published to determine V_T and L ,⁷⁶ and different techniques give different values; this is primarily because the details of the *model* that underlies each technique is different.

Parameter Extraction Examples Describing a complete parameter extraction procedure is a complex task and is beyond the scope of this book. Hence, we will only provide the flavor of the procedure here by giving examples of the extraction of some key parameters.

The first step in parameter extraction is to determine t_{ox} . For thick oxides, this is done by measuring the capacitance of a large-area device and computing t_{ox} from (2.2.4). Because the MOS C-V relationship depends on bias (Chap. 2), the voltage applied to the MOS transistor must ensure it is in strong inversion or strongly in accumulation for (2.2.4) to be a reasonable approximation. For thin oxides, C-V measurements are affected by gate leakage and parasitic series resistances and the oxide capacitance must be determined from more detailed two-frequency or RF measurements.^{93,94} The capacitance of thin oxides is, as we saw in Chap. 5, affected by polysilicon depletion and quantum mechanical (QM) effects. In this case, t_{ox} cannot be directly computed using (2.2.4) but must be determined as the value of the model parameter t_{ox} that, when used to simulate MOS C-V characteristics taking polysilicon depletion and QM effects into account, best matches the measurements. Of necessity, this requires determining the polysilicon doping concentration at the same time, so optimization is generally used.

Once t_{ox} is known, some first-order parameters, e.g., substrate doping concentration, flatband voltage, and mobility, can be estimated from dc measurements of a wide/long transistor, as such geometries are affected least by parasitic series resistance and narrow- and short-channel effects (remember “divide and conquer”). Although for many circuit applications MOS transistors are operated in saturation, DIBL and velocity saturation effects increase with drain voltage, so measurements and parameter extraction techniques to elucidate core MOS behavior are often done at low V_{DS} , generally 0.05 or 0.1 V, to minimize the influence of these effects.

The most commonly used model for first-order analysis of long-channel MOS transistors in strong-inversion nonsaturation is (4.7.24a), with the effective mobility model (4.11.31). These give (assuming negligible body and gate current, so $I_D = I_{DS}$)

$$I_D = \frac{W}{L} \frac{\mu_0 C'_{\text{ox}}}{1 + \theta(V_{GS} - V_T) + \theta_B V_{SB}} (V_{GS} - V_T - 0.5\alpha V_{DS}) V_{DS} \quad (10.7.5)$$

The slope of the plot of I_D vs. V_{GS} is the transconductance, g_m (Sec. 7.2). Differentiating with respect to V_{GS} leads to (we shall need this later)

$$g_m = \frac{W}{L} \mu_0 C'_{\text{ox}} \frac{1 + 0.5\alpha\theta V_{DS} + \theta_B V_{SB}}{\left[1 + \theta(V_{GS} - V_T) + \theta_B V_{SB}\right]^2} V_{DS} \quad (10.7.6)$$

For small V_{DS} , $V_{SB} = 0$ V, and V_{GS} is somewhat above threshold, high enough so that the device is out of moderate inversion but not so high that the effective mobility is significantly reduced from its low-field value μ_0 , a reasonable approximation to (10.7.5) is

$$I_D = \frac{W}{L} \mu_0 C'_{\text{ox}} (V_{GS} - V_T) V_{DS} \quad (10.7.7)$$

This predicts that the drain current should increase linearly with V_{GS} . Figure 10.4 shows measured I_D - V_{GS} characteristics. Although the deviation from (10.7.7) is apparent at low and high V_{GS} due to moderate inversion and effective mobility dependence on gate voltage, respectively, there is a region where the linear approximation is reasonable. Selecting the two points that give the highest value for the slope g_m (computed by numerical differencing) defines the region where the inaccuracies from our assumptions (that we are above moderate inversion but the mobility is still close to μ_0) are small. The extrapolation of the line fitted to the two points to $I_D = 0$ intercepts the horizontal axis at approximately V_T ; more precisely, from (10.7.5), the intercept is $V_T - 0.5\alpha V_{DS}$, and commonly this is simplified to $V_T - 0.5V_{DS}$ (for small V_{DS} the error introduced from this approximation is small).

Other techniques to determine threshold voltage exist. From (4.7.24b) it can be seen that in saturation $\sqrt{I_D}$ vs. V_{GS} is predicted to be a straight line. Although the field dependence of effective mobility will cause deviation from exact straight line behavior, extrapolation as was done in Fig. 10.4 for I_D - V_{GS} in nonsaturation can also be done for $\sqrt{I_D}$ vs. V_{GS} in saturation. We have noted that it is desirable to use biases typical of those in circuits when determining parameter values, and from this point of view, operation in saturation is preferable to operation in nonsaturation. However, as we have also noted, operation in saturation is affected by DIBL and

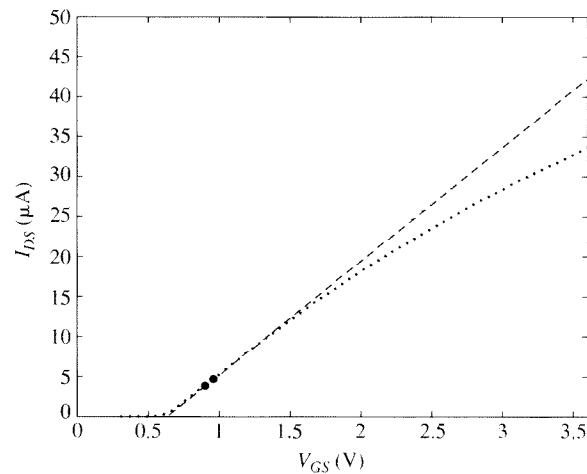


FIGURE 10.4

I_D vs. V_{GS} , $V_{SB} = 0.0$ V, $V_{DS} = 0.1$ V, $W = L = 25$ μ m. Points: measurements. Bold points: biases of maximum g_m . Broken line: linear fit to points for the maximum g_m .

velocity saturation, so to avoid confounding the value of V_T with the effects of other parameters, for the purpose of modeling, V_T is commonly extracted from measurements done in nonsaturation.[†] V_T can also be estimated by determining the value of V_{GS} required to produce a certain current, usually chosen to be in moderate inversion.⁹⁵ Unlike the peak- g_m extrapolation technique, the V_T extracted from this constant current approach depends on mobility and therefore can confound mobility with threshold voltage, which is undesirable, especially from the point of view of physically separating sources of statistical variability. However, the single constant current measurement can be performed much faster than the gate voltage sweep required for the extrapolation methods, and so to minimize test cost, it is often used for production measurements in manufacturing lines.

By repeating the peak g_m extrapolation process for I_D - V_{GS} sweeps at different body biases, the $V_T(V_{SB})$ characteristic of a transistor can be derived. At low drain bias, for which there is no DIBL effect, recall from (4.7.20a) that

$$V_T = V_{T0} + \gamma \left(\sqrt{\phi_0 + V_{SB}} - \sqrt{\phi_0} \right) \quad (10.7.8)$$

where

$$V_{T0} = V_{FB} + \phi_0 + \gamma \sqrt{\phi_0} \quad (10.7.9)$$

with $\phi_0 = 2\phi_F + \Delta\phi$, $\phi_F = \phi_r \ln(N_A/n_i)$, and $\gamma = \sqrt{2q\epsilon_s N_A}/C'_{ox}$. These expressions are valid for wide/long transistors with uniform-substrate doping and are modified for nonuniform transverse doping (see Sec. 9.3) and for the short- and narrow-channel effects and poly doping and QM effects discussed in Chap. 5. Nevertheless, these expressions are still useful for generating initial estimates of parameter values. Problems 10.6 through 10.9 give examples of different approaches used to determine N_A , and from this and an appropriate value for $\Delta\phi$, (10.7.9) can be used to compute V_{FB} (in many circuit simulation models, $\Delta\phi$ is taken to be zero, although as discussed in Sec. 2.6.2 this is inaccurate). Different MOSFET models have different expressions and parameters for threshold voltage variation with nonuniform doping and geometry, so a specific technique to determine them cannot be given here. However, the general procedure to fit $V_T(V_{SB})$ over geometry is to either develop a model-specific approach to get an initial estimate of some key parameters, or start from typical values and use optimization to adjust the threshold voltage model parameters to minimize the differences between the model and the measured $V_T(V_{SB})$ data. This is part of the “divide-and-conquer” strategy, as only parameters that enter in the threshold voltage model are included to fit the derived $V_T(V_{SB})$ data.

The deviation of the measured data from the extrapolated straight-line behavior in Fig. 10.4 above the maximum slope region is because of the field dependence of effective mobility, which is modeled by the terms in θ and θ_B in (10.7.5). This

[†] The value of V_T for operation in saturation, when affected by DIBL, is important for monitoring short-channel and leakage effects. So it is often measured in manufacturing lines, along with measurement of the nonsaturation value of V_T .

deviation can be used to determine θ (Prob. 10.10). Knowing V_T , θ , N_A , and C'_{ox} (computed from t_{ox}) from these steps and W and L from the layout of the transistor (assuming this is wide and long so the transistor width and length are the same as the mask width and length) and using α computed based on the specific form used in a model, μ_0 can be directly calculated from (10.7.5) for the measured drain current at one of the points of peak g_m .

The quantities W and L used in (10.7.5) and in many equations in this book are the “electrical” channel width and length, which are different from the corresponding “mask” (or “drawn”) values, W_m and L_m , respectively, specified during the design process.[†] The electrical and mask values are related by (1.6.1) and (1.6.2), repeated here:

$$W = W_m - \Delta W \quad (10.7.10)$$

$$L = L_m - \Delta L \quad (10.7.11)$$

The values of W_m and L_m are known from layout information; however, ΔW and ΔL depend on many aspects of the manufacturing process and need to be determined from measurement. This can be done as follows. In (10.7.6), use (10.7.10) to obtain

$$g_m = \left[\frac{\mu_0 C'_{ox} (1 + 0.5\alpha\theta V_{DS} + \theta_B V_{SB}) V_{DS}}{L [1 + \theta(V_{GS} - V_T) + \theta_B V_{SB}]^2} \right] (W_m - \Delta W) \quad (10.7.12)$$

Consider the same peak g_m points $g_{m,\max}$, that were used to determine V_T . One can thus plot $g_{m,\max}$ vs. W_m for several devices with different values of W_m , and the abscissa intercept of a least squares fit of peak will be ΔW , as indicated in Fig. 10.5. Note that the points used are at low $V_{GS} - V_T$, so the $\theta(V_{GS} - V_T)$ term in the denominator of (10.7.12) will be small and, more importantly, turns out to be close to the same value for each device width because it occurs at relatively the same position in the I_D - V_{GS} sweep. This separates the effect of threshold voltage from that of mobility-related parameters; such separation is preferred for parameter determination.[‡]

[†] Although called “mask” values, W_m and L_m are not necessarily the dimensions of the patterns formed on the masks to define MOSFET gates. Imperfections in modern manufacturing procedures cause final on-wafer sizes and shapes to differ significantly from those of the masks used during photolithography steps. Sophisticated shape and structure manipulation of the masks is done, including making them nonrectangular, to compensate for the manufacturing imperfections so that final on-wafer dimensions of the gate are as close as possible to the desired values W_m and L_m .

[‡] It would appear that one could instead determine ΔW using a similar procedure based on the drain current directly. From (10.7.5), it is apparent that I_D should be proportional to $W = W_m - \Delta W$; thus, plotting I_D vs. W , as obtained for several devices of different channel width at a given bias point, should give a linear relationship. The abscissa intercept of a least squares fit of I_D vs. W_m would therefore give ΔW . However, the data for this analysis need to be at constant $V_{GS} - V_T$, yet V_T can vary with width (Sec. 5.4); this means that some interpolation of the raw I_D - V_{DS} data is needed, and the procedure requires that V_T be known. This problem is avoided by using a g_m -based approach, as described above.

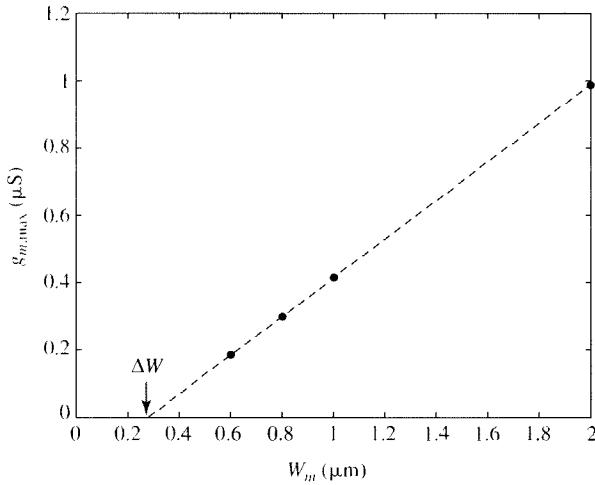


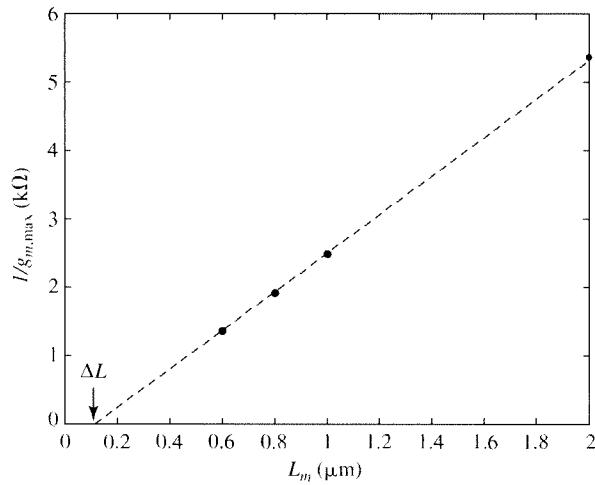
FIGURE 10.5
Peak g_m vs. W_m for ΔW extraction. Symbols are measured data; broken line is least squares linear fit.

Similarly, inverting (10.7.6) and using (10.7.11), we have

$$\frac{1}{g_m} = \left[\frac{\left[1 + \theta(V_{GS} - V_T) + \theta_B V_{SB} \right]^2}{W \mu_0 C_{\text{ox}} (1 + 0.5\alpha\theta V_{DS} + \theta_B V_{SB}) V_{DS}} \right] (L_m - \Delta L) \quad (10.7.13)$$

Again, the points of peak g_m occur at low and similar $V_{GS} - V_T$, so the x -axis intercept of a least squares fit of $1/g_{m,\max}$ vs. L_m is ΔL , as shown in Fig. 10.6. Other techniques use drain-source conductance, rather than transconductance, to determine ΔL and ΔW .⁹⁶ Related techniques can also be used to determine the source and drain parasitic resistances.^{96–98}

Complications due to Halo Regions The preceding approach is reasonable for devices that have a constant doping along the channel, such as higher-voltage devices, but is inaccurate for devices that have nonuniform lateral doping, which includes modern technologies with halo implants. In deriving (10.7.5), we assumed that the low-field surface mobility μ_0 does not vary with position along the channel. This mobility decreases with increasing doping concentration, and therefore is nonuniform along the channel for halo-implanted devices; the doping is higher adjacent to the source and drain regions than in the middle of a long-channel device. We have seen in Sec. 9.4 that there are details of the behavior of halo-implanted devices that can only be captured using multisection models. However, because of computational cost, these are usually avoided in practice; rather, a single-section model, derived assuming constant lateral doping, is used with appropriate length dependence of key parameters, such as threshold voltage and body effect coefficient, to capture the most important first-order effects of the nonuniform lateral doping (Sec. 9.4). As part of this approach, a single “effective” value of μ_0 is assumed to be constant for

**FIGURE 10.6**

Reciprocal of peak g_m vs. L_m for ΔL extraction. Symbols are measured data, broken line is least squares linear fit.

a given length, and its value will decrease as L decreases because the halo regions, which have a lower mobility, form an increasingly larger proportion of the overall channel length, increasing the average channel doping concentration. This means that in (10.7.13), μ_0 also varies with L_m and therefore $1/g_{m,\max}$ vs. L_m is not linear and cannot be used to determine ΔL as suggested previously. It is possible to introduce a three-section model, with separate mobilities for the central and halo regions and, from this more complex model, to determine ΔL from a plot of $1/g_{m,\max}$ vs. L_m .⁹⁹ However, this is not consistent with the single-section model used for simulation. Also, as we have discussed in Sec. 9.4, the peak g_m of halo-implanted devices does not depend solely on $\mu_0 C'_\text{ox} W/L$, because the device behaves as if it had an “effective” channel length that is approximately the sum of the halo region lengths near turn-on, and gradually changes to the total channel length as the gate bias is increased (Sec. 9.4). This confounds extraction based on the assumption that only mobility and L variations affect the peak g_m value.

Given that dc drain current, and hence g_m , depends on mobility, and the “effective” mobility of a device can vary with channel length, capacitance methods, which are not influenced by mobility, are also used to determine ΔL . The gate capacitance of a MOSFET includes intrinsic components, whose values are proportional to L , and extrinsic overlap and fringing capacitance components. C_{gb} measured in accumulation consists of only an intrinsic component, and is thus proportional to $L = L_m - \Delta L$. Therefore, if this capacitance is measured for devices of different lengths, the abscissa intercept of a least squares fit of C_{gb} vs. L_m is ΔL .^{100,101} However, capacitance measurement of leaky dielectrics, as encountered in thin oxide devices, can be difficult. As an alternative, it has been noted that the gate-to-bulk leakage current in accumulation is also proportional to L ; so ΔL can similarly be calculated as the abscissa intercept of a plot of this leakage current vs. L_m .¹⁰² These two

approaches are not in line with our stated preference to determine parameters from measurements at biases commonly used in circuits (MOSFETs are rarely biased in accumulation); however, they are necessary because of the difficulty of extracting ΔL , one of the most important MOSFET model parameters, for halo-implanted devices based on I_D - V_{GS} data.

Parameter Value Accuracy The preceding descriptions give some flavor of the types of analyses that are used to determine model parameters. Modern “industrial strength” MOSFET models include many hundreds of parameters, some highly physical in nature and some empirical, and determining values for these typically involves a sequence of many steps, some of which generate estimates of parameter values from direct calculation or from least squares fitting. These are followed by nonlinear optimization steps in which a small subset of parameters is adjusted to minimize some form of residual (i.e., difference) between model calculations and measurements. Several commercial parameter extraction systems and programs are available, and although early on these systems provided limited capability, such as only being able to optimize based on drain current residuals, they are now flexible and programmable. Thus, user-defined quantities important for analog design, such as I_D/g_{sd} and g_m/I_D , can also be defined as targets for optimization, and multiple targets can be specified and weighted differently based on geometry and bias.

Care needs to be exercised when using optimization to determine parameter values. Modern models have so many parameters and account for so many interacting effects that it is possible for incorrect values for one or more parameters to be compensated for by adjustments to other parameters. Stated another way, there are many different sets of parameter values that give essentially the same quality of fit of a model to measured data. For this reason, parameter extraction is sometimes abused, with serious consequences. *It is extremely important for the most physical model parameters to have physical values, and for all parameters not to have unreasonable values.* Otherwise, the unphysical values assigned can predict very wrong behavior in ranges or aspects not considered during the parameter extraction process. Parameter values for the optimization steps therefore are, or should be, constrained to lie within reasonable ranges.

Despite the widespread use of these sophisticated parameter extraction programs, it is *extremely* difficult to arrive at a set of parameters that does not have problems modeling some aspects of a particular device. This may be because the model being characterized has some unphysical behavior, such as asymmetry or a capacitance with the wrong algebraic sign, that cannot be overcome for any parameter values. Even without any unphysical behavior, no model is a perfect representation of a real device, and therefore trade-offs are made during characterization. These emphasize accuracy of fitting some facets of device performance, which are deemed most important for the target application areas of a technology, at the expense of increased errors in fitting other device performances that are considered less important. There has often been a focus on accuracy for digital CMOS circuit simulation with the compromise being accuracy for analog modeling. Even with a good model and great care being taken during parameter extraction, it is difficult to fit everything

well. Models are supposed to represent dc currents (from the leakage level up), capacitances and conductances (output conductance is often a very difficult quantity to model accurately), large-signal behavior for distortion modeling, noise, and statistical variations and to do this over a wide range of geometries (including adjacency to well and trench isolation edges), biases, and temperatures. If engineers are going to use a transistor in some uncommon manner or have an extremely high precision requirement, they should not just assume that the simulation models provided are accurate; instead, they need to verify the model accuracy and understand the limitations of the model and/or parameter set being used.

An example of the results of fitting measured data is shown in Fig. 10.7. Overall biases shown, the rms error in fitting I_D is 0.12 percent and the maximum error in fitting I_D is 1.0 percent. For the output conductance, which was also specified as a fitting target for optimization, the rms error is 2.2 percent, but the maximum error is 79.3 percent, for the point identified by the enlarged symbol in Fig. 10.7c. Visually the overall fit to the measured data looks adequate, and this is indeed an example of a very good fit; however, clearly at this point there is a very large error. Note that this is invisible in the plot of g_{sd} vs. V_{DS} on a linear scale, Fig. 10.7b, which highlights the need to view such plots on a log-linear scale. Even on the log-linear plot of Fig. 10.7c, the large error for this point does not visually “leap out.” This is because for this point the eye is drawn to match the solid line, which is from the model, by looking horizontally rather than vertically (which is how the fitting error is computed). In practice this error is of minimal consequence if a design is done properly. There is a rapid change in g_{sd} through the transition from nonsaturation to saturation, and the exact point at which this happens will vary with statistical fluctuations in devices and with temperature. If a circuit critically depends on a specific value for g_{sd} , then it should not be designed to operate at or near this point. An automated optimization program does not know if accuracy of fitting g_{sd} at that point is important (unless the specified weighting is zero or small); therefore it is essential for the optimization results to be reviewed and verified by an experienced characterization engineer. This is often an iterative process whereby, after an initial parameter extraction sequence is performed, “playback” plots of modeled vs. measured results are generated and reviewed, and weighting coefficients for specific targets (I_D , g_{sd} , g_m/I_D , etc.) for specific geometry and bias values are updated to balance trade-offs in fitting these quantities based on a knowledge of what is most important for the design of products targeted for the manufacturing process.

Another example of plots that may be misleading is shown in Fig. 10.8. Model fits to data in weak inversion are often presented on a logarithmic ordinate axis scale, and this visually shrinks the apparent error; the difference between the two curves in Fig. 10.8 seems to be reasonably small visually even though the percentage error in I_{DS} is up to 50 percent.[†] This error can easily be minimized by shifting the curve by only 15 mV, which can be accomplished by changing the flatband voltage

[†] In weak inversion, currents vary approximately exponentially with threshold voltage, and so, although, the error is large in relative terms, it is small compared with the statistical variation of the manufacturing process. Thus, the particular problem exhibited in Fig. 10.8 is not too great a concern in practice.

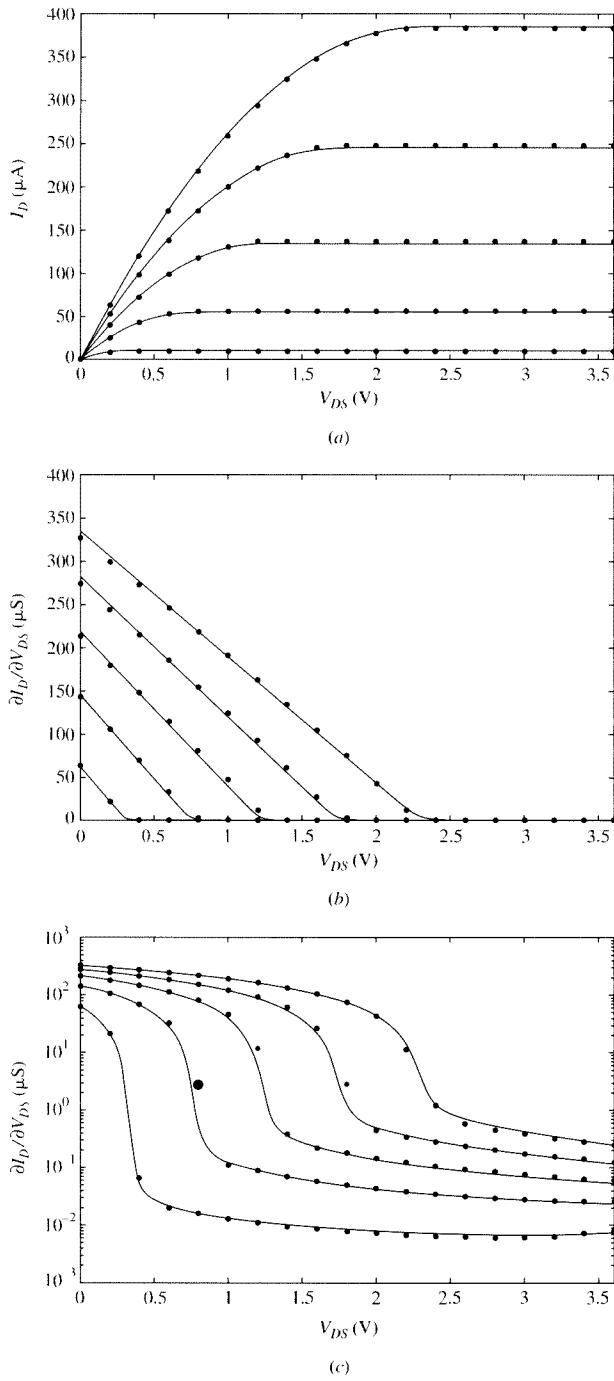
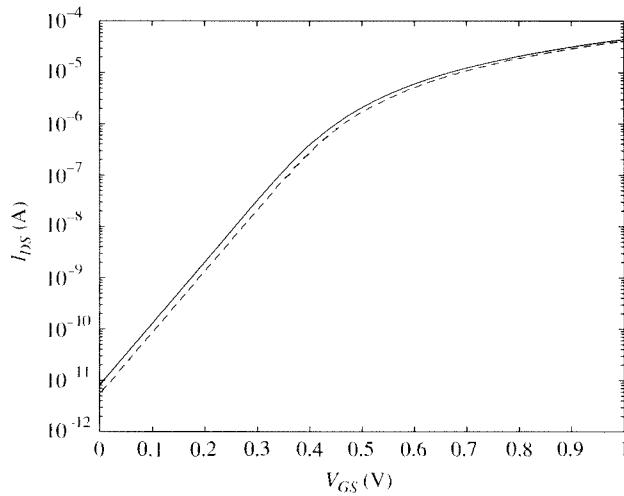


FIGURE 10.7
Model fit (solid line) to measured data (symbols). (a) I_D vs. V_{DS} , (b) dI_D/dV_{DS} vs. V_{DS} , (c) dI_D/dV_{DS} on a log scale vs. V_{DS} .

**FIGURE 10.8**

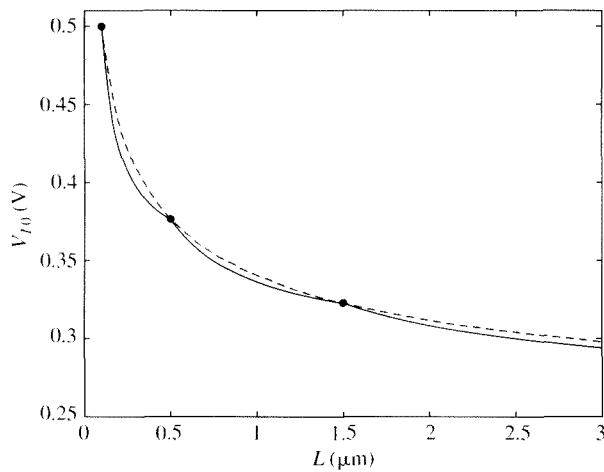
$\log I_{DS}$ vs. V_{GS} for $V_{DS} = 1.2$ V. V_{TB} differs by 15 mV; I_{DS} differs by up to 50 percent.

by this amount. The error between the two curves would certainly be caught by an appropriate error criterion, such as the mean square error in (10.7.1).

Binning Given the difficulty of accurately fitting a wide variety of device electrical behaviors over geometry, bias, and temperature, a technique known as “binning” has been developed to help improve the accuracy of modeling over geometry.^{103,104} In this approach, a sequence of .model statements (see Sec. 10.5) is provided, for the same device type, and each .model statement includes the additional parameters wmin, wmax, lmin, and lmax; these define a rectangular “bin” in the $W-L$ geometry space, for which the other parameters listed as part of that .model statement apply. In this way, different parameter sets can be provided for separate geometry ranges to help improve the overall accuracy of a model over geometry. Each parameter P in each bin is formulated as¹⁰⁴

$$P = P_0 + \frac{P_W}{W} + \frac{P_L}{L} + \frac{P_{WL}}{WL} \quad (10.7.14)$$

and the values of P_0 , P_W , P_L , and P_{WL} are calculated to ensure continuity at bin boundaries. Binning is widely used, especially by foundries, to try to help provide as accurate simulation models as possible, but has some drawbacks. Figure 10.9 shows how a binning model introduces a “cusp” effect when modeling geometry variation,^{104,105} which is clearly less desirable than a smooth and physical model for geometry dependence. More important, it has been shown¹⁰⁴ that binning does not improve overall modeling accuracy, even if the devices chosen to evaluate modeling accuracy are statistically as close as possible to the devices that were used

**FIGURE 10.9**

"Cusping" resulting from binning. Symbols are at bin boundaries; broken line is $V_{T0}(L)$; solid line is binning approximation.

for parameter extraction. Nevertheless, there is one extremely useful side effect of including binning for model parameters: even if the separation into bins is not done, and `wmin`, `wmax`, `lmin`, and `lmax` are not used and only a single `.model` card is provided for all geometries for one device type, the mapping (10.7.14) can be used to provide additional degrees of freedom in fitting experimental data. For example, analysis of DIBL generally concentrates on its length dependence, as discussed in Sec. 5.5. Experimental data show that it also depends on width, becoming smaller as W decreases. Even though this width scaling is not explicitly included in DIBL models, it can conveniently be implemented through the flexibility added because of the binning mapping (10.7.14).

Structural Effects Today, MOSFET parameter extraction is even more complex than in the past because of well proximity and stress effects and because of statistical variations (all these effects are discussed in Chap. 9). The first two cause shifts in key device parameters, such as V_T and mobility, and also in the body effect. Models are generally formulated as a core, for an ideal device not affected by well or trench proximity, and the (layout-dependent) parameter shifts are then added on top of the core model. Because there can be many well and trench edges near a device, a detailed proximity effect model can require more than 100 layout parameters. This level of complexity is much greater than warranted and slows down the process of extracting these parameters from the mask levels that define a circuit. This layout extraction process could be implemented so that, rather than determining many layout dimensions, it computed shifts in parameters such as V_{T0} , γ , and μ_0 . However this is potentially very dangerous, as it splits the modeling task into two independent parts, which significantly increases the likelihood of inconsistencies and errors. This is why an "interface" between layout extraction and models for proximity effects has

been defined, as outlined in Secs. 9.5 and 9.6. This results in a small number of layout parameters that efficiently yet reasonably embody the combined edge proximity.

Well proximity effects can be characterized separately from the core model by having test structures with the well edges far from the devices being measured. But because the width of a MOSFET in many technologies is defined by shallow trench isolation (STI), the STI stress effect cannot be removed by test structure design; it influences the electrical behavior of every device and cannot be characterized in isolation from a core device. Therefore either shallow trench isolation stress parameters should be characterized first, and the STI model taken into account during parameter extraction, or the core model should be characterized first, with stress effects turned off and STI characterization subsequently done as the difference with respect to the measured layouts. Note that this introduces a subtle confounding of parameters; if V_T varies with geometry, as does the STI effect, then determining the values of the parameters that characterize the geometry dependence of V_T without prior explicit characterization of STI (which is commonly what is done, as it is the most convenient approach) means that some portion of those parameters implicitly embodies the STI effect, even if the initial intent was to model charge sharing or nonuniform doping phenomena. Although the overall model may still fit the measured data sufficiently well, and so may seem to be acceptable, when statistical variations are added to parameters, there is not a clean separation as to the physical source of the variation in each parameter.

Measurement Errors Parameter extraction is predicated on having good measurement data, so care needs to be taken to ensure this is the case. Device measurements are affected by parasitic resistances and capacitances in the instrument cables and connectors, probe cards or needles, probe contact to on-wafer pads, and probe and interconnect metal lines on-wafer up to the device. If the expected currents are going to be large, the test structures should be connected using a Kelvin configuration to remove the effects of parasitic series resistance,¹⁰⁶ and capacitance and *s*-parameter measurements should include calibration and de-embedding structures and procedures to remove the effects of parasitic capacitances.¹⁰⁷

Data Quality and Effect of Random Variations The majority of the statistical variation of small devices is from local variation (i.e., mismatch) rather than global variation (i.e., between dice and wafers). This can make it difficult to extract model parameters that characterize trends over L and W , because the expected trends can be obscured by the random mismatch variation between devices. Rather than selecting a “golden” die site and measuring all device geometries from that one site to get the data to be used for extraction, multiple die sites need to be measured and selection of the most “typical” data needs to be done, for each device geometry, from all measured dice and data. This should be done based on the median of the data.¹⁰⁸ It may at first seem that putting down many test structures in parallel would be beneficial as this would naturally “average” out the mismatch fluctuations. This is common for measuring capacitances of small structures, as certain minimum capacitance values, much greater than those of individual devices, may need to be measured by the test equipment used, making sure the device capacitance is not swamped by pad and

other parasitic capacitances. However, such averaging cannot be used for dc data because for operation in weak inversion, where currents vary approximately exponentially with threshold voltage, the statistical distribution of I_D is log-normal, rather than normal, so the mean (which is what is measured from multiple test structures in parallel) differs from the median. Corrections to the measured data can be applied;¹⁰⁸ however these corrections are based on up to sixth-order derivatives computed from an initial assumed model, so it is preferable to directly form and use median data. In either case, the message is clear: one of the most critical steps in parameter extraction is to ensure the quality and accuracy of the data.

As the nonideal behaviors of a MOSFET can generally be expressed in terms that are inversely proportional to L and W , parameters that control these behaviors are most observable if the geometries used for the test structures are spaced roughly uniformly in reciprocal dimensions, rather than being evenly spaced in L and W . There is significant scope for simplifying and improving characterization of MOSFET behavior and extraction of MOSFET model parameters by clever design of test structures, measurement strategies, and extraction algorithms. A good resource for information on these topics is Ref. 109.

10.8 Simulation and Extraction for RF Applications

There are no simple formulas that can be used to calculate all of the parasitics for RF transistors, and as we have seen in Chap. 8, there is no single equivalent network topology that is universally used for RF modeling. As transistors become smaller and operating frequencies become higher, the metal interconnect capacitances become increasingly important to take into account. With dense, complicated interconnect structures, even the lumped circuit element approach to modeling becomes questionable; each piece of metal is in effect acting as an antenna and communicating with every other piece of metal in the interconnect and silicon structure. It is impractical to model all details of such a distributed system. Therefore, rather than trying to derive analytical formulas for the interconnect, CAD tools are used to determine the parasitics.

Analog and RF circuit design generally comprises two phases. The first is graphical entry of the schematic that defines the circuit topology, the devices used in the circuit, and their lengths and widths and other layout parameters, such as the number of gate fingers for an RF transistor. Circuit simulations are run at this schematic level to verify the basic functionality of the circuit and to help select the optimum values of biases and device geometries. At this point the final layouts of the transistors and interconnect are not known, so estimates of the parasitics are included in the circuit simulation models. These form a large and complex macroscopic model, called a sub-circuit, which includes a selected equivalent network topology of parasitic elements along with embedded scaling rules for calculating estimates of the layout parasitics as functions of the geometric layout parameters of the device. Historically parasitics included in MOSFET models were limited to current and capacitance of source-bulk and drain-bulk junctions, source and drain series resistance, gate-drain and gate-source overlap and fringing capacitance, and gate-bulk overlap capacitance. For improved RF modeling, many MOSFET models now include gate resistance and some form

of a substrate resistance network. However, which substrate network it is best to use depends on the device layout and how the substrate is contacted. For modern devices the parasitic capacitance between the source and drain contact vias can be important (see Fig. 6.16), yet is not included in core MOSFET models. So the standard practice is still to use subcircuit models for RF MOSFET modeling, even if the core model around which the subcircuit is constructed has some RF modeling capabilities.

The development of the scaling rules has generally been done based on analysis of device layouts and measurement of devices with varied lengths, widths, number of gate fingers, and whatever other designable layout parameters are defined. RF test structures include added parasitics from the metal pads used for probe contact at the top of a wafer, and from the metal lines that connect the pads to the device being tested. These test structure parasitics are characterized from additional test structures, which include the probe pads and metal connection lines, but have the actual devices removed. This allows the properties of the device alone to be “de-embedded” from the measurements of the device plus pads and metal connection lines. However, the procedure is not completely error-free, and the uncertainty in the final results tends to be greater at higher frequencies. In addition, the measured data include contributions from both the core transistor and the parasitics to be characterized, so any measurement error can get amplified if an inappropriate parasitic element is “adjusted” to make a model fit the measured data. Consequently, three-dimensional electromagnetic simulation of the interconnect structure is also used to characterize the parasitics. There are no issues with de-embedding, as no probe pads and metal connection lines are needed in the simulations, and the core MOS device can be removed to simplify the simulations. However, the geometry and material properties of the interconnect structure need to be known.

Although analog and RF design starts at the schematic level, what is manufactured is based on layout. Therefore, simulations to verify that circuit performance meets specifications should be run on the final layout. To do this, a CAD program, called a layout parameter extractor, analyzes the overall circuit layout and based on the exact routing patterns of the metal and polysilicon layers, and on the shapes of diffusion regions in the silicon, computes capacitances and resistances of the parasitics (inductances can also be extracted, but this is less common). As we have noted, in essence, every metal interconnect line or via can electromagnetically couple to every other metal line or via, and the resulting network would be too complex to analyze if all of the couplings were included. Rules and heuristics are therefore used to prune what is included in the final network representation of the parasitics, which can still be quite complex. The extracted parasitics are, apart from a full three-dimensional electromagnetic simulation, the most accurate representation possible of the actual layout and are used for final verification simulations. Note that because the schematic-based simulation models include estimates of the parasitics, these need to be turned off for simulation of a layout extracted netlist; otherwise they will be double-counted. More on RF modeling and characterization can be found in Refs. 110–113.

There is one thing to be cautious of when using a circuit simulator to evaluate performance. Quantities such as the transition frequency $f_T = \omega_T/(2\pi)$ can be very useful to quickly evaluate if a transistor is sized or biased to give adequate performance.

An operating point analysis in a circuit simulator may generate a value for f_T (this is more likely for a bipolar transistor model than a MOS transistor model), but more likely will generate values for g_m and C_g , from which f_T can be calculated using (8.6.2). However, the quantities provided by the simulator or model are generally those of the *intrinsic* transistor, and they do *not* represent the terminal-to-terminal characteristic of the overall transistor including parasitics. The values that are generated may therefore be misleading, as parasitics tend to *degrade* most figures of performance. For example, the addition of a resistance R_S in series with the source reduces the terminal-to-terminal, or extrinsic, transconductance $g_{m,\text{extr}}$ to $g_{m,\text{extr}} = g_m / (1 + g_m R_S)$. Thus, it is better to estimate the transition frequency by running a simulation, paralleling the approach discussed in Sec. 8.6. However, parasitic effects may prevent the current gain from dropping off as expected (i.e., from being inversely proportional to frequency), so it is better to extrapolate from a region of gain vs. frequency where such an inverse proportionality is valid. There is also a simple and useful simulation technique that can be used to quickly determine the terminal-to-terminal small-signal y-parameters of a MOSFET, and hence also f_T . Consider a MOSFET biased with dc voltages on all terminals, and a voltage signal applied only to the gate terminal, i.e., superimposed on the dc bias. This corresponds to the situation in Fig. 8.34, with the input source replaced by a voltage source. The gate and drain current phasors are given by

$$I_g = j\omega C_g V_{g'} \quad (10.8.1)$$

$$I_d = (g_m - j\omega C_{dg}) V_{g'} \quad (10.8.2)$$

The transition frequency ω_T is given by (8.6.2). Using the preceding two equations, we see that this can be expressed in terms of the imaginary and real parts of I_g and I_d , respectively:

$$\omega_T = \frac{g_m}{C_g} = \frac{\text{Re}\{I_d\}}{\text{Im}\{I_g\}} \quad (10.8.3)$$

This can be easily calculated, as simulators directly print the real and imaginary components of current for small-signal analysis. This test can be performed at a convenient frequency, e.g., $\omega = 1$ rad/s (an unrealistically low value for measurement, but this is completely fine for circuit simulation). Also, any convenient input amplitude can be used, e.g., 1 V. (In circuit simulators, small-signal analysis is done by linearizing the large-signal network representation of a model. Therefore, although having a 1-V gate voltage phasor may not seem to be a “small-signal” excitation, for the purpose of ac circuit simulation, there are no nonlinearities that arise from using this, because the analysis is based on linearization.)

We stress that the transition frequency computed previously is the true terminal-to-terminal ω_T of the overall MOSFET, including all parasitics, and not just of the

intrinsic transistor. (The technique works for conductance and capacitance between any excitation terminal and any terminal current and also for current excitation.)

10.9 Common MOSFET Models Available in Circuit Simulators

There are many—more than 50—MOSFET models that have been made available in one or more commercial circuit simulators, and many more models that have been developed and implemented in proprietary simulators. Most of these are legacy models that are now used rarely or never. A small number of models are in widespread use; historically, large semiconductor companies used to develop some or all of their own CAD tools, including circuit simulators, so it was feasible to develop and support proprietary models optimized for specific technologies or design needs. With the rise of separate design houses and manufacturing foundries, and the increase in the use of different simulators for various aspects of a design flow (standard circuit simulators for analog design, specialty RF, fast MOS for large digital and mixed-signal circuit verification), there emerged a need for standard models in order to mitigate the huge engineering effort needed to extract parameters for and support multiple models.

This section gives a brief overview of the general characteristics and capabilities of some of the MOSFET models in common use. Some of the effects discussed in Chaps. 4, 5, and 9 that are included in these models are:

- standard effects, such as mobility reduction, velocity saturation, and short- and narrow-channel effects
- drain-induced barrier lowering
- vertical and lateral nonuniform doping modeling
- polysilicon depletion effects
- impact ionization current
- geometric scaling
- extrinsic parasitics (details vary with the different models)

Additional effects are included in particular models, and these will be mentioned in the corresponding descriptions that follow.

10.9.1 BSIM

The BSIM (Berkeley Short-channel IGFET Model) family of models has been the workhorse of CMOS IC design for the past decade. The original BSIM1¹¹⁴ and BSIM2¹¹⁵ models were intended as replacements for the SPICE Level-3 MOSFET model,¹¹⁶ and they were designed to improve accuracy of modeling short-channel devices and overcome issues with smoothness and continuity at boundaries between regions. However both of these models were based on regional formulations and had issues with smoothness and continuity near the region boundaries. Major

semiconductor companies continued to use their own proprietary MOSFET models and did not switch to BSIM1 or BSIM2.

With the structural shift in the semiconductor industry from integrated device manufacturers toward separate “fabless” IC design houses and wafer fabrication foundries, and with the adoption of multiple specialized circuit simulators for different parts of a design flow, the burden of support for both implementation and parameter extraction for multiple models in multiple simulators became untenable. There was a critical need for a common compact MOSFET model to become the *lingua franca* of the industry, to bridge the gap between technology and design, and the BSIM3 model,^{117,118} and later the BSIM4 model,¹¹⁹ stepped in to meet this need and were selected as standard models by the Compact Model Council.¹²⁰ It is likely that more ICs have been designed with these models than with any other compact MOSFET model.

Several books are available on the BSIM3 and BSIM4 models,^{26,121,122} and details of these models are widely known by characterization engineers and designers. The basis of BSIM3 and BSIM4 is the source-referenced strong-inversion non-saturation model of Sec. 4.7.3. A very large number of effects are implemented in BSIM models through V_T , through α , and through field-dependent mobility, which is available in several selectable forms and with more bias dependencies than in the simple form 4.11.31. Output conductance modeling is done via a multicomponent Early voltage approach (see Sec. 5.3) that includes CLM, DIBL, and the body current-induced body effect. The last effect can be modeled explicitly by including a separate body resistor, and any body current will then cause the body bias to change and thereby alter the threshold voltage via the body effect. However, introducing resistors in a model adds nodal voltages that the simulator must solve for, which increases simulation time, so implicitly including the effect via the Early voltage improves the computational efficiency of BSIM.

In addition to the effects in the list given earlier, BSIM3 and BSIM4 include RF and non-quasi-static modeling. The BSIM4 model also includes gate current and gate induced drain and source leakage, as well as STI and WPE modeling. The STI and WPE models in BSIM4 have become the industry standards and have been adopted by other models. The layout extraction instance parameters (Secs. 9.4 and 9.5) were defined for BSIM4. The unified $1/f$ noise model described in Sec. 7.5, which has also been widely adopted, stems from BSIM.

The BSIM family represents the pinnacle of development of source-referenced threshold voltage-based models. However its Achilles’ heel stems from this basic formulation: it leads to a lack of proper symmetry or, equivalently, a singularity at $V_{DS} = 0$. That is, in BSIM, higher-order derivatives do not exist at $V_{DS} = 0$, and derivatives evaluated for very slightly positive and very slightly negative V_{DS} are significantly different. This is of little concern for modeling digital CMOS circuits, or for modeling analog CMOS circuits operating in saturation, for which BSIM3 and BSIM4 have proven extremely effective. However it is of crucial importance for RF CMOS circuits, such as passive mixers, MOSFET-C filters, and other precision circuits that involve MOSFETs with varying drain-source voltages that pass through $V_{DS} = 0$.

The BSIM3 and BSIM4 models are developed and released in C code for the SPICE3F5 simulator,¹²³ and complete code and documentation are available from Ref. 119.

10.9.2 EKV

The EKV (Enz-Krummenacher-Vittoz) MOSFET model^{28,47,124–126} is unique among mainstream MOSFET models in that it was developed by experienced circuit designers and emphasized accurate modeling of characteristics important for analog and RF IC design. The EKV model was primarily developed for the design of low-power analog ICs. These operate in weak and moderate inversion, i.e., precisely where the prevalent strong-inversion nonsaturation-based models have the least physical basis and accuracy (see Sec. 10.4); EKV was developed to address this deficiency.

Initially EKV was formulated based on interpolation between weak and strong inversion,^{47,53} and since version 2.8, it has been reformulated as an inversion-charge-based model, in terms of Q'_{I0} and Q'_{IL} (Sec. 4.4.4).¹²⁴ These are the fundamental quantities that are computed for model evaluation. Version 3.0,¹²⁵ EKV3, has kept this inversion-charge basis but has added modeling of many effects needed for modern devices. In addition to the effects mentioned in the list at the beginning of this section, EKV3 includes charge-based models for mobility reduction, velocity saturation, CLM, and effective mobility; QM effects; gate tunneling current; gate and drain noise correlation; and bias-dependent overlap and inner fringing capacitances.

Non-quasi-static effects are modeled via channel segmentation. Shallow trench isolation MOSFETs can exhibit a parasitic conduction path along the sides of the channel, because the fringing field causes the surface to become weakly inverted there. This parasitic conduction path can show up as a “hump” in the I_D - V_{GS} characteristics in weak or moderate inversion and can be modeled by EKV3.

The core of EKV is still the inversion-charge-based approach of (4.4.21) and (4.4.22), which give emphasis on accurate modeling in weak and moderate inversion, especially for g_m/I_D . The EKV3 model was developed and released in Verilog-A, and online information is available from Ref. 126.

10.9.3 PSP

The PSP MOSFET model^{30,127} (the name is not an acronym, but emphasizes the symmetry of the model) is a surface potential-based MOSFET model. It was formed by merging and developing the best features of two other advanced surface potential-based MOSFET models, SP¹²⁸ and MM11,¹²⁹ and is a Compact Model Council standard model. Two key features that distinguish the core of PSP from other MOSFET models are that it is based on the symmetric linearization approach, reviewed in Chap. 4, and that it uses a highly accurate noniterative solution for the surface potential.

The intrinsic PSP model solution for ψ_s is based on the surface potential equation

$$(V_{GB} - V_{FB} - \psi_s)^2 = \gamma^2 \left[\phi_t e^{-\psi_s/\phi_t} + \psi_s - \phi_t + e^{-(2\phi_t + V_{FB})/\phi_t} (\phi_t e^{\psi_s/\phi_t} - \psi_s - \phi_t - \chi(\psi_s)) \right] \quad (10.9.1)$$

which follows from the expression (3.2.11) for the charge in the semiconductor under the oxide, with the addition of the term $\chi(\psi_s)$. This term only affects ψ_s near flat-band, when the electron concentration is negligible, so it has essentially no effect on

modeled current and charges. However, using an appropriate form for $\chi(\psi)$ ³⁰ enables setting $\psi_{SL} = \psi_{s0}$ in accumulation, thereby eliminating the need to compute ψ_{SL} , yet it maintains continuity of capacitance coefficients at flat-band. This is a subtle modification, but important for a practical compact MOSFET model. Some aspects of the modeling of nonideal effects are handled through the “ideal” expression (10.9.1) by introducing physically based bias and/or geometry dependencies into quantities such as the flatband voltage and body effect coefficient.

The intrinsic core PSP model includes the effects in the list at the beginning of this section, as well as gate current (based on physical x/L partitioning), impact ionization current, STI and WPE, QM effects, bias-dependent overlap capacitances, and gate and drain noise correlation. The model also includes an option of using either physical geometry scaling rules or phenomenological binning scaling rules.

Because an original driver for PSP development was RF CMOS circuit simulation, the mobility models used in PSP were carefully formulated for accurate modeling of g_m/I_D and of distortion.¹³⁰ In addition, PSP provides detailed and accurate models for noise,¹³¹ including induced gate noise and the correlation between drain and gate noise, and for non-quasi-static effects. The non-quasi-static version of the model is based on spline collocation.^{68–70} In this procedure, the continuity equation for the inversion charge (10.4.8) is solved for at a number of points along the channel, self-consistently accounting for the full complexity of the physical effects included in PSP. Not only does this physically model the time dependence of the change in charge distribution in the channel, with no added model parameters, but the number of spline collocation points can be varied to allow a trade-off between accuracy and computational efficiency. Note that the spline collocation points are internal to the model; the technique is more computationally efficient than a multi-segment model.

For the overlap charge and gate tunneling models, the surface potential is computed for each of the source and drain overlap regions. For all reasonable biases, there are no holes in the overlap regions of an *n*MOS device, so for computational efficiency, a simplified surface potential equation, that includes only the electrons (the “accumulation” charge), is formulated and solved analytically. The bias dependence of the overlap capacitances is determined physically from the bias dependence of the surface potential in the overlap regions, with no empirical fitting parameters.

The junction diode model used in PSP is JUNCAP2,¹³² which at the time of writing is the most advanced junction model available. This model includes separate area, gate edge, and isolation edge components for charges and currents. The current model includes physical models for the ideal, Shockley-Read-Hall recombination/generation, trap-assisted-tunneling, band-to-band tunneling, and avalanche components of current. Because of its comprehensive physical basis, JUNCAP2 is accurate over bias and temperature; for improved computational efficiency, a simplified version, JUNCAP2Express, which works with the same parameter set as JUNCAP2, is also provided.¹³⁰

The PSP model has been developed and released in Verilog-A and has also been compiled in dynamically linked libraries. Complete code and documentation are available from Refs. 133 and 134. A PSP-compatible MOS varactor model has also been standardized by the Compact Model Council.¹³⁵

10.9.4 Other Models

We have described above three main CAD models that are available in several widely used simulators at the time of writing. We note that other circuit simulation models exist which have undergone considerable development. One such model is ACM,²⁷ which is conceptually related to the EKV model. Another is HiSIM (Hiroshima University Semiconductor Technology Academic Research Center IGFET Model) MOSFET model,^{29,52,136,137} a surface potential based MOSFET model that includes modeling of many physical effects. There have been several different versions of the model released; the interested reader can consult the above references.

References

1. L. W. Nagel, *SPICE2: A Computer Program to Simulate Semiconductor Circuits*, Memorandum no. ERL-M520, Electronics Research Laboratory, University of California, Berkeley, May 1975.
2. C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer-Verlag, Vienna, 1989.
3. S. E. Laux and M. V. Fischetti, "Transport models for advanced device simulation—Truth or consequences?" *Proceedings IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 27–34, October 1995.
4. T. Grassler, T.-W. Tang, H. Kosina, and S. Selberherr, "A review of hydrodynamic and energy-transport models for semiconductor device simulation," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 251–274, 2003.
5. S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Vienna, 1984.
6. P. Bendix, "SPICE model quality: Process development viewpoint," *IEEE International Symposium on Quality Electronic Design*, pp. 477–481, March 2001.
7. P. Yang and P. Chatterjee, "Statistical modeling of small geometry MOSFET," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 286–289, December 1982.
8. N. Herr, B. Garbs, and J. J. Barnes, "A statistical modeling approach for simulation of MOS VLSI circuit designs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 290–293, December 1982.
9. P. Cox, P. Yang, S. S. Mahant-Shetti, and P. Chatterjee, "Statistical modeling for efficient parametric yield estimation of MOS VLSI circuits," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 471–478, 1985.
10. C. J. B. Spanos and S. W. Director, "Parameter extraction for statistical IC process characterization," *IEEE Transactions on Computer-Aided Design*, vol. CAD-5, pp. 66–78, 1986.
11. P. Yang, D. E. Hocevar, P. F. Cox, C. Machala, and P. K. Chatterjee, "An integrated and efficient approach for MOS VLSI statistical circuit design," *IEEE Transactions on Computer-Aided Design*, vol. CAD-5, pp. 5–14, 1986.
12. M. J. B. Bolt, A. Trip, and H. J. Verhagen, "Statistical worst-case MOS parameter extraction," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 211–215, March 1989.
13. J. A. Power, A. Mathewson, and W. A. Lane, "MOSFET statistical parameter extraction using multivariate statistics," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 209–214, March 1991.
14. J. A. Power, B. Donnellan, K. Burke, K. Moloney, A. Mathewson, and W. A. Lane, "Generation of MOS model parameters covering statistical process variations," *Proceedings of the 25th European Solid State Device Research Conference*, p. 97, 1993.
15. J. A. Power, A. Mathewson, and W. A. Lane, "An approach for relating model parameter variabilities to process fluctuations," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, vol. 6, pp. 63–68, 1993.
16. C. C. McAndrew, "Statistical modeling for circuit simulation," *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 357–362, March 2003.

17. J. B. Shyu, G. Temes, and F. Krummenacher, "Random error effects in matched MOS capacitors and current sources," *IEEE Journal of Solid-State Circuits*, vol. SC-19, pp. 948–959, 1984.
18. S. Inohira, T. Shimmi, M. Nagata, T. Toyabe, and K. Iida, "A statistical model including parameter matching for analog integrated circuits simulation," *IEEE Transactions on Computer-Aided Design*, vol. CAD-4, pp. 621–628, 1985.
19. K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterization and modeling of mismatch in MOS transistors for precision analog design," *IEEE Journal of Solid-State Circuits*, vol. SC-21, pp. 1057–1066, 1986.
20. M. Peigrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistor," *IEEE Journal of Solid-State Circuits*, vol. SC-24, pp. 1433–1439, 1989.
21. C. Michael and M. Ismail, "Statistical modeling of device mismatch for analog MOS integrated circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-27, pp. 154–166, 1992.
22. P. Drennan and C. C. McAndrew, "Understanding MOSFET mismatch for analog design," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 3, pp. 450–456, 2003.
23. R. S. Vogelsong and C. Brzezinski, "Simulation of thermal effects in electrical systems," *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, pp. 353–356, March 1989.
24. G. Gildenblat, H. Wang, T.-L. Chen, X. Gu, and X. Cai, "SP: An advanced surface-potential-based compact MOSFET model," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1394–1406, 2004.
25. W. Liu, X. Jin, J. Chen, M.-C. Jeng, Z. Liu, Y. Cheng, K. Chen, et al., *BSIM3v3.2 MOSFET Model User's Guide* [online]. Available at <http://www-device.eecs.berkeley.edu/~bsim3/>. Accessed Oct. 1, 2009.
26. W. Liu and C. Hu, *BSIM4: Theory and Engineering of MOSFET Modeling for IC Simulation*, World Scientific, Hackensack, N.J., 2008.
27. C. Galup-Montoro, M. C. Schneider, A. I. A. Cunha, F. R. de Sousa, H. Klimach, and O. F. Siebel, "The advanced compact MOSFET (ACM) model for circuit analysis and design," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 519–526, September 2007.
28. C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling: The EKV Model for Low-Power and RF IC Design*, Wiley, Hoboken, N.J., 2006.
29. M. Miura-Mattausch, H. J. Mattausch, and T. Ezaki, *The Physics and Modeling of MOSFETs: Surface-Potential Model HiSIM*, World Scientific, Hackensack, N.J., 2008.
30. G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1979–1993, 2006.
31. T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, 1990.
32. T. Sakurai, "Alpha power-law MOS model," *IEEE Solid-State Circuits Society Newsletter*, vol. 9, no. 4, pp. 4–5, 2004.
33. K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang, and J. D. Meindl, "A physical alpha-power law MOSFET model," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 10, pp. 1410–1414, 1999.
34. J. Barby, J. Vlach, and K. Singhal, "Polynomial splines for MOSFET model approximation," *IEEE Transactions on Computer-Aided Design*, vol. 7, pp. 557–565, 1988.
35. H. K. Dirks and K.-M. Eickhoff, "Numerical models and table models for MOS circuit analysis," *Proceedings of the International Conference on Numerical Analysis of Semiconductor Devices*, J. J. H. Miller (editor), pp. 13–24, June 1985.
36. W. M. Coughran, E. H. Grosse, and D. J. Rose, "Variation diminishing splines in simulation," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 2, pp. 696–705, 1986.
37. E. J. Prendergast, P. Lloyd, and H. K. Dirks, "The extraction of terminal charges from two-dimensional simulation of MOS transistors," *International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 6, no. 2, pp. 107–114, 1987.
38. W. M. Coughran, W. Fichtner, and E. Grosse, "Extracting transistor charges from device simulations by gradient fitting," *IEEE Transactions on Computer-Aided Design*, vol. 8, no. 4, pp. 380–394, 1989.
39. Y. P. Tsividis and K. Suyama, "MOSFET modeling for analog circuit CAD: Problems and prospects," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 210–216, 1994.

40. G. Groenewold and W. J. Lubbers, "Systematic distortion analysis for MOSFET integrators with use of a new MOSFET model," *IEEE Transactions on Circuits and Systems—II*, vol. 41, pp. 569–580, September 1994.
41. E. A. M. Klumperink, C. H. J. Mensik, and P. M. Stroet, "Comment on low-voltage CMOS transdundance cell based on parallel operation of triode and saturation transconductors," *Electronics Letters*, vol. 30, pp. 1824–1825, 1994.
42. R. van Langevelde and F. M. Klaassen, "Effect of gate-field dependent mobility degradation on distortion analysis in MOSFET," *IEEE Transactions on Electron Devices*, vol. ED-44, no. 11, pp. 2044–2052, 1997.
43. R. van Langevelde and F. M. Klaassen, "Accurate drain conductance modeling for distortion analysis in MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 313–316, December 1997.
44. P. Wambaeq and W. Sansen, *Distortion Analysis of Analog Integrated Circuits*, Kluwer, Dordrecht, The Netherlands, 1998.
45. C. C. McAndrew, "Practical modeling for circuit simulation," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, pp. 439–448, 1998.
46. P. Bendix, P. Rakers, P. Wagh, L. Lemaitre, W. Grabinski, C. C. McAndrew, X. Gu, and G. Gild-enblat, "RF distortion analysis with compact MOSFET models," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 9–12, October 2004.
47. C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 9, pp. 83–114, 1995.
48. A. I. A. Cunha, O. C. Gouveia-Filho, M. C. Schneider, and C. Galup-Montoro, "A current-based model for the MOS transistor," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1608–1611, June 1997.
49. A. R. Boothroyd, S. W. Tarasewicz, and C. Slaby, "MISNAN—A physically based continuous MOSFET model for CAD applications," *IEEE Transactions on Computer-Aided Design*, vol. 10, pp. 1512–1529, 1991.
50. B. Iñiguez and E. G. Moreno, "A physically based C_x -continuous model for small-geometry MOSFETs," *IEEE Transactions on Electron Devices*, vol. 42, pp. 283–287, 1995.
51. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An explicit physical model for the long-channel MOS transistor including small-signal parameters," *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
52. M. Miura-Mattausch, U. Feldmann, A. Rahm, M. Bollu, and D. Savignac, "Unified complete MOSFET model for analysis of digital and analog circuits," *IEEE Transactions on Computer-Aided Design*, vol. 15, pp. 1–7, 1996.
53. J. Oguey and S. Cserveny, "MOS modelling at low current density," Presented at summer course on Process and Device Modelling, ESAT Leuven-Heverlee, Belgium, June 1983.
54. G. T. Wright, "Physical and CAD models for the implanted-channel VLSI MOSFET," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 4, pp. 823–833, 1987.
55. B. Ankele, W. Holzl, and P. O'Leary, "Enhanced MOS parameter extraction and SPICE modeling for mixed analogue and digital circuit simulation," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 73–78, March 1989.
56. S. Veeraraghavan, "SSIM: A new charge-based MOSFET model," presented at the MCNC Circuit Simulation Workshop, November 1990.
57. C. C. McAndrew, B. K. Bhattacharrya, and O. Wing, "A single piece C_x -continuous MOSFET model including subthreshold conduction," *IEEE Electron Device Letters*, vol. 12, no. 10, pp. 565–567, 1991.
58. K. Joardar, K. K. Gullapalli, C. C. McAndrew, M. E. Burnham, and A. Wild, "An improved MOSFET model for circuit simulation," *IEEE Transactions on Electron Devices*, vol. 45, no. 1, pp. 134–148, 1998.
59. G. T. Wright, "Threshold modelling of MOSFETs for CAD of CMOS-VLSI," *Electronics Letters*, vol. 21, no. 6, pp. 223–224, 1985.

60. P. Yang, B. D. Epler, and P. K. Chatterjee, "An investigation of the charge conservation problem for MOSFET circuit simulation," *IEEE Journal of Solid-State Circuits*, vol. SC-18, no. 1, pp. 128–138, 1983.
61. M. A. Cirit, "The Meyer model revisited: Why is charge not conserved?" *IEEE Transactions on Computer-Aided Design*, vol. 8, no. 10, pp. 1033–1037, October 1989.
62. J. G. Fossum, H. Jeong, and S. Veeraraghavan, "Significance of the channel-charge partition in the transient MOSFET model," *IEEE Transactions on Electron Devices*, vol. ED-33, pp. 1621–1623, 1986.
63. M. F. Sevat, "On the channel charge division in MOSFET modeling," *Proceedings of the IEEE International Conference on Computer-Aided Design*, pp. 208–210, November 1987.
64. M. Chan, K. Hui, R. Neff, C. Hu, and P. K. Ko, "A relaxation time approach to model the non-quasi-static transient effects in MOSFETs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 169–172, December 1994.
65. K. Machida, D. Navarro, M. Miyake, R. Inagaki, N. Sadachika, T. Ezaki, H. J. Mattausch, and M. Miura-Mattausch, "Efficient non-quasi-static MOSFET model for both time-domain and frequency-domain analysis," *Proceedings of the Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems*, pp. 73–76, January 2006.
66. H. Wang, T.-L. Chen, and G. Gildenblat, "Quasi-static and nonquasi-static compact MOSFET models based on symmetric linearization of the bulk and inversion charges," *IEEE Transactions on Electron Devices*, vol. 50, no. 11, pp. 2262–2272, 2003.
67. P. Mancini, C. Turchetti, and G. Masetti, "A non-quasi-static analysis of the transient behavior of the long-channel MOST valid in all regions of operation," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 2, pp. 325–334, 1987.
68. D. H. Kwon, Y. S. Yu, K. H. Kim, T. W. Yoon, and S. W. Hwang, "A new CAD-compatible non-quasi-static MOS transient model with direct inclusion of depletion charge variations," *Journal of the Korean Physical Society*, vol. 33, pp. 192–195, 1998.
69. S. W. Hwang, T.-W. Yoon, D. H. Kwon, Y. S. Yu, and K. H. Kim, "A physics-based, SPICE (Simulation Program with Integrated Circuit Emphasis)-compatible non-quasi-static MOS (metal-oxide-semiconductor) transient model based on the collocation method," *Japanese Journal of Applied Physics*, vol. 37, no. 2A, pp. L119–L121, 1998.
70. H. Wang, X. Li, W. Wu, G. Gildenblat, R. van Langevelde, G. D. J. Smit, A. J. Scholten, and D. B. M. Klaassen, "A unified nonquasi-static MOSFET model for large-signal and small-signal simulations," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2035–2043, 2006.
71. K. Singhal and J. Vlach, *Computer Methods for Circuit Analysis and Design*, 2nd ed., Springer, New York, 1993.
72. K. Kundert and A. Sangiovanni-Vincentelli, "Simulation of nonlinear circuits in the frequency domain," *IEEE Transactions on Computer-Aided Design*, vol. 5, no. 4, pp. 521–535, 1986.
73. Version 2.2 of the Verilog-A language reference manual (LRM) [online]. Available at <http://www.eda.org/verilog-ams/htmlpages/compact.html>. Accessed Oct. 1, 2009.
74. L. Zhou, B. P. Hu, B. Wan, and C.-J. Shi, "Rapid BSIM model implementation in VHDL-AMS/Verilog-AMS and MCAST compact model compiler," *Proceedings of the IEEE International Systems-on-Chip Conference*, pp. 285–286, September 2003.
75. N. Arora, *MOSFET Modeling for VLSI Circuit Simulation-Theory and Practice*, World Scientific, Hackensack, N.J., 2007.
76. K. K. Ng and J. R. Brews, "Measuring the effective channel length of MOSFETs," *IEEE Circuits and Devices Magazine*, vol. 6, pp. 33–38, 1990.
77. B. J. Sheu and P. K. Ko, "A capacitance method to determine channel lengths for conventional and LDD MOSFETs," *IEEE Electron Device Letters*, vol. EDL-5, pp. 491–493, 1984.
78. Y. Taur, D. S. Zicherman, D. R. Lombardi, P. J. Restle, C. H. Hsu, H. I. Hanafi, M. R. Wordeman, B. Divari, and G. G. Shahidi, "A new 'shift and ratio' method for MOSFET channel-length extraction," *IEEE Electron Device Letters*, vol. 13, no. 5, pp. 267–269, 1992.
79. P. R. Karlsson and K. O. Jeppson, "An efficient parameter extraction algorithm for MOS transistor models," *IEEE Transactions on Electron Devices*, vol. 39, pp. 2070–2076, 1992.

80. C. C. McAndrew and P. A. Layman, "MOSFET effective channel length, threshold voltage and series resistance determination by robust optimization," *IEEE Transactions on Electron Devices*, vol. 39, no. 10, pp. 2298–2311, 1992.
81. M. Bucher, C. Lallement, and C. C. Enz, "An efficient parameter extraction methodology for the EKV MOST model," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 145–150, 1996.
82. P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.
83. J. E. Dennis, Jr., D. M. Gay, and R. E. Welsch, "An adaptive nonlinear least-squares algorithm," *ACM Transactions on Mathematical Software*, vol. 7, no. 3, pp. 348–368, 1981; also [online]. Available at <http://www.netlib.org/port/dn2fb.f>. Accessed Oct. 1, 2009.
84. K. Doganis and D. L. Scharfetter, "General optimization and extraction of IC device model parameters," *IEEE Transactions on Electron Devices*, vol. ED-30, no. 9, pp. 1219–1228, 1983.
85. M. S. Sharma and N. D. Arora, "OPTIMA: A nonlinear model parameter extraction program with statistical confidence region algorithms," *IEEE Transactions on Computer-Aided Design*, vol. 12, no. 7, pp. 982–987, 1993.
86. S. R. Nassif and K. Singhal, personal communication.
87. N. R. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 1981.
88. "Least Square Fitting" [online]. Available at <http://mathworld.wolfram.com/LeastSquaresFitting.html>. Accessed Oct. 1, 2009.
89. Y. Tsividis and G. Masetti, "Problems in precision modeling of the MOS transistor for analog applications," *IEEE Transactions on Computer-Aided Design*, vol. CAD-3, pp. 72–79, 1983.
90. J. L. D'Arez and R. C. Rennick, "MOSFET parameter optimization for accurate output conductance modeling," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 512–515, May 1985.
91. W. Maes, K. M. De Meyer, and L. H. Dupas, "SIMPAR: A versatile technology-independent parameter extraction program using a new optimized fit strategy," *IEEE Transactions on Computer-Aided Design*, vol. CAD-5, no. 2, pp. 320–325, 1986.
92. H. Ikeda, "An elegant method for measuring MOST drain-source conductance in the saturation current region," *IEEE Transactions on Instrumentation and Measurement*, vol. IM-21, no. 3, pp. 234–236, 1972.
93. K. J. Yang and C. Hu, "MOS capacitance measurements for high-leakage thin dielectrics," *IEEE Transactions on Electron Devices*, vol. 46, no. 7, pp. 1500–1501, 1999.
94. J. Schmitz, F. N. Cubaynes, R. J. Havens, R. de Kort, A. J. Scholten, and L. F. Tiemeijer, "RF capacitance-voltage characterization of MOSFETs with high leakage dielectrics," *IEEE Electron Device Letters*, vol. 24, no. 1, pp. 346–348, 2003.
95. H.-G. Lee, S.-Y. Oh, and G. Fuller, "A simple and accurate method to measure the threshold voltage of an enhancement-mode MOSFET," *IEEE Transactions on Electron Devices*, vol. ED-29, no. 2, pp. 134–148, 1982.
96. H. C. de Graaff and F. M. Klaassen, *Compact Transistor Modelling for Circuit Design*, Springer-Verlag, Vienna, 1990.
97. G. J. Hu, C. Chang, and Y. T. Chia, "Gate-voltage-dependent effective channel length and series resistance of LDD MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-34, no. 12, pp. 2469–2475, 1987.
98. K. K. Ng and W. T. Lynch, "Analysis of the gate-voltage dependence series resistance of MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-33, no. 7, pp. 965–972, 1986.
99. M. Minondo, G. Gouget, and A. Juge, "New length scaling of current gain factor and characterization method for pocket implanted MOSFETs," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, pp. 263–267, March 2001.
100. T. S. Hsieh, Y. W. Chang, W. J. Tsai, and T. C. Lu, "A new Leff extraction approach for devices with pocket implants," *Proceedings of the International Conference on Microelectronic Test Structures*, pp. 15–18, March 2001.
101. A. J. Scholten, R. Duffy, R. van Langevelde, and D. B. M. Klaassen, "Compact modelling of pocket-implanted MOSFETs," *Proceedings of the European Solid-State Device Research Conference*, pp. 311–314, September 2001.

102. R. van Langevelde, A. J. Scholten, R. Duffy, F. N. Cubaynes, M. J. Knitel, and D. B. M. Klaassen, "Gate current: Modeling, ΔI , extraction and impact on RF performance," *Proceedings of the International Electron Devices Meeting*, pp. 289–292, December 2001.
103. D. Foty, *MOSFET Modeling with SPICE: Principles and Practice*, Prentice Hall, Upper Saddle River, N.J., 1997.
104. P. Bendix, D. Foty, and D. Pachura, "Practical aspects of MOS transistor model 'accuracy' in modern CMOS technology," *Proceedings of the IEEE International Conference on Electronic Circuits and Systems*, pp. 642–645, December 2004.
105. R. Rios, W.-K. Shih, A. Shah, S. Mudanai, P. Packan, T. Sandford, and K. Mistry, "A three-transistor threshold voltage model for halo processes," *Proceedings of the International Electron Devices Meeting*, pp. 113–116, December 2002.
106. A. J. Walton, "Test structure fundamentals," tutorial short course, at *IEEE International Conference on Microelectronic Test Structures*, March 2007.
107. P. J. Van Wijnen, *On the Characterization and Optimization of High-Speed Silicon Bipolar Transistors*, Cascade Microtech, Inc., Beaverton, Ore., 1995.
108. J. Watts, W. Pino, and H. Trombley, "Modeling small MOSFETs using ensemble devices," *Proceedings of the Nanotechnology Conference and Trade Show*, vol. 3, pp. 703–707, 2006.
109. [Online]. Available at http://www.sce.ed.ac.uk/research/IMNS/test_structures/publications.html. Accessed Oct. 1, 2009.
110. C. C. Enz and Y. Cheng, "MOS transistor modeling for RF IC design," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, pp. 186–201, 2000.
111. Y. Cheng, M. J. Deen, and C.-H. Chen, "MOSFET modeling for RF IC design," *IEEE Transactions on Electron Devices*, vol. 52, no. 7, pp. 1286–1303, 2005.
112. M. J. Deen and T. A. Fjeldly, *CMOS RF Modeling, Characterization and Applications*, World Scientific, Hackensack, N.J., 2002.
113. T. Ytterdal, Y. Cheng, and T. A. Fjeldly, *Device Modeling for Analog and RF CMOS Circuit Design*, Wiley, Hoboken, N.J., 2003.
114. B. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. SC-22, no. 4, pp. 558–566, 1987.
115. M.-C. Jeng, "Design and modeling of deep submicrometer MOSFETs," *Memorandum No. UCB/ERL M90/90*, University of California, Berkeley, 1990.
116. A. Vladimirescu and S. Liu, "The simulation of MOS integrated circuits using SPICE2," *Memorandum No. UCB/ERL M80/7*, University of California, Berkeley, Electronics Research Laboratory, 1980.
117. J. H. Huang, Z. H. Liu, M.-C. Jeng, P.-K. Ko, and C. Hu, "A robust physical and predictive model for deep-submicrometer MOS circuit simulation," *Memorandum No. UCB/ERL M93/57*, University of California, Berkeley, Electronics Research Laboratory, 1993.
118. Y. Cheng, M. Chan, K. Hui, M.-C. Jeng, Z. Liu, J. Huang, K. Chen, et al., "BSIM 3v3 manual (final version)," *Memorandum No. UCB/ERL M97/2*, University of California, Berkeley, Electronics Research Laboratory, 1997.
119. BSIM website [online]. Available at <http://www-device.eecs.berkeley.edu/~bsim3/>. Accessed Oct. 1, 2009.
120. [Online]. Available at <http://www.geia.org/index.asp?bid=597>. Accessed Oct. 1, 2009.
121. Y. Cheng and C. Hu, *MOSFET Modeling and BSIM3 User's Guide*, Springer, New York, 1999.
122. W. Liu, *MOSFET Models for SPICE Simulation, Including BSIM3v3 and BSIM4*, Wiley, New York, 2001.
123. [Online]. Available at <http://embedded.eecs.berkeley.edu/pubs/downloads/spice/index.htm>. Accessed Oct. 1, 2009.
124. M. Bucher, *Analytical MOS Transistor Modelling for Analog Circuit Simulation*, Ph.D. thesis, no. 2114, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 1999.
125. M. Bucher, A. Bazigos, F. Krummenacher, J.-M. Sallese, and C. C. Enz, "EKV3.0: An advanced charge-based MOS transistor model", pp. 67–95, in W. Grabinski, B. Nauwelaers, and D. Schreurs (editors), *Transistor Level Modeling for Analog/RF IC Design*, Springer, New York, 2006.
126. [Online]. Available at <http://legwww.epfl.ch/ekv/>. Accessed Oct. 1, 2009.

127. R. van Langevelde and G. Gildenblat, "PSP: An advanced surface-potential-based MOSFET model," in W. Grabinski, B. Nauwelaers, and D. Schreurs (editors), *Transistor Level Modeling for Analog/RF Design*, pp. 29–66, Springer Verlag, 2006.
128. G. Gildenblat, W. Wang, T.-L. Chen, X. Gu, and X. Cai, "SP: An advanced surface-potential-based compact MOSFET model," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1394–1406, September 2004.
129. R. van Langevelde, A. J. Scholten, and D. B. M. Klaassen, "Physical background of MOS model 11," Philips Nat. Lab. unclassified report 2003/00239, April 2003 [online]. Available at http://www.nxp.com/models/mos_models/model11/index.html. Accessed Dec. 6, 2009.
130. A. J. Scholten, G. D. J. Smit, B. A. De Vries, L. F. Tiemeijer, J. A. Croon, D. B. M. Klaassen, R. van Langevelde, X. Li, W. Wu, and G. Gildenblat, "The new CMC standard compact MOS model PSP: Advantages for RF applications," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 5, pp. 1415–1424, 2009.
131. A. J. Scholten, R. van Langevelde, L. F. Tiemeijer, and D. B. M. Klaassen, "Compact modeling of noise in CMOS," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 711–716, September 2006.
132. A. J. Scholten, G. D. J. Smit, M. D. Durand, R. van Langevelde, and D. B. M. Klaassen, "The physical background of JUNCAP2," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2098–2107, 2006.
133. [Online]. Available at <http://pspmodel.asu.edu/>. Accessed Oct. 1, 2009.
134. [Online]. Available at http://www.nxp.com/models/mos_models/psp/index.html. Accessed Oct. 1, 2009.
135. J. Victory, Z. Zhu, Q. Zhou, W. Wu, G. Gildenblat, Z. Yan, J. Cordovez, C. McAndrew, F. Anderson, J. C. J. Paasschens, R. van Langevelde, P. Kolev, R. Cherne, and C. Yao, "PSP-based scalable MOS varactor model," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 495–502, 2007.
136. M. Miura-Mattausch, N. Sadachika, D. Navarro, G. Suzuki, Y. Takeda, M. Miyake, T. Warabino, Y. Mizukane, R. Inagaki, T. Ezaki, H. J. Mattausch, T. Ohguro, T. Iizuka, M. Taguchi, S. Kumashiro, and S. Miyamoto, "HiSIM2: Advanced MOSFET model valid for RF circuit simulation," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 1994–2007, 2006.
137. D. Navarro, Y. Takeda, M. Miyake, N. Nakayama, K. Machida, T. Ezaki, H. J. Mattausch, and M. Miura-Mattausch, "A carrier-transit-delay-based nonquasi-static MOSFET model for circuit simulation and its application to harmonic distortion analysis," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2025–2034, 2006.
138. www.oup.com/tividis_mcandrew

Problems

- 10.1.** Plot I_D - V_{DS} characteristics from the alpha-power law model (10.2.1). Use $V_T = 0.4$ V, $a = 1.2$, $P_C = 0.0005$, and $P_V = 1.0$ (units are omitted). Assume that $V_{SB} = 0.0$ V, $W/L = 1.0$, and the supply voltage V_{DD} is 1.2 V, and use four evenly spaced values of gate voltage from a point somewhat above V_T to V_{DD} . How do the I_D - V_{DS} characteristics of this model compare with those of the more physical models we have considered? Besides the overall shape for a fixed V_{GS} , comment on the spacing of the curves in saturation. Why do you think such a simple model can be reasonable for modeling the switching behavior of digital circuits?
- 10.2.** Consider a MOSFET operating in nonsaturation, around $V_{DS} = 0.0$ V. For such biases, a MOSFET behaves like a resistor whose conductance G is modulated by the applied biases. If $V_D = +V_X$ and $V_S = -V_X$ are the biases on the source and drain, i.e., the transistor is driven in a balanced fashion, the drain current must be an *odd* function of V_X . The conductance G must be an *even* function of V_X . If a limiting function $V_{DS,\text{eff}}(V_{DS})$ is used to smoothly limit V_{DS} to V'_{DS} , to incorporate the effects of velocity saturation, what

type of function must $V_{DS,\text{eff}}(V_{DS})$ be? Evaluate (10.4.3) and (10.4.4) numerically to determine if these functions meet this requirement.

- 10.3. Download the Verilog-A code for the surface potential-based MOSFET model provided on this book's website.¹³⁸ Read and understand the code; most of it is like that of standard programming languages. The `define operator defines a *macro*, which is a textual substitution block (not a function or subroutine), and the back-tick operator ` is used to instantiate a macro. The *access function* $V(n1, n2)$ references the voltage between nodes $n1$ and $n2$, and the *contribution assignment* $I(n1, n2) <+ expression$ indicates that the value of *expression* is a current flowing between nodes $n1$ and $n2$. Contribution statements to currents flowing between the same two nodes are additive, and charging current contributions are specified as time derivatives of charges, using the `ddt()` operator. With an understanding of this Verilog-A-specific syntax, the rest of the code should be readily understandable. If you have access to a Verilog-A compiler, write a netlist to run the Verilog-A model and generate the $I_D - V_{DS}$ characteristics.
- 10.4. In Sec. 10.7, we discussed a method for obtaining the value of V_T using measured data in the nonsaturation region. Show that, under certain conditions, V_T can also be determined using data in the saturation region, by plotting $\sqrt{I_D}$ vs. V_{GS} . [Hint: See (4.7.24).] What problems do you see with this technique, for practical devices?
- 10.5. Show that a procedure for graphically determining the values of γ , ϕ_0 , and V_{FB} in (4.7.19) from measured data is as follows: (a) Plot V_T vs. $\sqrt{V_{SB} + \phi_0}$ using an initial guess for the value of ϕ_0 (based on an assumed value for N_A). (b) If the plot turns out to be not a straight line, modify the value of ϕ_0 and try again, until approximately a straight line is obtained. Explain how, from the final plot, you can obtain the values of these parameters.
- 10.6. Assuming constant substrate doping, based on the difference between V_T at $V_{SB} = 0$ and V_T at a large value of V_{SB} , develop an iterative procedure to determine N_A . Assume that $\phi_0 = 2\phi_F + m\phi_t$, where m is specified in advance. What is a reasonable range of values to be expected for m ? Assume that the value of the oxide thickness is known.
- 10.7. We know from Chap. 9 that the substrate doping in transistors is not uniform but varies with depth into the substrate. However, assuming that the doping variation with depth is small, show how the differential quantity $\partial V_T / \partial V_{SB}$ evaluated at $V_{SB} = 0$ can be used to estimate the doping concentration. Assume that $\phi_0 = 2\phi_F + m\phi_t$, where $m = 4$.
- 10.8. To determine initial estimates of parameter values during the parameter extraction process, it is desirable to manipulate data based on simplified analyses and assumptions that lead to approximate linear relations whose slope and/or intercept enable a parameter value to be calculated. Assume the substrate doping is uniform with depth. From (4.7.20a) develop a relation that is linear between some function of $\partial V_T / \partial V_{SB}$ and V_{SB} . Assume that this relation is plotted from measured data, and that it turns out not to be a straight line. What does this mean?
- 10.9. Write an expression for N_A in terms of the slope of the relation derived in Prob. 10.8. What happens if N_A is not uniform? Can you determine the depth dependence of N_A (i.e., the vertical doping profile) from variations in this slope? Explain why or why not.
- 10.10. Consider $I_D - V_{DS}$ characteristics at low V_{DS} and $V_{SB} = 0.0$ V, as in Fig. 10.4, and assume W , L , α , and C'_{ox} are known. (a) From (10.7.5) develop an expression for the ratio of the value of I_D at the maximum V_{GS} to its value at the point of peak g_m , and from this develop an expression for direct calculation of θ . (b) Assuming that the variation of

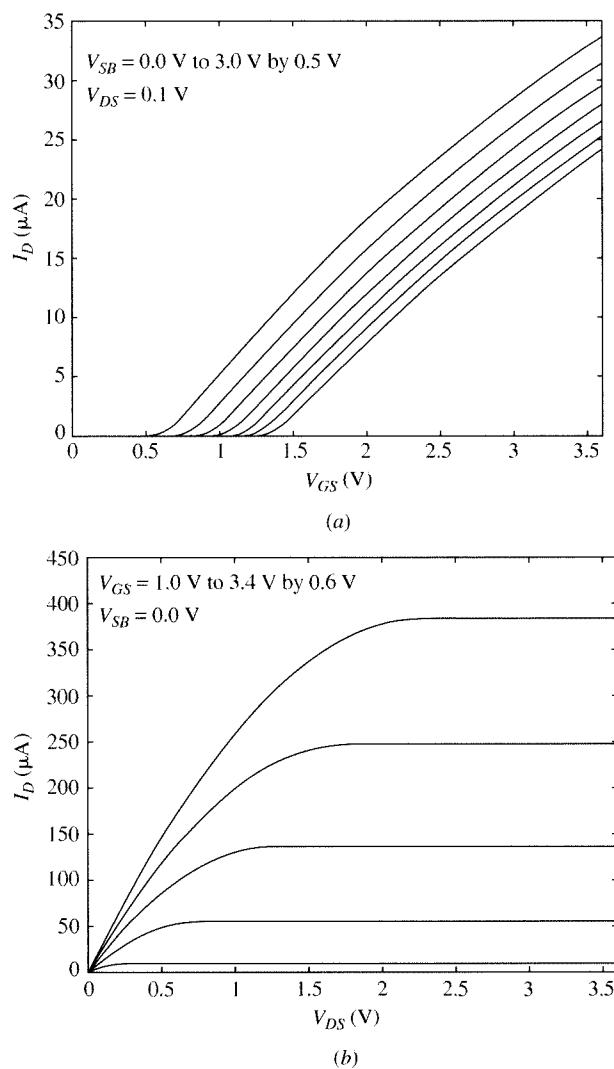


FIGURE P10.1

mobility with effective field is small, derive a relationship between the ratio of the extrapolated current from the points that define the peak of g_m to the current from (10.7.5), and show how this can be used to calculate θ . (c) Manipulate (10.7.5) to derive an expression that is linear in θ and describe a plot whose slope and intercept can be used to calculate θ and μ_0 . (d) Apply these three techniques to the data for the wide/long device provided on the book's website¹³⁸ (use $t_{ox} = 10.5$ nm and $\alpha = 1.2$). (e) Which of the techniques works best?

- 10.11.** Propose a method to determine θ_B in (10.7.5).
- 10.12.** Analyses related to the peak g_m (see Sec. 10.7) are useful because they separate threshold-related parameters from parameters related to mobility and mobility dependence on gate field. Discuss the imprecision involved in using quantities related to peak g_m for strong-inversion nonsaturation-based interpolation models.
- 10.13.** Propose a method for obtaining the value of α in (4.7.24) from measured data.
- 10.14.** Shown in Fig. P10.1 are measured characteristics for an *n*-channel transistor. Detailed numerical data can be found on the book's website.¹³⁸ Choose a *single* value for each of the quantities μ_0 , V_{T0} , γ , α , ϕ_0 , and θ so that the source-referenced, simplified strong-inversion model equations (Sec. 4.7.3), together with (4.11.31), are reasonably matched to the measurements. $W = L = 25$ μm , $t_{ox} = 10.5$ nm. Assume for simplicity that the parameter θ_B is zero. One way to approach the task is to use the I_D - V_{GS} data to determine V_{T0} , γ , and ϕ_0 ; then use the I_D - V_{GS} for $V_{SB} = 0.0$ to determine approximate values for μ_0 and θ ; and finally use the I_D - V_{DS} data to determine α and to refine the values of μ_0 and θ .
- 10.15.** Assuming that at the point of peak g_m the effect of θ is small, use (10.7.6), (10.7.10), and (10.7.11) to develop an expression for $g_{m,\max}/(C'_{ox}V_{DS})$. Show how this can be rewritten so that for three separate values of W_m and L_m it leads to a set of three linear equations that can be solved for μ_0 , ΔW , and ΔL . (*Hint:* you can solve for only two of these directly and will first need to formulate the linear equations so one unknown is some function of these parameters.) What are the restrictions on the selections of lengths and widths for the three devices, so that this procedure gives reliable results? Qualitatively, what do you think are optimum values of these geometries, and why? Why is it better in practice to use more than this minimum of three geometries?
- 10.16.** Show that (10.4.6) is equivalent to (10.4.5), yet it does not suffer from exponential overflow problems.

APPENDIX**A****Basic
Laws of Electrostatics
in One Dimension****A.1 Relations between Charge Density,
Electric Field, and Potential**

We consider the *static* case, in which charges either are fixed or move at a constant rate, so that charge densities and currents are independent of time. The results will also be valid approximately if a time dependence is involved, as long as it is sufficiently slow. We give relations for one-dimensional analysis only, i.e., for cases in which the quantities involved vary with position in only one direction (here assumed to be x). In such cases the electric field, which is a vector, will only have a component in that direction, and is commonly represented by its signed magnitude. The following symbols will be used:

$\rho(x)$ = charge density per unit volume at point x .

$\mathcal{E}(x)$ = electric field intensity at point x (in V/cm), with the field defined in the *positive x* direction.

$\psi(x)$ = electric potential at point x with respect to an arbitrary reference.

ϵ = permittivity of material; it is equal to $k\epsilon_0$, where k is the dielectric constant of the material and ϵ_0 the permittivity of free space (8.854×10^{-14} F/cm). The material is assumed to be characterized by a single value of ϵ everywhere, unless noted otherwise.

The preceding quantities are related as follows¹ (Fig. A.1):

$$\frac{d\mathcal{E}}{dx} = \frac{\rho(x)}{\epsilon} \quad (\text{A.1})$$

$$\frac{d\psi}{dx} = -\mathcal{E}(x) \quad (\text{A.2})$$

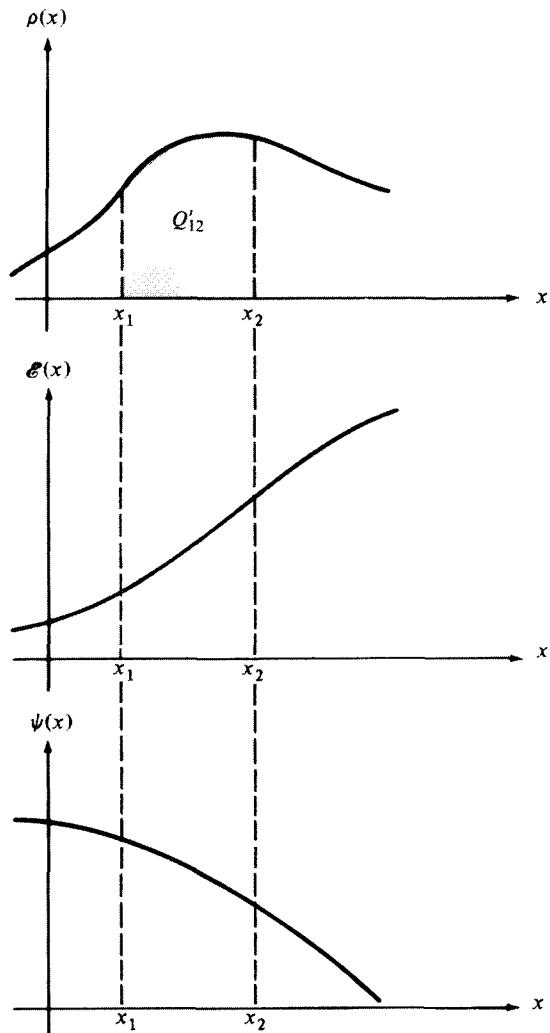


FIGURE A.1

or, in integral form:

$$\mathcal{E}(x_2) - \mathcal{E}(x_1) = \frac{1}{\epsilon} \int_{x_1}^{x_2} \rho(x) dx \quad (\text{A.1}')$$

$$\psi(x_2) - \psi(x_1) = - \int_{x_1}^{x_2} \mathcal{E}(x) dx \quad (\text{A.2}')$$

Combining (A.1) with (A.2), we obtain Poisson's equation:

$$\frac{d^2\psi}{dx^2} = - \frac{\rho(x)}{\epsilon} \quad (\text{A.3})$$

A.2 Relation between Electric Field and Charge per Unit Area

Consider a parallelepiped as shown in Fig. A.2. The charge in a vanishingly thin, vertical slice of length Δx around point x will be $\rho(x)A\Delta x$. [As stated in the beginning of this appendix, $\rho(x)$ is assumed constant over any plane perpendicular to the x axis.] The total charge Q in the parallelepiped will be the integral of $\rho(x)Adx$ from x_1 to x_2 . Dividing this by A gives the charge per unit area Q' as seen from the side i.e., on a plane perpendicular to the x direction. Thus (A.1') can be written

$$\mathcal{E}(x_2) - \mathcal{E}(x_1) = \frac{Q'_{12}}{\epsilon} \quad (\text{A.4})$$

where Q'_{12} is the charge per unit area contained between two vertical planes at x_1 and x_2 , as shown in Fig. A.1. The previous equation is referred to as *Gauss' law*.

If points x_1 and x_2 in (A.4) belong to two different materials joined at a plane perpendicular to the x dimension at a point x_0 somewhere between x_1 and x_2 , and if ϵ_1 and ϵ_2 are the corresponding permittivities, we have, in lieu of (A.4),¹

$$\epsilon_2 \mathcal{E}(x_2) - \epsilon_1 \mathcal{E}(x_1) = Q'_{12} \quad (\text{A.5})$$

A.3 Discontinuities in Electric Field

As can be deduced from (A.5), by letting x_1 and x_2 approach the same value x_0 , the electric field can be discontinuous at $x = x_0$ because of (1) a change of permittivity at an interface and/or (2) a charge sheet of zero thickness located at point x_0 , with a nonzero charge per unit area.

Thus, for example, assume that the change of permittivity from ϵ_1 to ϵ_2 occurs at point x , and that the charge density is zero at that point. If we denote by x^- and x^+ two points immediately to the left and to the right of x , respectively, (A.5) dictates

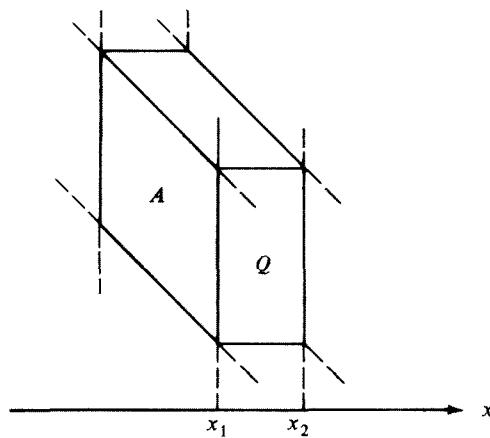


FIGURE A.2

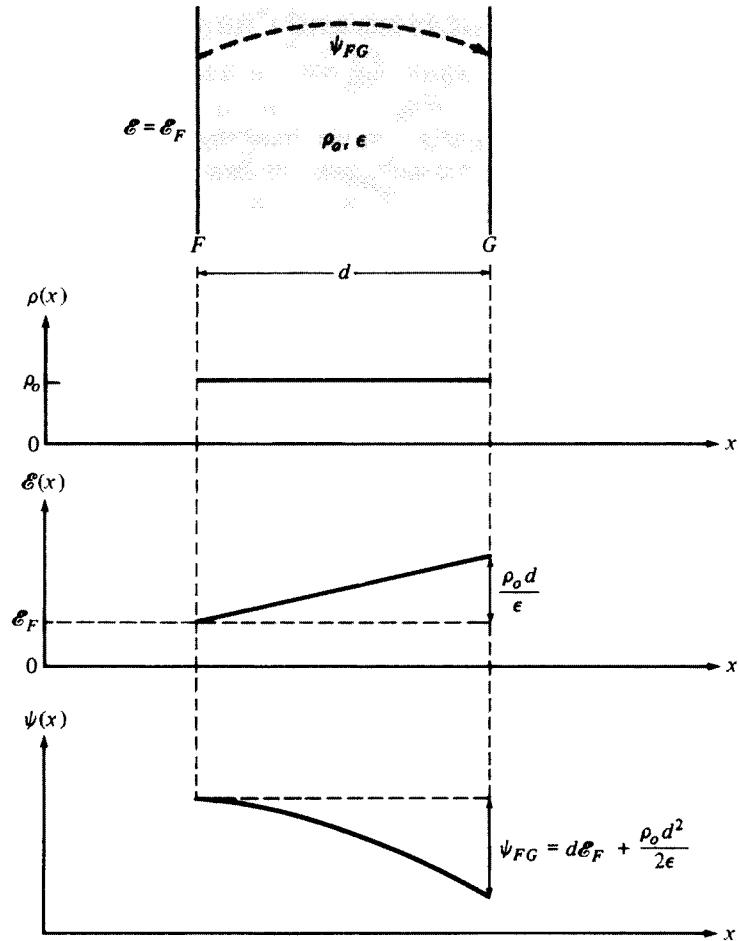


FIGURE A.3

$$\frac{E(x^+)}{E(x^-)} = \frac{\epsilon_1}{\epsilon_2} \quad (A.6)$$

From the equations given previously, one can derive the following two useful results that can be applied in several cases considered in this book.

A.4 Result 1

Consider a region characterized by a *uniform* charge density ρ_o and permittivity ϵ enclosed between two planes F and G perpendicular to the x dimension, as shown in Fig. A.3, separated by a distance d . Assume $E = E_F$ at the left plane. Then the potential drop ψ_{FG} between the two planes is given by

$$\psi_{FG} = E_F d + \frac{\rho_o d^2}{2\epsilon} \quad (A.7)$$

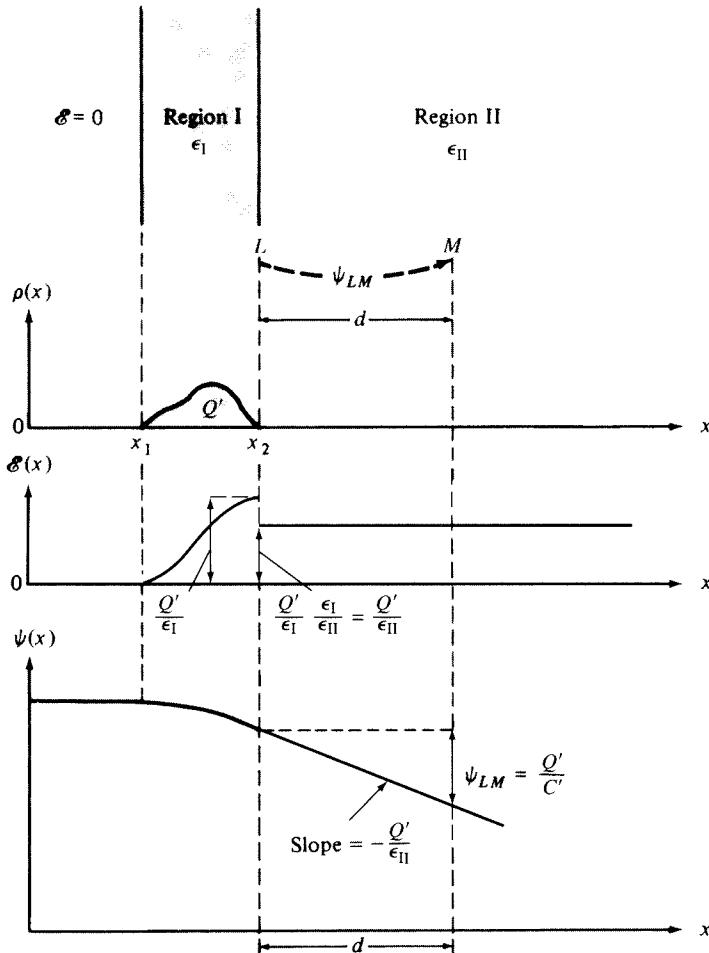


FIGURE A.4

The above result is a straightforward application of (A.1) and (A.2). The plots in Fig. A.3 are shown only for points between planes F and G .

A.5 Result 2

Consider the situation pictured in Fig. A.4. A region with permittivity ϵ_{II} and no charge in it (region II) is adjacent to a region with permittivity ϵ_I and with some charge in it (region I). The two regions are assumed separated by a plane perpendicular to the x dimension at point L as shown. Assume $\mathcal{E} = 0$ to the left of region I. Then the total potential drop from the boundary between the two regions to a point at a distance d inside region II will be

$$\psi_{LM} = \frac{Q'}{C'} \quad (\text{A.8})$$

where Q' is the charge per unit area in region I as seen from the side in Fig. A.4 (i.e., on a plane perpendicular to the x direction) and C' is the capacitance per unit area between the planes at L and M , given by

$$C' = \frac{\epsilon_{\text{II}}}{d} \quad (\text{A.9})$$

The preceding result is again an application of (A.1) and (A.2). Note that the result in (A.8) is *independent* of the details of the shape of $\rho(x)$ in region I.

Reference

1. J. D. Kraus, *Electromagnetics*, McGraw-Hill, New York, 1992.

B

Quasi-Fermi Levels and Currents

The quasi-fermi levels E_{Fn} and E_{Fp} introduced in Sec. 1.2.5 can be related to electron current and hole current, respectively, in a simple manner. Thus, consider one-dimensional current flow in the x direction, uniformly distributed throughout a cross-sectional area A .[†] Let $\psi(x)$ and $n(x)$ be the electrostatic potential and the electron concentration at x . We can express the total current due to electrons $I_n(x)$, caused by both drift and diffusion, by using (1.3.10') with (1.3.6) and (1.3.17). If the current is defined in the *positive* x direction (*opposite* from that in Fig. 1.12 or 1.15), we will have

$$I_n(x) = qA \left[-\mu_n n(x) \frac{d\psi}{dx} + D_n \frac{dn}{dx} \right] \quad (\text{B.1})$$

where μ_n and D_n are the electron mobility and the diffusion constant, respectively. As follows from (1.2.17), if $\psi(x)$ varies with x so will $E_c(x)$. Dividing both sides of that equation by Δx and letting Δx approach zero, we obtain:

$$\frac{d\psi}{dx} = -\frac{1}{q} \frac{dE_c}{dx} \quad (\text{B.2})$$

From (1.2.20b), we obtain

$$\frac{dn}{dx} = \frac{1}{kT} n(x) \left(\frac{dE_{Fn}}{dx} - \frac{dE_c}{dx} \right) \quad (\text{B.3})$$

[†] If desired, one can let A shrink to zero around a point and define a *current density* at that point as dI/dA .

Using now (B.2), (B.3), and (1.3.18) in (B.1), we obtain

$$I_n(x) = A\mu_n n(x) \frac{dE_{Fn}}{dx} \quad (\text{B.4})$$

Thus the spatial variation of the electron quasi-Fermi level depends on the *total* electron current (drift plus diffusion components). Zero total electron current implies a constant electron quasi-Fermi level, and vice versa.

In a similar manner, the total hole current can be expressed in terms of drift and diffusion components by

$$I_p(x) = qA \left[-\mu_p p(x) \frac{d\psi}{dx} - D_p \frac{dp}{dx} \right] \quad (\text{B.5})$$

with μ_p being the hole mobility and D_p the hole diffusion constant given by $D_p = \mu_p \phi_t$ in analogy with (1.3.18). The algebraic signs in (B.5) are easily discernible from the discussion in Secs. 1.3.2 and 1.3.3, remembering that the current direction considered there was opposite from the one here. Using (1.2.20a) and proceeding as previously, we obtain

$$I_p(x) = A\mu_p p(x) \frac{dE_{Fp}}{dx} \quad (\text{B.6})$$

which gives the relation between the *total* hole current (drift plus diffusion components) and the hole quasi-Fermi potential. Zero hole current implies a constant E_{Fp} , and vice versa.

Since in thermal equilibrium $E_{Fn} = E_{Fp} = E_F$ and E_F is constant (see Sec. 1.2.4.), it is clear from (B.4) and (B.6) that thermal equilibrium implies zero total electron current and zero total hole current.

C

General Analysis of the Two-Terminal MOS Structure

General Solution and Expressions Involving Acceptor and Donor Atoms

In Sec. 2.4, we have shown how to set up Poisson's equation in the general case, allowing for the presence of both acceptors and donors, and for the presence of both electrons and holes throughout the semiconductor, with distributions dictated from semiconductor physics. This equation, (2.4.16), is repeated here for convenience:

$$\frac{d^2\psi}{dy^2} = -\frac{q}{\epsilon_s} \left(p_o e^{-\psi(y)/\phi_t} - N_A + N_D - n_o e^{\psi(y)/\phi_t} \right) \quad (\text{C.1})$$

This equation can be solved as follows. Multiply both sides of this equation by $2(d\psi/dy)$; the resulting left-hand side can be recognized as $(d/dy)(d\psi/dy)^2$. Replace y by a dummy variable \hat{y} and integrate from a point deep in the bulk (theoretically at infinity, where $\psi = 0$ and $d\psi/dy = 0$) to a point y . Solve for $d\psi/dy$ at point y , and recall from (1.2.27) that $\epsilon(y) = -d\psi/dy$. The following is then obtained:

$$\mathcal{E}(y) = -\frac{d\psi}{dy} = \text{sgn}(\psi_s) \sqrt{\frac{2q}{\epsilon_s}} \sqrt{\underbrace{p_o \phi_t (e^{-\psi/\phi_t} - 1)}_{\text{Hole contribution}} + \underbrace{(N_A - N_D)\psi}_{\text{Dopant ion contribution}} + \underbrace{n_o \phi_t (e^{\psi/\phi_t} - 1)}_{\text{Electron contribution}}} \quad (\text{C.2})$$

where $\psi = \psi(y)$ and $\text{sgn}(\psi_s)$ is the “signum” function, defined as +1 for $\psi_s > 0$, zero for $\psi_s = 0$, and -1 for $\psi_s < 0$. This choice of signs corresponds to the fact that the signs of \mathcal{E} and ψ agree, as follows from the discussion in Sec. 2.4. The preceding equation is very general and is valid for both p -type and n -type substrates.

To find the total semiconductor charge per unit area Q'_c , we can apply (A.4) of Appendix A, taking point 1 at the surface and point 2 deep in the bulk, where $\mathcal{E} = 0$. This gives $-\mathcal{E}_{\text{surface}} = Q'_c/\epsilon_s$. Evaluating $\mathcal{E}_{\text{surface}}$ from (C.2) with $\psi = \psi_s$ (the surface potential), we obtain (2.4.17), repeated here:

$$Q'_c = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} \underbrace{\sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1)}}_{\text{Hole contribution}} + \underbrace{(N_A - N_D)\psi_s}_{\text{Dopant ion contribution}} + \underbrace{n_o\phi_t(e^{\psi_s/\phi_t} - 1)}_{\text{Electron contribution}} \quad (\text{C.3})$$

Using this equation in (2.4.24) gives

$$V_{GB} = V_{FB} + \psi_s + \text{sgn}(\psi_s) \frac{\sqrt{2q\epsilon_s}}{C'_{\text{ox}}} \sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1) + (N_A - N_D)\psi_s + n_o\phi_t(e^{\psi_s/\phi_t} - 1)} \quad (\text{C.4})$$

where the individual contributions are easily recognized as in (C.3).

Using the preceding equation in (2.7.7) gives

$$\begin{aligned} C'_c &= \frac{-\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} (p_o e^{-\psi_s/\phi_t} - N_A + N_D - n_o e^{\psi_s/\phi_t})}{2\sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1) + (N_A - N_D)\psi_s + n_o\phi_t(e^{\psi_s/\phi_t} - 1)}}, \quad \psi_s \neq 0 \\ &= \sqrt{q\epsilon_s(p_o + n_o)/\phi_t}, \quad \psi_s = 0 \end{aligned} \quad (\text{C.5})$$

where we have used l'Hôpital's rule to determine the value at $\psi_s = 0$, as the value of the first branch of the equation is indeterminate at that point, using the procedure outlined in Prob. 2.18.

Expressions in Terms of Equilibrium Carrier Densities Using (2.4.18) in (C.2)–(C.5) gives equally valid expressions for the general case:

$$\mathcal{E}(y) = -\frac{d\psi}{dy} = \text{sgn}(\psi_s) \sqrt{\frac{2q}{\epsilon_s}} \sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1) + (p_o - n_o)\psi_s + n_o\phi_t(e^{\psi_s/\phi_t} - 1)} \quad (\text{C.6})$$

$$Q'_c = -\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} \sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1) + (p_o - n_o)\psi_s + n_o\phi_t(e^{\psi_s/\phi_t} - 1)} \quad (\text{C.7})$$

$$V_{GB} = V_{FB} + \psi_s + \text{sgn}(\psi_s) \frac{\sqrt{2q\epsilon_s}}{C'_{\text{ox}}} \sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1) + (p_o - n_o)\psi_s + n_o\phi_t(e^{\psi_s/\phi_t} - 1)} \quad (\text{C.8})$$

$$C'_c = \frac{-\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} [p_o(e^{-\psi_s/\phi_t} - 1) - n_o(e^{\psi_s/\phi_t} - 1)]}{2\sqrt{p_o\phi_t(e^{-\psi_s/\phi_t} - 1) + (p_o - n_o)\psi_s + n_o\phi_t(e^{\psi_s/\phi_t} - 1)}} , \quad \psi_s \neq 0$$

$$= \sqrt{q\epsilon_s(p_o + n_o)/\phi_t} , \quad \psi_s = 0 \quad (\text{C.9})$$

Expressions for *p* Substrates in Terms of the Fermi Potential The corresponding expressions for *p*-type substrates in terms of Fermi potential can now be found using (2.4.20) in the above. From (C.2), we get

$$\mathcal{E}(y) = -\frac{d\psi}{dy} = \text{sgn}(\psi_s) \frac{\sqrt{2q\epsilon_s N_A}}{\epsilon_s} \sqrt{\phi_t e^{-\psi/\phi_t} + \psi - \phi_t + e^{-2\phi_F/\phi_t} (\phi_t e^{\psi/\phi_t} - \psi - \phi_t)} \quad (\text{C.10})$$

and the expressions for Q'_C , V_{GB} , and C'_c have been given in (2.4.21), (2.4.25), and (2.7.9).

Relating Potential to Depth The relation between ψ and y can be obtained from (C.6) by separating variables and integrating from a point at y to a point at the surface. This gives

$$\int_{\psi(y)}^{\psi} \frac{d\hat{\psi}}{\mathcal{E}[\hat{\psi}]} = y - y_{\text{surface}} \quad (\text{C.11})$$

where $\hat{\psi}$ is a dummy variable of integration, and $\mathcal{E}[\psi] = \mathcal{E}[\psi(y)]$ is given by the right-hand side of (C.2), (C.6), or (C.10). From this, one can determine ψ for a given y (and a given ψ_s) numerically. With $\psi(y)$ obtained in this way, one can then determine $\rho(y)$, $n(y)$, $p(y)$ from (2.4.13) to (2.4.15), and $\mathcal{E}(y)$ from (C.2), (C.6), or (C.10).

The rest of the analysis proceeds as explained in Sec. 2.4.4.

Inversion

In inversion it is easy to see that for *p*-type substrates, (C.3) reduces practically to (2.6.1), and (C.5) reduces to (2.7.11). The charges and capacitances (per unit area) corresponding to the inversion region and the depletion region can be determined as follows. In (2.6.4), we use $n(y)$ from (2.4.13) and in the result we use (1.2.13). We take y_c deep in the bulk (theoretically at infinity, where $\psi = 0$). We perform a change of variables from y to ψ . The resulting factor $(d\psi/dy)^{-1}$ in the integrand can be obtained from (C.6). This procedure gives

$$Q'_I = -qN_A e^{-2\phi_F/\phi_t} \int_0^{\psi_c} \frac{e^{\psi/\phi_t}}{\mathcal{E}[\psi]} d\psi \quad (\text{C.12})$$

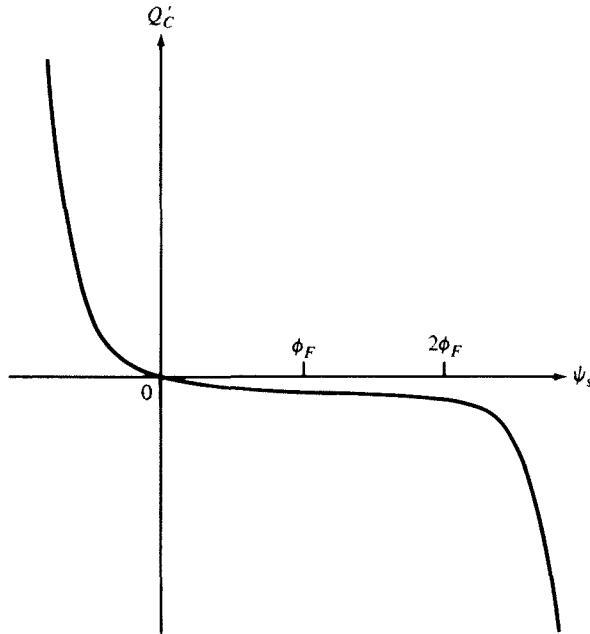


FIGURE C.1

A similar procedure gives the bulk charge per unit area Q'_B , consisting of ionized acceptor atoms and holes:

$$Q'_B = -qN_A \int_0^{\psi_s} \frac{1 - e^{-\psi/\phi_t}}{\mathcal{E}[\psi]} d\psi \quad (\text{C.13})$$

The preceding integrals can be evaluated numerically. However, determining the corresponding capacitances is easy. Using (C.12) and (C.13) in (2.7.15) and (2.7.14) and performing the differentiation results in *explicit* expressions for C'_i and C'_b . After dropping the terms that are negligible in the inversion region, we easily obtain (2.7.16) and (2.7.17).

Sometimes an intermediate level of approximation is used: the charge sheet assumption (i.e., infinitesimal thickness) for the inversion layer is taken to hold, but holes are *still* allowed to exist in the depletion region. Then $n(y)$ is taken as zero in (2.4.13) at any point below the inversion layer. Thus the sum of the last two terms in (C.1) will be zero as will the fourth term in the sum in (C.10). The *bulk* charge per unit area Q'_B can then be found by applying (A.4) as was done earlier, taking point 2 again in the bulk but point 1 immediately below the inversion layer. This gives, neglecting $\exp(-\psi_s/\phi_t)$:

$$Q'_B \approx -\sqrt{2q\epsilon_s N_A} \sqrt{\psi_s - \phi_t} \quad (\text{C.14})$$

which could have also been obtained from (2.5.1).

Finally, if the depletion approximation is used and thus not even holes are allowed in the depletion region, the only term that will be present in the right-hand side of (C.1) (for any y below the inversion layer) is N_A . Using this directly in Poisson's equation and integrating gives (2.6.7), which is the approximation widely used in this book.

Bibliography

- E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.
S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.

D

Careful Definitions for the Limits of Moderate Inversion

Two-Terminal Structure Throughout this appendix, uniform substrates are assumed. In Sec. 2.6, the lower limit of the moderate-inversion region was taken to coincide with the upper limit of weak inversion. The latter is traditionally defined at $\psi_s = 2\phi_F$. At this point, the surface electron concentration becomes equal to the bulk doping concentration N_A , as seen from (2.4.11b). Also, at this point $C'_i = C'_b$, as seen from (2.7.16) and (2.7.17). These two facts hold independently of the value of C'_{ox} . However, neither of these facts says anything about the validity of common “weak-inversion approximations” [such as $Q'_i(V_{GB})$ being an exponential] at this point. This is because the validity of such approximations cannot be discussed carefully without evoking the value of C'_{ox} . For detailed work, then, it makes practical sense to take C'_{ox} into account and redefine the onset of moderate inversion at some point beyond which common weak-inversion approximations become unacceptable. In weak inversion, and at points where C'_i is negligible (Fig. 2.23), C'_b varies little; thus, $d\psi_s/dV_{GB}$ in (2.7.21) will be approximately constant. That is necessary for an exponential dependence of Q'_i on V_{GB} , as can be seen from the development leading from (2.6.28) to (2.6.34). Significant departure from such behavior will be observed if C'_i starts becoming significant in comparison to $C'_{ox} + C'_b$ in (2.7.21). Thus, let us redefine the onset of moderate inversion as follows:¹

ϕ_{M0} and \hat{V}_{M0} are the values of ψ_s and V_{GB} , respectively, at which

$$\frac{C'_i}{C'_{ox} + C'_b} = 0.1 \quad (\text{D.1})$$

Keeping in mind that C'_b varies slowly, and using (2.7.21), it is easy to see that at this point the slope $d\psi_s/dV_{GB}$ in Fig. 2.16 drops to about 91 percent of its value

deep in weak inversion. The same is true for the slope of $\ln|Q'_I|$ vs. V_{GB} (Fig. 2.19), this slope being a measure of the “exponentiality” of $Q'_I(V_{GB})$.

Accurate evaluation of ϕ_{M0} requires using (2.7.16) and (2.7.17) in (D.1), and solving iteratively for ψ_s ; the corresponding \hat{V}_{M0} can then be found from (2.6.12). The result depends on oxide thickness and substrate doping. It can be shown that, for practical values of these parameters, the value of ϕ_{M0} as defined above will not differ by more than about $1\phi_i$ from $2\phi_F$. The corresponding value of \hat{V}_{M0} will differ from that given in (2.6.13b) by at most a few tens of millivolts. For most practical cases, we can continue using the value of $2\phi_F$ and (2.6.13b) for simplicity.

We now turn to the upper limit of moderate inversion. For detailed work, it makes practical sense to define this at a point below which common strong-inversion approximations, such as (2.6.20), become unacceptable. Let us look at the slope of the $Q'_I(V_{GB})$ plot, given by (2.7.23). Deep in strong inversion, C'_i is very large and (2.7.23) reduces to $d|Q'_I|/dV_{GB} \approx C'_{ox}$; this agrees with (2.6.20). If C'_i is not much larger than $C'_{ox} + C'_b$, $d|Q'_I|/dV_{GB}$ will be less than C'_{ox} and (2.6.20) will not hold. Thus, let us define the upper limit of moderate inversion as follows:¹ ϕ_{H0} and V_{H0} are the values of ψ_s and V_{GB} , respectively, at which

$$\frac{C'_i}{C'_{ox} + C'_b} = 10 \quad (\text{D.2})$$

From (2.7.23), it is easy to see that at this point the slope in the plot of $|Q'_I|$ vs. V_{GB} (Fig. 2.17) is reduced to about 91 percent of its theoretical maximum value of C'_{ox} and, from (2.7.21), it can be seen that the slope of the $\psi_s(V_{GB})$ plot (Fig. 2.16) drops to about 9 percent of its maximum value.

To find ϕ_{H0} one must use (2.7.16) and (2.7.17) in (D.2) and solve for ψ_s iteratively. One finds that ϕ_{H0} is several ϕ_i above ϕ_{M0} (about $6\phi_i$), the exact value being dependent on oxide thickness and substrate doping. If ϕ_{H0} is accurately known, it can be used in (2.6.12) to find V_{H0} . (Note that a small error in ϕ_{H0} will result in a large error in V_{H0} because of the exponential term in that equation; that term is now large.)

Using the previous calculations, the width of the moderate-inversion region in terms of V_{GB} can be found to vary significantly with process parameters. However, not all combinations of oxide thickness and substrate doping are practical. For example, in MOS transistor fabrication, large doping concentrations are usually combined with thin oxides. For practical cases, then, the width of the moderate inversion does not differ much; an average value is about 0.6 V at room temperature assuming a uniform substrate.

Three-Terminal Structure For the three-terminal MOS structure discussed in Chap. 3, the same rationale can be used. Thus, the surface potential ϕ_M at the boundary between weak and moderate inversion for a given V_{CB} can be found from (D.1), using in it (3.2.22) and (3.2.23) and solving iteratively for $\psi_s = \phi_M$. The solution turns out to be close to $2\phi_F$ (within about $1\phi_i$, depending on process parameters). The surface potential ϕ_H at the boundary between moderate and strong inversion can be found from (D.2) in a similar manner and turns out to be $5\phi_i$ to $6\phi_i$ higher. The corresponding V_{GB} values can be found by using the preceding surface potential values

in (3.2.21). Calculated this way, the onset of moderate inversion is very close to the value of V_{MB} given in Table 3.1 (within a few tens of millivolts). The onset of strong inversion is above this by several hundreds of millivolts. The width of moderate inversion in terms of V_{GB} tends to decrease somewhat with V_{CB} (e.g., by about 0.1 V or less, for V_{CB} increasing from 0 to 3 V, for practical devices).

Rather than finding ϕ_M and ϕ_H for a given V_{CB} , as already suggested, one can find these limits for a given V_{GB} . One can use (3.2.22) and (3.2.23) in the definitions (D.1) and (D.2), and then utilize (3.2.21) to eliminate the exponential terms. The resulting equations can be solved *explicitly* for ϕ_M and ϕ_H . These values can then be substituted back in (3.2.21) to find explicitly the corresponding V_{CB} values, V_W and V_Q respectively, for the given V_{GB} (Prob. 3.10).

Reference

1. Y. Tsividis, "Moderate inversion in MOS devices," *Solid-State Electronics*, vol. 25, pp. 1099–1104, 1982; see also Erratum, *Solid-State Electronics*, vol. 26, p. 823, 1983.

E

General Analysis of the Three-Terminal MOS Structure

General Solution and Expressions Involving Acceptor and Donor Atoms For the three-terminal structure, consider the total charge density from (1.2.21), with $n(y)$ and $p(y)$ from (3.2.5a) and (3.2.5b). We have

$$\rho(y) = q \left[p_o e^{-\psi(y)/\phi_t} - N_A + N_D - n_o e^{-V_{CB}/\phi_t} e^{\psi(y)/\phi_t} \right] \quad (\text{E.1})$$

This leads to Poisson's equation (3.2.6), repeated here for convenience:

$$\frac{d^2\psi}{dy^2} = -\frac{q}{\epsilon_s} \left(p_o e^{-\psi(y)/\phi_t} - N_A + N_D - n_o e^{-V_{CB}/\phi_t} e^{\psi(y)/\phi_t} \right) \quad (\text{E.2})$$

As has been mentioned in Sec. 3.2, V_{CB} is a constant independent of y and ψ . This equation can be solved using exactly the same procedure outlined for the two-terminal structure in Appendix C; in the various integrations involved in such a solution, e^{-V_{CB}/ϕ_t} can be treated as a multiplicative constant to n_o . Not surprisingly, then, the solution leads to the same form of equations as in Appendix C, only with n_o replaced by $n_o e^{-V_{CB}/\phi_t}$:

$$\mathcal{E}(y) = -\frac{d\psi}{dy} = \text{sgn}(\psi_s) \sqrt{\frac{2q}{\epsilon_s}} \underbrace{\sqrt{p_o \phi_t (e^{-\psi/\phi_t} - 1) + (N_A - N_D) \psi}}_{\text{Hole contribution}} + \underbrace{n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi/\phi_t} - 1)}_{\text{Dopant ion contribution}} + \underbrace{n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi/\phi_t} - 1)}_{\text{Electron contribution}} \quad (\text{E.3})$$

$$Q'_C = -\operatorname{sgn}(\psi_s) \sqrt{2q\epsilon_s} \underbrace{\sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1)}}_{\text{Hole contribution}} + \underbrace{(N_A - N_D) \psi_s}_{\text{Dopant ion contribution}} + \underbrace{n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi_s/\phi_t} - 1)}_{\text{Electron contribution}} \quad (\text{E.4})$$

$$V_{GB} = V_{FB} + \psi_s + \operatorname{sgn}(\psi_s) \frac{\sqrt{2q\epsilon_s}}{C'_{ox}} \sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (N_A - N_D) \psi_s + n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi_s/\phi_t} - 1)} \quad (\text{E.5})$$

$$\begin{aligned} C'_c &= \frac{-\operatorname{sgn}(\psi_s) \sqrt{2q\epsilon_s} [p_o e^{-\psi_s/\phi_t} - N_A + N_D - n_o e^{-V_{CB}/\phi_t} e^{\psi_s/\phi_t}]}{2\sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (N_A - N_D) \psi_s + n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi_s/\phi_t} - 1)}}, \quad \psi_s \neq 0 \\ &= \sqrt{q\epsilon_s (p_o + n_o e^{-V_{CB}/\phi_t}) / \phi_t}, \quad \psi_s = 0 \end{aligned} \quad (\text{E.6})$$

Expressions in Terms of Equilibrium Carrier Densities In Sec. 2.4 and Appendix C, we used charge neutrality at $\psi = 0$ to write (2.4.18), which was used to eliminate $N_A - N_D$ from the equations. We adopt this procedure for the present case. When $\psi_s = 0$ (flatband case), one expects charge neutrality throughout, and thus (E.1) gives (this equation will be discussed shortly)

$$p_o - N_A + N_D - n_o e^{-V_{CB}/\phi_t} = 0 \quad (\text{E.7})$$

Using this to eliminate $N_A - N_D$ in (E.3)–(E.6), we obtain

$$\mathcal{E}(y) = -\frac{d\psi}{dy} = \operatorname{sgn}(\psi_s) \sqrt{\frac{2q}{\epsilon_s}} \sqrt{p_o \phi_t (e^{-\psi/\phi_t} - 1) + (p_o - n_o e^{-V_{CB}/\phi_t}) \psi + n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi/\phi_t} - 1)} \quad (\text{E.8})$$

$$Q'_C = -\operatorname{sgn}(\psi_s) \sqrt{2q\epsilon_s} \sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (p_o - n_o e^{-V_{CB}/\phi_t}) \psi_s + n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi_s/\phi_t} - 1)} \quad (\text{E.9})$$

$$\begin{aligned} V_{GB} &= V_{FB} + \psi_s + \operatorname{sgn}(\psi_s) \frac{\sqrt{2q\epsilon_s}}{C'_{ox}} \\ &\times \sqrt{p_o \phi_t (e^{-\psi_s/\phi_t} - 1) + (p_o - n_o e^{-V_{CB}/\phi_t}) \psi_s + n_o e^{-V_{CB}/\phi_t} \phi_t (e^{\psi_s/\phi_t} - 1)} \end{aligned} \quad (\text{E.10})$$

$$C'_c = \frac{-\text{sgn}(\psi_s) \sqrt{2q\epsilon_s} [p_o(e^{-\psi_s/\phi_t} - 1) - n_o e^{-V_{CB}/\phi_t} \phi_t(e^{\psi_s/\phi_t} - 1)]}{2\sqrt{p_o \phi_t(e^{-\psi_s/\phi_t} - 1) + (p_o - n_o e^{-V_{CB}/\phi_t}) \psi_s + n_o e^{-V_{CB}/\phi_t} \phi_t(e^{\psi_s/\phi_t} - 1)}}, \quad \psi_s \neq 0$$

$$= \sqrt{q\epsilon_s (p_o + n_o e^{-V_{CB}/\phi_t}) / \phi_t}, \quad \psi_s = 0 \quad (\text{E.11})$$

The corresponding expressions for *p*-type substrates in terms of Fermi potential can be found by using (2.4.20) in the preceding equations and have been given in (3.2.8), (3.2.11), and (3.2.13).

Differences from Classical Treatment The preceding four equations do not agree exactly with classical results.^{1,2} Such results are derived based on (2.4.18) ($p_o - N_A + N_D - n_o = 0$) instead of (E.7). It has been pointed out³ that such results lead to numerical problems (e.g., a quantity under a square root can become negative, or the square of the electric field can become negative). Although such problems occur over a very small ψ_s range, they can result in serious numerical issues in computer simulation. This can be traced to the fact that, under our assumptions leading to (3.2.5a), insisting that $p_o - N_A + N_D - n_o = 0$, although correct for the two-terminal case, in the three-terminal case, it violates charge neutrality under our assumptions; indeed, it is clear that doing so violates (E.7). Because we have made the approximate assumption of constant quasi-Fermi level in deriving (E.1), if we are to use this approximate equation, we might as well make it consistent with charge neutrality at flat-band; (E.7) is the only way to ensure this. It is comforting to know that the solution that results, as given previously, is consistent with that proposed in Ref. 3, which, as mentioned there, avoids numerical problems.

Nonconstant Quasi-Fermi Level It is known that the validity of assuming a constant electron quasi-Fermi level is only an approximation. Although such an assumption is consistent with the qualitative arguments we have provided in Chap. 3 and in this appendix, such arguments use inversion concepts, and they cannot necessarily be expected to be valid in depletion and accumulation. A more general equation is

$$n(y) = n_o e^{[\psi(y)-V]/k\phi_t} = p_o e^{[\psi(y)-2\phi_t-V]/k\phi_t} \approx N_A e^{[\psi(y)-2\phi_t-V]/k\phi_t} \quad (\text{E.12})$$

where V is the actual quasi-Fermi potential split; this quantity is equal to V_{CB} at or near the inverted surface, but can otherwise be different from V_{CB} . Two-dimensional semiconductor device simulations³ show that V tends to zero in accumulation and even in inversion at places away from the surface.[†]

[†] A semiempirical approximation of the results in Ref. 3 is:

$$V = \psi_t - \phi_t \ln [e^{(\psi - V_{CB})/\phi_t} + 1 - e^{-V_{CB}/\phi_t}]$$

Using this more general approach we have, instead of (E.1):

$$\rho(y) = q \left[p_o e^{-\psi(y)/\phi_t} - N_A + N_D - n_o e^{(\psi(y)-V)/\phi_t} \right] \quad (\text{E.13})$$

This equation can now provide a very accurate charge density, if the variation of V is known.[†]

The preceding approaches do not pay much attention to what is the exact electron density in accumulation. This is because that density is completely negligible there, and its exact value is immaterial.

Expressions for p Substrates in Terms of Fermi Potential Expressions for p -type substrates in terms of Fermi potential can now be found using (2.4.20) in the previous approach. From (E.3), we get

$$\mathcal{E}(y) = -\frac{d\psi}{dy} = \text{sgn}(\psi_s) \frac{\sqrt{2q\epsilon_s N_A}}{\epsilon_s} \sqrt{\phi_t e^{-\psi/\phi_t} + \psi - \phi_t + e^{-(2\phi_F + V_{CB})/\phi_t} (\phi_t e^{\psi/\phi_t} - \psi - \phi_t)} \quad (\text{E.14})$$

and the expressions for Q'_C , V_{GB} , and C'_c have been given in (3.2.8), (3.2.11), and (3.2.13).

Relating Potential to Depth As in Appendix C, we have

$$\int_{\psi(y)}^{\psi_s} \frac{d\hat{\psi}}{\mathcal{E}[\hat{\psi}]} = y - y_{\text{surface}} \quad (\text{E.15})$$

where $\mathcal{E}[\psi]$ is given by the right-hand side of (E.3), (E.8), or (E.14).

Inversion

For the inversion region, proceeding as in Appendix C, we obtain:

$$Q'_I = -qN_A e^{(-2\phi_F - V_{CB})/\phi_t} \int_0^{\psi_s} \frac{e^{\psi/\phi_t}}{\mathcal{E}[\psi]} d\psi \quad (\text{E.16})$$

$$Q'_B = -qN_A \int_0^{\psi_s} \frac{1 - e^{-\psi/\phi_t}}{\mathcal{E}[\psi]} d\psi \quad (\text{E.17})$$

[†] Using in (E.12) the expression for V given in the previous footnote, and eliminating $N_A - N_D$ by using the classical ($p_o - N_A + N_D - n_o = 0$) (which we have already used for the same purpose for the two-terminal structure), and working as before, we obtain exactly (E.7)–(E.10) again. Note that in this case it *does* make sense to use this charge neutrality equation, as the expression used for V predicts $V = 0$ when $\psi = 0$, which, from (E.12), results in the classical charge neutrality equation.

Using (E.16) in (2.7.15) and (E.17) in (2.7.14) results in explicit expressions for C'_i and C'_b . From these expressions, after deleting terms that are negligible in inversion, we obtain (3.2.23) and (3.2.22).

References

1. H. C. Pao and C. T. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator) semiconductor transistors," *Solid-State Electronics*, vol. 9, pp. 927–937, 1966.
2. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley-Interscience, Hoboken, N.J., 2007.
3. W. Wu, T.-L. Chen, G. Gildenblat, and C. C. McAndrew, "Physics-based mathematical conditioning of the MOSFET surface potential equation," *IEEE Transactions on Electron Devices*, vol. 51, no. 7, pp. 1196–1200, 2004.

F

Drain Current Formulation Using Quasi-Fermi Potentials

In Sec. 4.3, we evaluated the drain current caused by both drift and diffusion. The current components caused by each of these two phenomena were kept separate. In this appendix, we present an alternative approach, resulting in a compact formulation that combines the two effects.¹ We consider the flow of electrons as laminar flow. That is, we assume that the total current I_{DS} is the sum of elemental currents ΔI , each flowing horizontally in an inversion layer slice parallel to the surface, having width W and depth Δy , centered at point y , as shown in Fig. F.1. In the general case, we will have for the current in each slice, allowing Δy to become a differential,

$$dI_{DS} = dI_{\text{drift}}(x,y) + dI_{\text{diff}}(x,y) \quad (\text{F.1})$$

where it should be kept in mind that our direction for the current is from right to left, as shown in Fig. F.1. The drift component is, from (1.3.12),

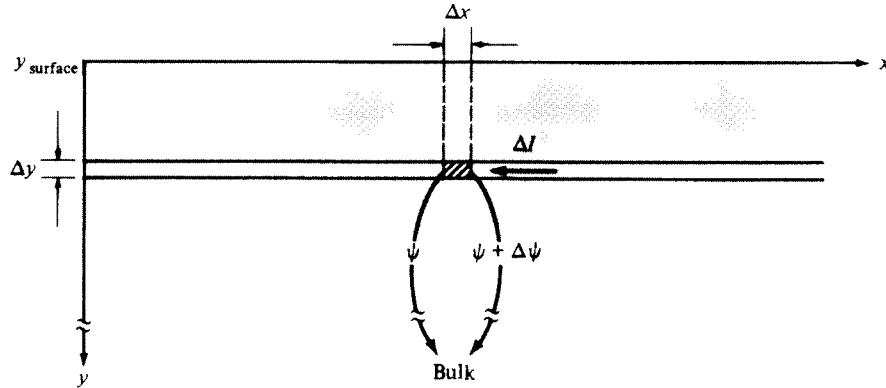
$$dI_{\text{drift}}(x,y) = (W dy) q \mu n(x,y) \frac{\partial \psi(x,y)}{\partial x} \quad (\text{F.2})$$

The diffusion component is, from (1.3.17) and (1.3.18)

$$dI_{\text{diff}}(x,y) = -(W dy) q \mu \phi_t \frac{\partial n(x,y)}{\partial x} \quad (\text{F.3})$$

$n(x,y)$ can be expressed from (3.2.5a), as

$$n(x,y) = n_o e^{[\psi(x,y) - V_{CB}(x)]/\phi_t} \quad (\text{F.4})$$

**FIGURE F.1**

where $V_{CB}(0) = V_B$ and $V_{CB}(L) = V_{DB}$. At points in the channel between source and drain $V_{CB}(x)$ takes the value required for (F.4) to provide the correct value of $n(x,y)$. As we know from Chap. 3, $V_{CB}(x)$ can be interpreted as the potential corresponding to the difference between the electron quasi-Fermi level in the inversion layer and the hole quasi-Fermi level in the bulk (Chap. 3). This quantity is assumed independent of y in the inversion layer, as discussed in Chap. 3 and Appendix E.

Differentiating (F.4) with respect to x we obtain

$$\frac{\partial n(x,y)}{\partial x} = \frac{n(x,y)}{\phi_t} \left[\frac{\partial \psi(x,y)}{\partial x} - \frac{dV_{CB}(x)}{dx} \right] \quad (\text{F.5})$$

Using this in (F.3) gives

$$dI_{\text{diff}}(x,y) = -(W dy) q \mu n(x,y) \left[\frac{\partial \psi(x,y)}{\partial x} - \frac{dV_{CB}(x)}{dx} \right] \quad (\text{F.6})$$

Using this and (F.2) in (F.1) we see that the $\partial\psi/\partial x$ terms cancel out, resulting in

$$dI_{DS} = (W dy) q \mu n(x,y) \frac{dV_{CB}(x)}{dx} \quad (\text{F.7})$$

The total drain current can be obtained by integrating over depth from $y = y_{\text{surface}}$ to a point $y = y_c$, below which the electron concentration is negligible. Since V_{CB} is assumed independent of y as mentioned earlier, this gives

$$I_{DS} = W \mu \frac{dV_{CB}(x)}{dx} q \int_{y_{\text{surface}}}^{y_c} n(x,y) dy \quad (\text{F.8})$$

where we have assumed that μ is independent of y . The integral times q can be recognized from (2.6.4) as $-Q'_I$ at position x ; thus

$$I_{DS} = \mu W (-Q'_l) \frac{dV_{CB}(x)}{dx} \quad (\text{F.9})$$

Integrating over the length of the channel and recognizing $\int_0^L I_{DS} dx$ as $I_{DS}L$, we have

$$I_{DS} = \frac{W}{L} \int_{V_{SB}}^{V_{FB}} \mu (-Q'_l) dV_{CB} \quad (\text{F.10})$$

Note that the preceding two relations are of the same form as (4.7.5) and (4.7.6), which were developed by starting from strong-inversion assumptions. However, it is clear from the preceding development that (F.9) and (F.10) are actually very general. They are valid in all regions of operation, since they include both drift and diffusion effects. The development in Sec. 4.7.1 following (4.7.6) cannot be used here, since the approach there is only valid in strong inversion. A more general expression for Q'_l must be used. Depending on the expression used for Q'_l in conjunction with (F.10), different formulations are possible. These are shown below.

All-Region (Charge Sheet) Model

In the all-region charge sheet model (Sec. 4.3), Q'_l is known as a function of the surface potential ψ_s [see (4.3.13)]. Thus a change of variables from V_{CB} to ψ_s is made in (F.10):

$$I_{DS} = \frac{W}{L} \int_{\psi_{s0}}^{\psi_s} \mu (-Q'_l) \frac{dV_{CB}}{d\psi_s} d\psi_s \quad (\text{F.11})$$

where V_{CB} is related to ψ_s by an equation analogous to (4.3.15):

$$\psi_s = V_{GB} - V_{FB} - \gamma \sqrt{\psi_s + \phi_t e^{(\psi_s - 2\phi_t - V_{CB})/\phi_t}} \quad (\text{F.12})$$

The quantity $dV_{CB}/d\psi_s$ can be found from this and substituted in (F.11) to determine the current. This results in a rather lengthy development. Several derivations of the all-region charge sheet model have been proposed.²⁻⁵ The technique we have presented in Sec. 4.3 is the most straightforward.

Pao-Sah Model

In this model,¹ no simplifying assumptions are made, in the sense that both electrons and holes are allowed to exist in the depletion region. This results, as expected from (E.16), in

$$Q'_l = -q N_A e^{(-2\phi_t - V_{CB})/\phi_t} \int_{\psi_t}^{\psi_s} \frac{e^{\psi/\phi_t}}{\mathcal{E}[\psi]} d\psi \quad (\text{F.13})$$

where, consistent with our present convention, we consider the electrons between the surface and a depth beyond which the electron concentration becomes negligible. The

potential at that point with respect to the substrate is denoted by ψ_c . A convenient depth is that at which $n = n_i$. From (F.4) and (1.2.13b), we easily see that the corresponding potential is $\psi_c = \phi_F + V_{CB}$. The quantity $\mathcal{E}[\psi]$ in (F.13) is given in (E.14). Using (F.13) in (F.10), we obtain

$$I_D = \frac{W}{L} q N_A \int_{V_{SB}}^{V_{DB}} \mu \int_{\psi_c}^{\psi} \frac{e^{(\psi - 2\phi_F - V_{CB})/\phi_i}}{\mathcal{E}[\psi]} d\psi dV_{CB} \quad (\text{F.14})$$

This double integral can be evaluated numerically. The required computation time is long, so this formulation is mainly of theoretical interest. The all-region charge sheet model gives nearly the same results and is much simpler. A technique to reduce (F.14) to a single-integral formula (which must still be evaluated numerically) has been proposed,⁶ and an extension to handle nonconstant mobility has been developed.⁷

References

1. H. C. Pao and C. T. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors," *Solid-State Electronics*, vol. 10, pp. 927–937, 1966.
2. G. Baccarani, M. Rudan, and G. Spadini, "Analytical i.g.f.e.t. model including drift and diffusion currents," *IEE Journal on Solid-State and Electron Devices*, vol. 2, pp. 62–68, 1978.
3. J. R. Brews, "A charge sheet model for the MOSFET," *Solid-State Electronics*, vol. 21, pp. 345–355, 1978.
4. F. Van de Wiele, "A long-channel MOSFET model," *Solid-State Electronics*, vol. 22, pp. 991–997, 1979.
5. J. R. Brews, "Physics of the MOS transistor," Chap. 1, *Silicon Integrated Circuits, Part A*, D. Kahng (editor), Applied Solid-State Science Series, Academic Press, New York, 1981.
6. R. F. Pierret and J. A. Shields, "Simplified long-channel MOSFET theory," *Solid-State Electronics*, vol. 26, pp. 143–147, 1983.
7. M. Persi and G. Gildenblat, "Computationally efficient version of the Pao-Sah model with variable mobility," *Solid-State Electronics*, vol. 38, pp. 1461–1463, 1995.

G

Modeling Based on Pinchoff Voltage and Related Topics

The pinchoff voltage V_p has been defined in Sec. 3.5.2 and is given by (3.5.6). This equation can be written, using (3.4.15) to eliminate V_{FB} , as follows

$$V_p = V_{GB} - V_{T0} - \gamma \left[\sqrt{V_{GB} - V_{T0} + \left(\sqrt{\phi_0} + \frac{\gamma}{2} \right)^2} - \left(\sqrt{\phi_0} + \frac{\gamma}{2} \right) \right] \quad (\text{G.1})$$

A plot of V_p vs. V_{GB} is shown in Fig. G.1. The fact that the plot crosses the horizontal axis at $V_{GB} = V_{T0}$ is consistent with the interpretation of V_p in (3.5.5).

As we have mentioned, in some treatment ϕ_0 is replaced by $2\phi_F$; also, somewhat different definitions for V_p are sometimes used.¹ The quantity V_p , as defined earlier, is related to ψ_{sa} in (3.2.24) by (3.5.7).

The inverse of the slope of ψ_{sa} vs. V_{GB} was denoted in (3.2.25) by n . Since $V_p(V_{GB})$ and $\psi_{sa}(V_{GB})$ differ by a constant, they have the same slope, so we can write

$$n = \left(\frac{dV_p}{dV_{GB}} \right)^{-1} \quad (\text{G.2})$$

In some treatments,¹ this is used as the definition of n . The value of n in terms of V_p can be found from (3.2.26) with (3.5.7):

$$n = 1 + \frac{\gamma}{2\sqrt{\phi_0 + V_p(V_{GB})}} \quad (\text{G.3})$$

Note that n has a well-defined value for a given V_{GB} , independent of V_{CB} .

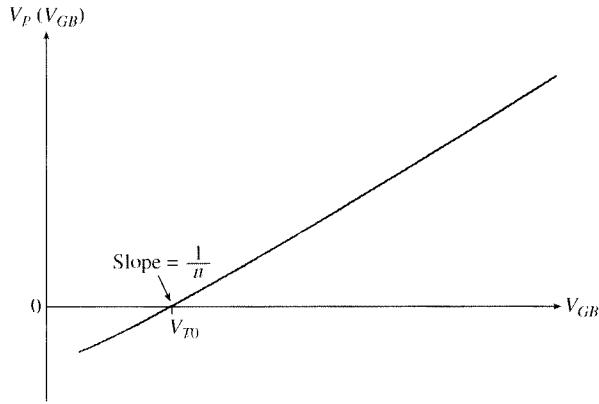


FIGURE G.1
The parameter V_p as a function of the gate-body voltage.

We now develop a simple approximation¹ for V_p . We see from (G.2) that the slope of $V_p(V_{GB})$ (Fig. G.1) is equal to $1/n$. Thus, if the plot is approximated by a straight line in Fig. G.1, we can write:¹

$$V_p \approx \frac{V_{GB} - V_{T0}}{n} \quad (\text{G.4})$$

where, strictly speaking, n should be evaluated at the point $V_{GB} = V_{T0}$. However, the approximation in (G.4) is used¹ with $n = n(V_{GB})$, as evaluated from (G.3) with (3.5.6). The slowly decreasing $n(V_{GB})$ (see Fig. 3.12) imparts a curvature to the plot of V_p from (G.4) and approximates the curve in Fig. G.1 rather closely.

We now consider expressions for Q'_I as a function of V_{CB} and V_p . These are used in conjunction with several models based on the “ V_{CB} control” point of view.¹⁻³

Charge Modeling

Strong Inversion The strong-inversion approximation for Q'_I was given by (3.5.3) and (3.5.4), and its magnitude was shown by the broken line in Fig. 3.16. From those equations and (G.3), we have:

$$\left. \frac{dQ'_I}{dV_{CB}} \right|_{V_{CB}=V_p} = C'_{ox} n \quad (\text{G.5})$$

Therefore, the strong-inversion approximation can be approximated by a first-order expansion around $V_{CB} = V_p$ as follows:¹

$$Q'_I \approx \left(\left. \frac{dQ'_I}{dV_{CB}} \right|_{V_{CB}=V_p} \right) (V_{CB} - V_p) \quad (\text{G.6})$$

$$Q'_I \approx -nC'_{\text{ox}}(V_P - V_{CB}) \quad (\text{G.7})$$

Weak Inversion Using ψ_{sa} from (3.5.7) in (3.4.20), we obtain

$$Q'_I = -\frac{\sqrt{2q\epsilon_s N_A}}{2\sqrt{\phi_0 + V_P}} \epsilon_t e^{(\phi_0 - 2\phi_F)/\phi_t} e^{(V_P - V_{CB})/\phi_t} \quad (\text{G.8})$$

where, again, V_P and n depend on V_{GB} , as already discussed.

Current Modeling

Expansion around ψ_{sa} Consider the following choice for ψ_{se} in (4.4.1):

$$\psi_{se} = \psi_{sa} = V_P + \phi_0 \quad (\text{G.9})$$

This leads to the approximation shown by the broken line in Fig. 4.16. We remind the reader that ψ_{sa} is the maximum possible surface potential for a value given V_{GB} (see Fig. 4.6). The choice in (G.9) gives an accurate Q'_B and thus, when $\psi_s \approx \psi_{sa}$, an accurate Q'_I . It will be recalled that at such values of ψ_s , Q'_I becomes negligible in comparison to Q'_B (Fig. 3.15), and we are in weak inversion or depletion. Thus, this approximation provides best accuracy for Q'_B when Q'_B is dominant. At the other extreme, i.e., when Q'_I is dominant, inaccuracies in Q'_B matter little anyway, since this charge contributes a negligible amount to the total semiconductor charge.⁴ Using (G.9) in (4.4.2), one finds that a α becomes equal to the quantity n in (G.3):

$$\alpha = n \quad (\text{G.10})$$

Using (G.9) and (G.10) in (4.4.5), we obtain

$$I_{DS1} = \frac{W}{L} \mu C'_{\text{ox}} \left[\left(V_{GB} - V_{FB} - \frac{\gamma}{2} \sqrt{\psi_{sa}} \right) (\psi_{sL} - \psi_{s0}) - \frac{n}{2} (\psi_{sL}^2 - \psi_{s0}^2) \right] \quad (\text{G.11a})$$

$$I_{DS2} = \frac{W}{L} \mu C'_{\text{ox}} n \phi_t (\psi_{sL} - \psi_{s0}) \quad (\text{G.11b})$$

These equations make apparent the symmetric role of the source and drain, just like the complete all-region model in Sec. 4.3.1. The values of ψ_{s0} and ψ_{sL} can be calculated, for given V_{GB} , V_{SB} , and V_{DB} , from (4.3.15).

Strong-Inversion Models Assuming all current in strong inversion is due to drift, we have, from (4.4.17a) and (G.10)

$$I_{DS} = \frac{W}{L} \frac{\mu}{2nC'_{\text{ox}}} (Q'^2_{l0} - Q'^2_{lL}) \quad (\text{G.12})$$

where Q'_{I0} and Q'_{IL} are the inversion charges per unit area at the source and drain ends of the channel, respectively. In nonsaturation, *both* ends are strongly inverted, so we will use for these charges the strong-inversion approximation (G.7). This gives the following expression for the *nonsaturation* current:^{1,5,6}

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \frac{n}{2} \left[(V_P - V_{SB})^2 - (V_P - V_{DB})^2 \right] \quad (\text{G.13})$$

To extend the preceding results into the saturation region, we follow an approach similar to that in Sec. 4.7. It can be easily verified that, as V_{DB} is raised, saturation is obtained at $V_{DB} = V_P$, at which point the slope dI_{DSN}/dV_{DB} becomes zero. Using $V_{DB} = V_P$ in (G.13), we obtain the value of the forward saturation current as follows:

$$I'_{DS} = \frac{W}{L} \mu C'_{ox} \frac{n}{2} (V_P - V_{SB})^2 \quad (\text{G.14})$$

Similarly, if V_{SB} is raised instead, we obtain reverse saturation when $V_{SB} = V_P$, and the reverse saturation current is

$$I''_{DS} = -\frac{W}{L} \mu C'_{ox} \frac{n}{2} (V_P - V_{DB})^2 \quad (\text{G.15})$$

These results can be combined into one strong-inversion model as in Fig. 4.21.

An even simpler model can be derived as follows. Using (G.4) in (G.13) gives a model derived by different means elsewhere:^{5-7,†}

$$I_{DSN} = \frac{W}{L} \mu C'_{ox} \left[(V_{GB} - V_{T0})(V_{DB} - V_{SB}) - \frac{n}{2} (V_{DB}^2 - V_{SB}^2) \right] \quad (\text{G.16})$$

By setting $dI_{DSN}/dV_{DB} = 0$, it can be easily checked that forward saturation is reached at $V_{DB} = V_P$, with V_P as given by (G.4). Using this value for V_{DB} in (G.16), we obtain the forward saturation current as follows:

$$I'_{DS} = \frac{W}{L} \mu C'_{ox} \frac{1}{2n} (V_{GB} - V_{T0} - nV_{SB})^2 \quad (\text{G.17})$$

Similarly, if V_{SB} is raised instead, reverse saturation is attained at $V_{SB} = V_P$, and the reverse saturation current is

$$I''_{DS} = -\frac{W}{L} \mu C'_{ox} \frac{1}{2n} (V_{GB} - V_{T0} - nV_{DB})^2 \quad (\text{G.18})$$

Again, the complete model is as in Fig. 4.21.

[†] This equation can be related to the source-referenced simplified strong-inversion model of Sec. 4.7.3 as follows. In (G.16), use $V_{GB} = V_{GS} + V_{SB}$ and $V_{DB} = V_{DS} + V_{SB}$. Then the equation can be manipulated to take the form of (4.7.24a), with $V_T = V_{T0} + (n - 1) V_{SB}$ [in lieu of (4.7.20)] and with n in lieu of α .

Weak-Inversion Models From (4.8.6) and (4.8.7), by using (3.4.3), (3.5.7), and (G.3), we obtain^{1,5,6,8}

$$I_{DS} = \frac{W}{L} \mu C'_{ox} (n-1) e^{(\phi_0 - 2\phi_t)/\phi_t} \phi_t^2 \left[e^{(V_p - V_{SB})/\phi_t} - e^{(V_p - V_{DB})/\phi_t} \right] \quad (\text{G.19})$$

An alternative expression is obtained by using in (G.19) the approximation for V_p in (G.4). This gives

$$I_{DS} = \frac{W}{L} \mu C'_{ox} (n-1) e^{(\phi_0 - 2\phi_t)/\phi_t} \phi_t^2 \left[e^{(V_{GB} - V_{T0} - nV_{SB})/\phi_t} - e^{(V_{GB} - V_{T0} - nV_{DB})/\phi_t} \right] \quad (\text{G.20})$$

with n as given by (G.3).

Single-Piece Models A single-piece model has been proposed as follows:^{1,6,8,9}

$$I_{DS} = \frac{W}{L} \mu C'_{ox} (2n) \phi_t^2 \left\{ \left[\ln \left(1 + e^{(V_p - V_{SB})/(2\phi_t)} \right) \right]^2 - \left[\ln \left(1 + e^{(V_p - V_{DB})/(2\phi_t)} \right) \right]^2 \right\} \quad (\text{G.21})$$

This “single-piece” model can be used in all regions, for both nonsaturation and saturation. It has a smooth behavior that approaches asymptotically the weak- and strong-inversion models in the respective regions. In weak inversion, both exponentials have a magnitude much smaller than 1; then by using the approximation $\ln(1+x) \approx x$ for $|x| \ll 1$, it is easily seen that the model reduces to the weak-inversion symmetric model of (G.19), only with $2n$ in place of $(n-1)e^{(\phi_0 - 2\phi_t)/\phi_t}$ (this problem will be discussed shortly). In strong inversion and deeply in nonsaturation, both exponentials are much larger than 1. Using now the approximation $[\ln(1+e^y)]^2 \approx (\ln e^y)^2 = y^2$ for $e^y \gg 1$, we see that these exponentials produce square-law terms and the model reduces to the strong-inversion expression (G.13). In moderate inversion, no simplification is possible, and the entire expression (G.21) must be used.

As V_{DB} is raised, the second exponential in (G.21) becomes negligible; saturation is then naturally obtained, and in strong inversion, the equation reduces to (G.14). If V_{SB} is raised instead, then it is the first exponential that becomes negligible and we enter reverse saturation.

If, in (G.21), we approximate V_p using (G.4), we obtain the following form:^{1,6,8}

$$I_{DS} = \frac{W}{L} \mu C'_{ox} (2n) \phi_t^2 \left\{ \left[\ln \left(1 + e^{(V_{GB} - V_{T0} - nV_{SB})/(2n\phi_t)} \right) \right]^2 - \left[\ln \left(1 + e^{(V_{GB} - V_{T0} - nV_{DB})/(2n\phi_t)} \right) \right]^2 \right\} \quad (\text{G.22})$$

Proceeding as already discussed, it is easy to verify that this equation reduces to (G.16) in strong-inversion nonsaturation, and to (G.20) in weak inversion, again with $(2n)$ in place of $(n-1)e^{(\phi_0 - 2\phi_t)/\phi_t}$.

Given the semiempirical nature of the preceding models, some error can be expected when correct physical values are used for their parameters. In such cases, a proper choice of parameter values is key to the success of the model. Consider, for example, (G.22), which in weak inversion reduces to (G.20), only with $2n$ in place of $(n-1)e^{(\phi_0-2\phi_F)/\phi_i}$. The latter disparity can cause an error in I_{DS} . If the value of V_{T0} in (G.20) is artificially increased, I_{DS} can be brought back to its correct value. Since V_{T0} appears in an exponential, it needs to be changed by only a few tens of millivolts for this to happen. One should be very cautious with such a practice, since such “fixes” could produce errors somewhere else. In particular, V_{T0} also affects the current in strong inversion; for example, in strong-inversion saturation, (G.22) reduces to (G.17). Fortunately, in strong inversion, the quantity in parentheses is rather large, so a small error in V_{T0} does not result in a large error in I_{DS} . In practice, during “parameter extraction,” several parameter values are adjusted for a best overall fit (Chap. 10). It has been shown that for the EKV model satisfactory results can be obtained in this way.¹

The preceding model has been converted to a source-referenced model in Ref. 10; the resulting expression has been given in Sec. 4.9.

References

1. C. C. Enz, F. Krummenacher, and E. A. Vittoz, “An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications,” *Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, 1995.
2. E. Vittoz and J. Fellrath, “CMOS analog circuits based on weak inversion operation,” *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 224–231, 1977.
3. A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, “An explicit physical model for the long-channel MOS transistor including small-signal parameters,” *Solid-State Electronics*, vol. 38, pp. 1945–1952, 1995.
4. A. I. A. Cunha, *Um Modelo do Transistor MOS para Projeto de Circuitos Integrados*, Ph.D. thesis, Universidade Federal de Santa Catarina, December 1996 (in Portuguese).
5. E. A. Vittoz, “Micropower Techniques,” pp. 104–143, in *Design of Analog-Digital VLSI Circuits for Telecommunications and Signal Processing*, J. E. Franca and Y. Tsividis (editors), Prentice Hall, Englewood Cliffs, N.J., 1994.
6. C. C. Enz, *High Precision CMOS Micropower Amplifiers*, Ph.D. thesis, no. 802. École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 1989.
7. K. A. Valiev, A. N. Karmazinskii, and A. M. Korolev, *Digital Integrated Circuits Using MOS Transistor*, Soviet Radio, Moscow, 1971.
8. H. Oguey and S. Cserveny, “Modèle du transistor MOS valable dans un grand domaine de courants,” *Sonderdruck aus dem Bulletin des SEV/VSE*, vol. 73, pp. 113–116, 1982.
9. G. T. Wright, “Simple and continuous MOSFET models for the computer-aided design of VLSI,” *IEE Proceedings*, vol. 132, part 1, pp. 187–194, 1985.
10. Y. Tsividis, K. Suyama, and K. Vavelidis, “A simple ‘Reconciliation’ MOSFET model valid in all regions,” *Electronics Letters*, vol. 31, pp. 506–508, 1995.

H

Evaluation of the Intrinsic Transient Source and Drain Currents

We present here a proof¹ of certain statements made in Sec. 6.3 for the intrinsic part of a MOS transistor. The continuity equation is, from (6.7.5),

$$\frac{\partial i(x,t)}{\partial x} = W \frac{\partial q'_I(x,t)}{\partial t} \quad (\text{H.1})$$

Integrating with respect to distance from the source to a point x in the channel, we obtain

$$i(x,t) - i(0,t) = W \int_0^x \frac{\partial q'_I(\hat{x},t)}{\partial t} d\hat{x} \quad (\text{H.2})$$

where \hat{x} is a dummy variable of integration. Recognizing $i(0,t)$ as $-i_S(t)$ (see the direction of i in Fig. 6.13), and using (6.7.6b) for $i(x,t)$ we obtain, from (H.2),

$$i_S(t) = \mu W q'_I(x,t) \frac{\partial v_{CB}(x,t)}{\partial x} + W \int_0^x \frac{\partial q'_I(\hat{x},t)}{\partial t} d\hat{x} \quad (\text{H.3})$$

Multiplying both sides by dx , integrating from $x = 0$ to $x = L$, and dividing both sides by L gives

$$i_S(t) = \frac{W}{L} \int_0^L \mu q'_I(x,t) \frac{\partial v_{CB}(x,t)}{\partial x} dx + \frac{W}{L} \int_0^L \int_0^x \frac{\partial q'_I(\hat{x},t)}{\partial t} d\hat{x} dx \quad (\text{H.4})$$

Interchanging the order of integration and differentiation in the second term of the right-hand side, we obtain

$$i_S(t) = \frac{W}{L} \int_0^L \mu q'_I(x, t) \frac{\partial v_{CB}(x, t)}{\partial x} dx + \frac{d}{dt} \left[\frac{W}{L} \int_0^L \int_0^x q'_I(\hat{x}, t) d\hat{x} dx \right] \quad (\text{H.5})$$

The double integral is of the form $\int_0^L G(x) dx$, where $G(x) = \int_0^x q'_I(\hat{x}, t) d\hat{x}$. Applying integration by parts to $\int_0^L G(x) dx$, with G and x the two variables involved, we can write (H.5) as follows:

$$i_S(t) = -\frac{W}{L} \int_0^L \mu [-q'_I(x, t)] \frac{\partial v_{CB}(x, t)}{\partial x} dx + \frac{d}{dt} \left[W \int_0^L \left(1 - \frac{x}{L}\right) q'_I(x, t) dx \right] \quad (\text{H.6})$$

Comparing the first term to the right-hand side of (4.7.6), we see that, in quasi-static operation, this term will produce the same current expressions as in Sec. 4.7.1, only with a minus sign and with the terminal voltages as functions of time. We can then write (H.6) as

$$i_S(t) = -i_T(t) + \frac{dq_S}{dt} \quad (\text{H.7})$$

where $i_T(t)$ will be of the form of (6.3.12), and

$$q_S = W \int_0^L \left(1 - \frac{x}{L}\right) q'_I dx \quad (\text{H.8})$$

To find $i_D(t) = i(L, t)$ in quasi-static operation, we use $x = L$ in (H.2) and substitute in it (H.7) and (H.8), which gives

$$i_D(t) = i_T(t) + \frac{dq_D}{dt} \quad (\text{H.9})$$

with

$$q_D = W \int_0^L \frac{x}{L} q'_I dx \quad (\text{H.10})$$

Since quasi-static operation is assumed, (H.8) and (H.10) can be evaluated by using the dc charge per unit area Q'_I instead of q'_I . The resulting quantities have been denoted by Q_S and Q_D , respectively, in (6.3.9b) and (6.3.9a). The detailed evaluation of Q_S and Q_D is considered in Sec. 6.4.

A Note on q_D and q_S

If desired, $q_D(t)$ can be defined as $\int'_{-\infty} [i_D(\tau) - i_T(\tau)] d\tau = \int'_{-\infty} i_{DA}(\tau) d\tau$, where τ is a dummy variable of integration, and $q_S(t)$ as $\int'_{-\infty} [i_S(\tau) + i_T(\tau)] d\tau = \int'_{-\infty} i_{SA}(\tau) d\tau$ (i_{DA} and i_{SA} are the “charging” currents discussed in Sec. 6.3). These definitions imply (6.3.6), and, by integrating (6.3.5), we see that they satisfy (6.3.8) (assuming that the channel was “empty” at $t = -\infty$). With these definitions, $q_D(t)$ can be interpreted as the part contributed to $q_I(t)$ by the deviation of i_D from i_T from the “beginning of time” to the instant t . Similarly, $q_S(t)$ can be interpreted as the part contributed to

$q_I(t)$ by the corresponding deviation of i_S from $-i_T(t)$. Although the preceding definitions may be elegant, they can confuse the issue: q_D and q_S cannot be found stored in any particular place, and neither can they be associated with specific charges that have been flowing by themselves through any terminal. This is because $i_{DA}(t)$ and $i_{SA}(t)$ are mere artifacts, each being the difference between *real* terminal currents and the value such terminal currents *would* have if one attempted to predict them by dc theory [see (6.3.4)]. Also, the previous definitions for q_D and q_S tend to assign specific significance to the *values* of these fictitious charges, when all that matters is their *derivatives* [see (6.3.6)]. Assume q_D and q_S as previously defined are given by two functions f_D and f_S , respectively [where f_D and f_S are the same functions as in (6.3.10)], giving the correct values for i_{DA} and i_{SA} from (6.3.6). The quantities q_D and q_S satisfy (6.3.8), as already noted. Now let us redefine q_D and q_S as $f_D + K$ and $f_S + L$, respectively, where K and L are arbitrary constants. Then (6.3.6) will still predict the correct values of i_{DA} and i_{SA} . Equation (6.3.8) will, in general, not hold now, but this is of *no* consequence and does *not* mean that charge conservation is violated. This is because, since i_{DA} and i_{SA} are still predicted correctly, they still satisfy (6.3.5) and (6.3.3). Integrating (6.3.3), one sees that the total charge that entered the device through source and drain up to time t is equal to $q_I(t)$, and thus the charge *is* conserved.

Reference

1. D. E. Ward, "Charge-based modeling of capacitance in MOS transistors," *Technical Report G201-11*, Integrated Circuits Laboratory, Stanford University, Stanford, Calif., June 1981.

Quantities Used in the Derivation of the Non-Quasi-Static y -Parameter Model

Expressions for y parameters, including non-quasi-static effects, have been given in (8.4.58) to (8.4.60). In these expressions, the numerators $N_{kl}(\omega)$ (where $k, l = d, g, b$) are of the form

$$N_{kl}(\omega) = n_{kl0} + (j\omega)n_{kl1} + (j\omega)^2n_{kl2} + \dots$$

and the denominator $D(\omega)$ is of the form

$$D(\omega) = d_0 + (j\omega)d_1 + (j\omega)^2d_2 + \dots$$

The coefficients in these relations can be found as discussed in Sec. 8.4.2; the results are given in the following. Some of these results apply to the three-terminal transistor;¹ all the results apply to the four-terminal transistor.²

$$n_{dd0} = \frac{W}{L}\mu C'_{ox} (V_{GS} - V_T)\eta = g_{sd}$$

$$n_{dg0} = \frac{W}{L}\mu C'_{ox} \frac{V_{GS} - V_T}{\alpha_1}(1 - \eta) = g_m$$

$$n_{db0} = (\alpha_1 - 1) n_{dg0} = g_{mb}$$

$$n_{gd0} = n_{gg0} = n_{bd0} = n_{gb0} = n_{bg0} = n_{bb0} = 0$$

$$n_{dd1} = \frac{2}{3}\alpha_1 C_{ox} \frac{\eta(2 + \eta)}{(1 + \eta)^2}$$

$$n_{dg1} = n_{gd1} = -\frac{2}{3}C_{ox} \frac{\eta(2 + \eta)}{(1 + \eta)^2}$$

$$n_{db1} = (\alpha_1 - 1) n_{dg1}$$

$$n_{gg1} = C_{ox} \left[\frac{2}{3\alpha_1} \frac{1 + 4\eta + \eta^2}{(1 + \eta)^2} + \frac{\alpha_1 - 1}{\alpha_1} \right]$$

$$n_{gb1} = n_{bg1} = -\frac{\alpha_1 - 1}{3\alpha_1} C_{ox} \left(\frac{1 - \eta}{1 + \eta} \right)^2$$

$$n_{bd1} = (\alpha_1 - 1) n_{gd1}$$

$$n_{bb1} = C_{ox} \left[\frac{\alpha_1 - 1}{\alpha_1} + \frac{2(\alpha_1 - 1)^2}{3\alpha_1} \frac{1 + 4\eta + \eta^2}{(1 + \eta)^2} \right]$$

$$n_{dd2} = \frac{C_{ox}}{\omega_o} \frac{2}{45} \alpha_1 \frac{\eta(5 + 8\eta + 2\eta^2)}{(1 + \eta)^4}$$

$$n_{dg2} = n_{gd2} = -\frac{C_{ox}}{\omega_o} \frac{2}{45} \frac{\eta(5 + 8\eta + 2\eta^2)}{(1 + \eta)^4}$$

$$n_{db2} = (\alpha_1 - 1) n_{dg2}$$

$$n_{gg2} = \frac{C_{ox}}{\omega_o} \left[\frac{2}{45} \frac{1}{\alpha_1} \frac{2 + 11\eta + 2\eta^2}{(1 + \eta)^3} + \frac{4}{15} \frac{\alpha_1 - 1}{\alpha_1} \frac{1 + 3\eta + \eta^2}{(1 + \eta)^3} \right]$$

$$n_{gb2} = n_{bg2} = -\frac{C_{ox}}{\omega_o} \frac{2}{45} \frac{\alpha_1 - 1}{\alpha_1} \frac{4 + 7\eta + 4\eta^2}{(1 + \eta)^3}$$

$$n_{bd2} = (\alpha_1 - 1) n_{gd2}$$

$$n_{bb2} = \frac{C_{ox}}{\omega_o} \left[\frac{4}{15} \frac{\alpha_1 - 1}{\alpha_1} \frac{1 + 3\eta + \eta^2}{(1 + \eta)^3} + \frac{2}{45} \frac{(\alpha_1 - 1)^2}{\alpha_1} \frac{2 + 11\eta + 2\eta^2}{(1 + \eta)^3} \right]$$

$$d_0 = 1$$

$$d_1 = \frac{4}{15} \frac{1}{\omega_o} \frac{1 + 3\eta + \eta^2}{(1 + \eta)^3}$$

$$d_2 = \frac{1}{45} \frac{1}{\omega_o^2} \frac{1 + 4\eta + \eta^2}{(1 + \eta)^4}$$

where

$$C_{ox} = C'_{ox} WL$$

$$\omega_o = \frac{\mu(V_{GS} - V_T)}{\alpha L^2}$$

$$\eta = \begin{cases} 1 - \frac{V_{DS}}{V'_{DS}}, & V_{DS} \leq V'_{DS} \\ 0, & V_{DS} > V'_{DS} \end{cases}$$

and, according to the simplifying assumptions stated in the beginning of Sec. 8.4.2,

$$\alpha_1 = 1 + \frac{\gamma}{2\sqrt{\phi_0 + V_{SB}}}$$

As explained in Sec. 8.4, the accuracy of the expressions containing α_1 can be improved if α_1 is replaced by an empirical parameter which, in general, would not be the same for all expressions; a related discussion can be found in Sec. 7.3.

References

1. J. J. Paulos and D. Antoniadis, "Limitations of quasi-static models for the MOS transistor," *IEEE Electron Device Letters*, vol. ED-4, pp. 221–224, 1983.
2. M. Bagheri and Y. Tsividis, "A small-signal non-quasi-static model for the four-terminal MOSFET valid in all regions of operation," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2383–2391, 1985.



Analysis of Buried-Channel Devices

We provide here a quantitative analysis to supplement the discussion of buried-channel devices in Sec. 9.3.4. Figure 9.12 is reproduced here as Fig. J.1, annotated with several symbols that will be used in the analysis that follows.

J.1 Charges and Threshold Voltage

Let Q'_J denote the magnitude of the charge in either side of the depletion region between the n implant and the p substrate in Fig. J.1. To find this charge, we will assume a two-sided step junction (as opposed to the one-sided junction emphasized in Sec. 1.5), because N_{DS} , as it results from (9.3.27), is in general not much larger than N_{AB} . Using the results in the footnote that follows (1.5.11), we have

$$Q'_J = \sqrt{2q\epsilon} \sqrt{\frac{N_{DS} N_{AB}}{N_{DS} + N_{AB}}} \sqrt{\phi_{bi} + V_{CB}} \quad (J.1)$$

where ϕ_{bi} is the built-in potential of the junction. Note that the charge per unit area is $+Q'_J$ in the top side of the junction and $-Q'_J$ in the bottom.

The gate charge is related to the oxide potential as in (2.4.22):

$$Q'_G = C'_{ox} \psi_{ox} \quad (J.2)$$

To write the charge balance equation, we note that the charges in the substrate-implant depletion region balance each other out and therefore do not contribute a net charge to the overall charge balance equation. Thus, the latter equation will involve

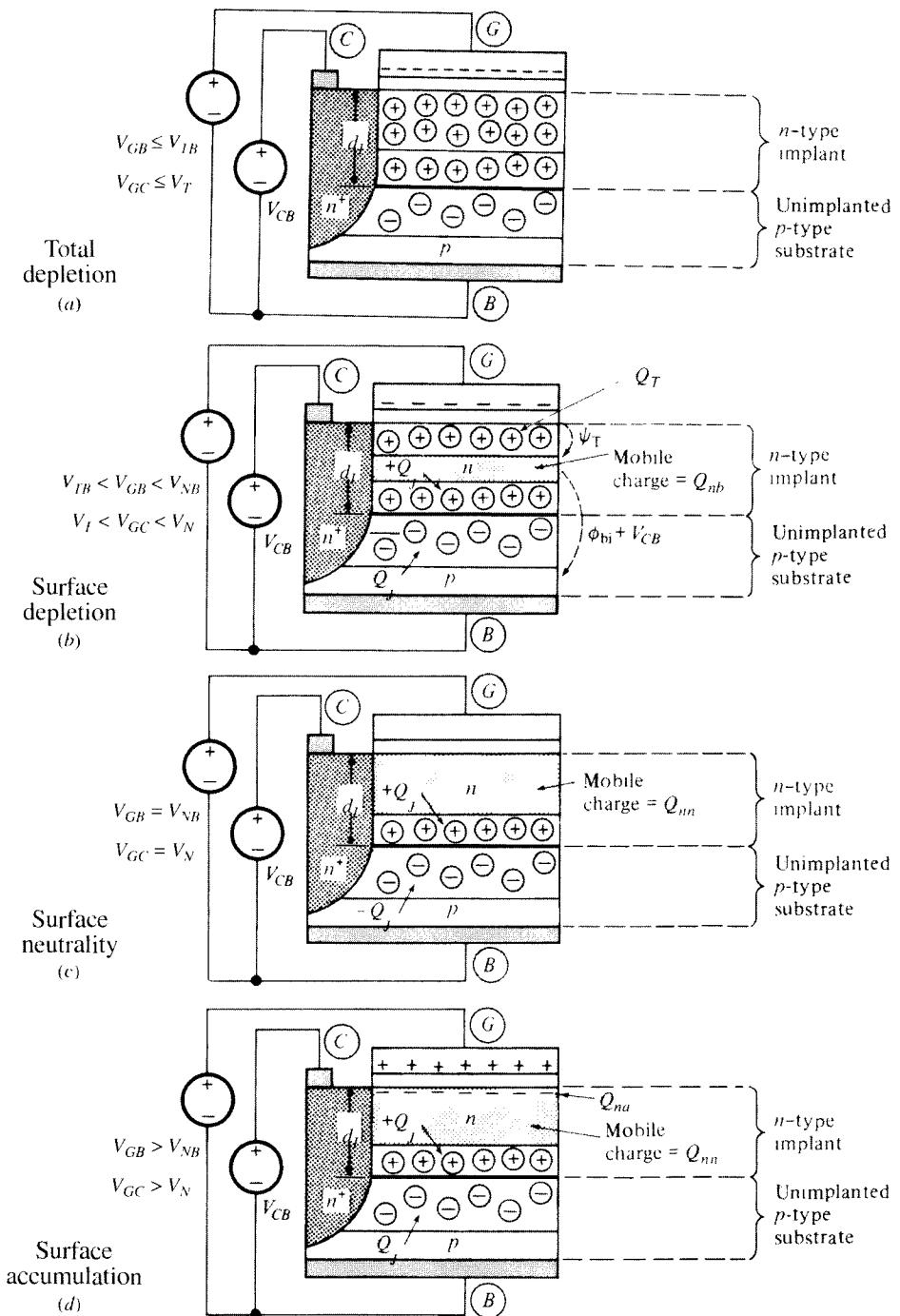


FIGURE J.1

A buried-channel device for various gate-substrate voltage values. The depletion region around the n^+ region boundary is not shown, and Q'_0 is assumed 0 for simplicity. (a) Total depletion; (b) surface depletion; (c) surface neutrality; (d) surface accumulation.

only the gate charge Q'_G , the interface charge Q'_o , and any charge that might be present at the top of the implant region. In order to write the charge balance and potential balance equations, it is convenient to distinguish three cases: surface depletion, surface neutrality, and surface accumulation.¹⁻³

Surface Depletion In Fig. J.1b, the charge at the top part of the implant, which will be denoted by Q'_T , is due to ionized donor atoms and is positive. This is opposite from the depletion condition we have encountered in *p*-type substrates, and the potential ψ_T across the top depletion region (from the surface to the neutral part of the *n* implant) indicated in the figure has a *negative* sign. We can use a relation like (3.2.17a), but with appropriate sign changes:

$$Q'_T = Q'_{T1} = +\sqrt{2q\epsilon_s N_{DS}} \sqrt{-\psi_T} \quad (\text{J.3})$$

With Q'_T as in (J.3), the charge balance equation is

$$Q'_G + Q'_o + Q'_T = 0 \quad (\text{J.4})$$

The undepleted implant region in Fig. J.1b does not contribute to the charge balance equation, since it is neutral: ionized impurity atoms in it are “covered” by mobile carriers of opposite charge.

The potential balance equation is obtained by going from the gate terminal through the device to the substrate terminal:

$$V_{GB} = \psi_{ox} + \psi_T + (\phi_{bi} + V_{CB}) + \phi_{MS} \quad (\text{J.5})$$

where $\phi_{bi} + V_{CB}$ is the total potential drop across the depletion regions of the implant-substrate junction in Fig. J.1b [see (1.5.3)], and ϕ_{MS} is the work function potential difference corresponding to the gate and the *unimplanted p*-substrate materials, as always.

Using (J.2) to (J.5) and solving for Q'_T , we obtain

$$Q'_T = -\frac{q\epsilon_s N_{DS}}{C'_{ox}} + \sqrt{\left(\frac{q\epsilon_s N_{DS}}{C'_{ox}}\right)^2 - 2q\epsilon_s N_{DS}(V_{GB} - V_{FB} - \phi_{bi} - V_{CB})} \quad (\text{J.6})$$

where V_{FB} is given by (9.3.4). We now calculate the value of the gate-body threshold voltage V_{TB} . This voltage was defined as the value of V_{GB} that causes the two depletion regions in the *n* implant to meet in Fig. J.1b, resulting in the situation of Fig. J.1a. In this case, the total charge per unit area in these two depletion regions is equal to the total donor charge per unit area. The latter is $qN_{DS}d_I$, with N_{DS} given by (9.3.27). Thus, we have

$$Q'_T|_{V_{GB}=V_{TB}} + Q'_J = qN_{DS}d_I \quad (\text{J.7})$$

Using in this expression (J.6) and (J.1), we can solve for V_{TB} . The expression obtained is of the form

$$V_{TB} = V_{CB} + V_T(V_{CB}) \quad (\text{J.8})$$

A detailed expression for $V_T(V_{CB})$ can be derived from the preceding analysis. That expression, if we assume $N_I \gg N_{AB}$, simplifies to the following:³

$$V_T(V_{CB}) = V_{T0} + \gamma_I \left(\sqrt{\phi_{bi} + V_{CB}} - \sqrt{\phi_{bi}} \right), \quad N_I \gg N_{AB} \quad (\text{J.9})$$

where

$$V_{T0} = V_{FB} + \phi_{bi} - \frac{qN_{DS}d_I}{C'_{ox}} \left(1 + \frac{d_I C'_{ox}}{2\epsilon_s} \right) + \gamma_I \sqrt{\phi_{bi}} \quad (\text{J.10})$$

and

$$\gamma_I = \left(1 + \frac{d_I C'_{ox}}{\epsilon_s} \right) \gamma \quad (\text{J.11})$$

with γ being the body effect coefficient of an *unimplanted* device on the same *p*-type substrate. It is thus seen from (J.9) that the threshold voltage V_T increases with V_{CB} in a way similar to the threshold voltage of unimplanted devices. However, the effective body effect coefficient γ_I is larger than γ .

In a transistor, the gate-source threshold voltage can be found by replacing the role of terminal C by the source terminal S. Thus, using V_{SB} in lieu of V_{CB} in (J.9), we obtain

$$V_T = V_{T0} + \gamma_I \left(\sqrt{\phi_{bi} + V_{SB}} - \sqrt{\phi_{bi}} \right), \quad N_I \gg N_{AB} \quad (\text{J.12})$$

with V_{T0} and γ_I as given previously.

In some contexts, modeling stops at this point. The resulting expression for V_T is used in the transistor equations for the unimplanted device (for example, those in Sec. 4.7.3), without much justification. For the interested reader, we will show how a more careful model can be developed. This will also be useful in order to show why, in some cases, models for unimplanted devices seem to do an adequate job for the implanted transistors we are considering.

The mobile charge per unit area in the surface depletion condition (Fig. J.1b) is due to the free electrons in the *undepleted* part of the implant and will be denoted by Q'_{nb} . If the depletion regions were absent, the number of mobile electrons would be equal to the effective number of donor atoms. Thus Q'_{nb} would be equal to $-qN_{DS}d_I$. However, now some donor atoms have already been depleted. Since the electron population is reduced by one electron for each depleted atom, the remaining mobile electron charge per unit area is

$$Q'_{nb} = -(qN_{DS}d_I - Q'_T - Q'_I) \quad (\text{J.13})$$

Surface Neutrality In Fig. J.1c the gate-substrate voltage is assumed to be such that the surface is neutral. Thus, the gate charge just balances the interface charge:

$$Q'_G + Q'_o = 0 \quad (\text{J.14})$$

Since the surface is neutral, there is no voltage drop associated with it. Thus we can write

$$V_{GB} = \psi_{\text{ox}} + \phi_{\text{bi}} + V_{CB} + \phi_{MS} \quad (\text{J.15})$$

Using (J.2) and (J.14) in this equation, we find that V_{GB} (the value of which is denoted by V_{NB} for this case of surface neutrality) is given by

$$V_{NB} = \phi_{MS} - \frac{Q'_o}{C'_{\text{ox}}} + \phi_{\text{bi}} + V_{CB} \quad (\text{J.16})$$

which becomes, using (9.3.4):

$$V_{NB} = V_N + V_{CB} \quad (\text{J.17})$$

where

$$V_N = V_{FB} + \phi_{\text{bi}} \quad (\text{J.18})$$

The mobile charge in the surface neutrality condition can be found as was done in order to obtain (J.13), only that now Q'_T has been reduced to zero. Denoting the mobile charge per unit area in this case by Q'_{mn} , we have:

$$Q'_{mn} = -(qN_{DS}d_I - Q'_j) \quad (\text{J.19})$$

Surface Accumulation[†] With V_{GB} raised above V_{NB} (Fig. J.1d), the charge at the top of the implant is due to mobile electrons in the accumulation layer. The thickness of this layer is very small and will be assumed to be infinitesimal; hence the potential needed across it to support it will be negligible (Appendix A). Thus (J.15) remains valid. The charge balance equation, though, must take into account the charge per unit area of the electrons in the accumulation layer; this charge will be denoted by Q'_{na} . Thus, we have

$$Q'_G + Q'_o + Q'_{na} = 0 \quad (\text{J.20})$$

Using in this equation (J.2) and (J.15) and solving for Q'_{na} , we obtain

[†] Also referred to as *surface enhancement*.

$$Q'_{na} = -C'_{ox}(V_{GB} - V_{FB} - \phi_{bi} - V_{CB}) \quad (J.21)$$

This is only the mobile charge in the accumulation layer; in addition, there is the mobile charge in the rest of the undepleted implant, which has already been calculated for the surface neutrality condition. That charge was found to be

$$Q'_{nn} = -(qN_{DS}d_f - Q'_j) \quad (J.22)$$

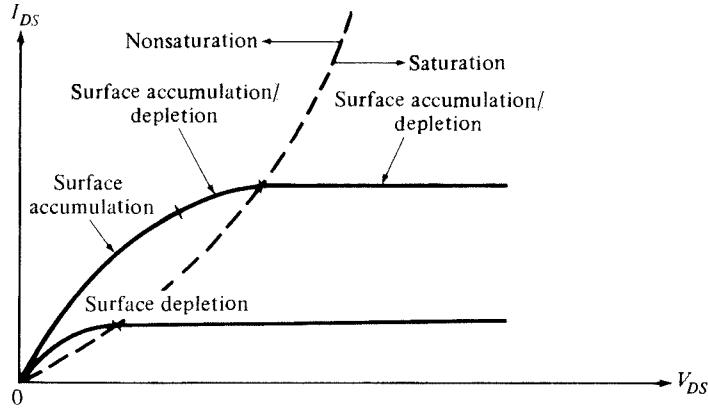
Conditions in Terms of V_{GB} and V_{GC} The relations that must be satisfied by V_{GB} and V_{GC} , for the various conditions discussed to be attained, are shown next to each part of Fig. J.1d. The relations in terms of V_{GB} follow directly from our discussion. The relations in terms of V_{GC} follow easily from those, by observing that $V_{GC} = V_{GB} - V_{CB}$, and using (J.8) and (J.17).

J.2 Transistor Operation

We now consider a complete transistor as in Fig. 9.3a, with $V_{GS} > V_T(V_{SB})$, so that a channel can exist next to the source (the various results obtained previously can be applied to the source end of the channel, by replacing V_{CB} by V_{SB}). We assume $V_{DB} > V_{SB}$ ($V_{DS} > 0$). We will denote the effective reverse bias between a point in the neutral part of the channel and the unimplanted substrate by V_{CB} . We will thus be able to apply directly the formulas derived previously. The value of V_{CB} increases from V_{SB} at the source to V_{DR} at the drain. Thus, the depletion region of the *np* junction will be deeper as we go toward the drain. Depending on the relative values of the terminal voltages, we can distinguish several regions of operation, which are discussed in the text that follows.

Assume first that surface depletion exists next to the source. From Fig. J.1b, this requires $V_T(V_{SB}) < V_{GS} < V_N$. Since V_{CB} increases toward the drain, $V_T(V_{CB})$ also increases, from (J.9), and tends to drive the drain end of the channel toward pinchoff. The two regions of operation in Fig. 9.13a and b are called *nonsaturation* and *saturation*, respectively; the critical value of V_{DS} at the transition between them is denoted by V'_{DS1} . In saturation, the electrons are assumed to travel from the tip of the pinched-off channel through the depletion region to the drain, through a mechanism analogous to that for unimplanted devices in saturation. In analogy to the corresponding simplified picture for those devices, it can be argued here that, if the channel in Fig. 9.13b is long, the drain current in saturation will be practically constant and equal to the nonsaturation current obtained as V_{DS} approaches V'_{DS1} . A typical I_{DS} - V_{DS} characteristic for this case is shown by the lower curve in Fig. J.2. A quantitative discussion of I_{DS} will follow later.

Assume now that V_{GB} is large enough to cause surface accumulation at the source. From Fig. J.1d, this requires $V_{GS} > V_N$. If V_{DS} is small, surface accumulation can exist throughout the length of the channel, as shown in Fig. 9.13c. As the drain potential is raised, though, the depletion region widens and $|Q'|$ increases near the drain. From (J.21), with $V_{CB} = V_{DB}$, $|Q'_{na}|$ there will be decreasing and will eventually disappear. For V_{DB} values above this the drain end of the channel will exhibit surface depletion, as shown in Fig. 9.13d. For this to happen, we must have, from Fig. J.1b, $V_{GD} < V_N$. Writing $V_{GD} = V_{GS} - V_{DS}$, this gives the condition $V_{DS} > V_{GS} - V_N$. Finally, further increase in the drain potential can increase $V_T(V_{DB})$ to the point where the channel becomes pinched off, as shown in Fig. 9.13e. The corresponding critical

**FIGURE J.2**

Relation of I - V characteristics to the modes of operation shown in Fig. 9.13.

value of V_{DS} is denoted by V'_{DS2} . An I_{DS} - V_{DS} characteristic corresponding to the cases in Fig. 9.13c, d, and e is shown by the upper curve in Fig. J.2.

We now show how the drain current can be calculated for the various regions summarized in Fig. 9.13.^{2,3} We first consider nonsaturation operation, corresponding to cases a, c, and d in Fig. 9.13. Here we cannot use (4.7.6), since there is no inversion layer involved. However, the role of Q'_f in that equation is now played by the total mobile charge per unit area, which, as we have already seen, can consist of either electrons flowing in the bulk of the implant, or of accumulation layer electrons flowing at its surface, or both. The mobilities involved will not be the same, since different scattering mechanisms are involved, as discussed in Sec. 4.11. Surface mobility is smaller than bulk mobility.

We now consider each of the three nonsaturation cases (see Figs. 9.13 and J.2) separately.

- 1. Surface depletion.**[†] In this case, shown in Fig. 9.13a, the total mobile charge per unit area is Q'_{nb} , given by (J.13). Thus instead of (4.7.6) we have

$$I_{DSN} = \frac{W}{L} \int_{V_{SB}}^{V_{DB}} \mu_B(-Q'_{nb}) dV_{CB} \quad (\text{J.23})$$

where μ_B is the value of the mobility *in the bulk* of the n region.

- 2. Surface accumulation,** Fig. 9.13c. Here the total mobile charge consists of the surface charge Q'_{na} given by (J.21), and the charge Q'_{nn} due to the mobile electrons in the undepleted part of the n-region bulk, given by (J.22). Hence, instead of (J.23) we have

$$I_{DSN} = \frac{W}{L} \int_{V_{SB}}^{V_{DB}} [\mu_S(-Q'_{na}) + \mu_B(-Q'_{nn})] dV_{CB} \quad (\text{J.24})$$

where μ_S is the *surface* mobility.

[†] This mode of operation is sometimes also referred to as *the buried-channel mode*.

3. Surface accumulation/depletion. This case is shown in Fig. 9.13d. Let V_{CBI} be the value of V_{CB} corresponding to that point in the channel where we pass from accumulation to depletion behavior. At this point, $Q'_{na} = 0$ and from (J.21), we obtain

$$V_{CBI} = V_{GB} - V_{FB} - \phi_{bi} \quad (J.25)$$

To the left of this point, we have a total mobile charge per unit area of $Q'_{na} + Q'_{nn}$, from (J.21) and (J.22); to the right, the mobile charge per unit area is Q'_{nb} , from (J.13). Thus we have

$$I_{DSN} = \frac{W}{L} \int_{V_{GS}}^{V_{CBI}} [\mu_s(-Q'_{na}) + \mu_B(-Q'_{nn})] dV_{CB} + \frac{W}{L} \int_{V_{CBI}}^{V_{DB}} \mu_B(-Q'_{nb}) dV_{CB} \quad (J.26)$$

Using the expressions developed for Q'_{nb} , Q'_{na} , and Q'_{nn} in the preceding equations provides the nonsaturation current I_{DSN} . The value of V'_{DS} can be obtained as usual by setting $dI_{DSN}/dV_{DS} = 0$ or by setting the total mobile charge at the drain end of the channel equal to zero. The resulting formulation is complicated but, as before, the expressions can be simplified through appropriate series expansions. Simplified formulas thus obtained are shown in Refs. 3 and 4.

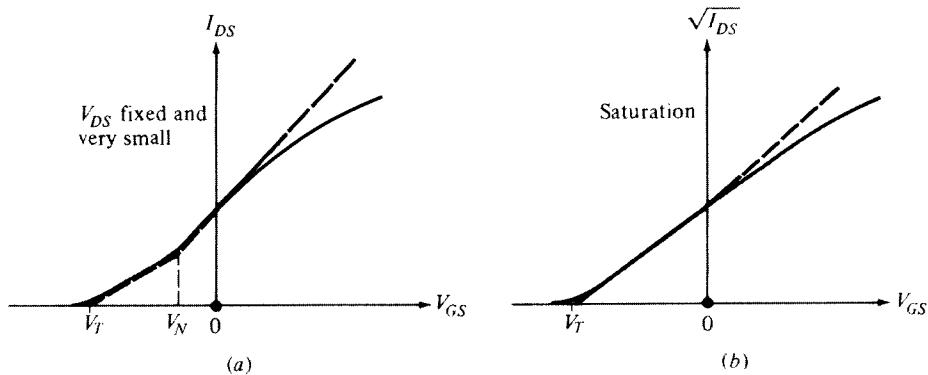
As before, an effective surface mobility can be defined and made dependent on V_{GS} , as in Sec. 4.11.^{3,5,6,†} The bulk mobility μ_B often is taken to be independent of V_{GS} .[‡]

For very small, fixed V_{DS} , the plot of I_{DS} vs. V_{GS} , assuming constant mobility, is as shown by the broken line in Fig. J.3. For buried channels ($V_T \leq V_{GS} < V_N$), the device behaves as an unimplanted device qualitatively. However, although μ_B is larger than the surface mobility, as shown in Fig. 9.13a the channel is further away from the gate than in the case of an unimplanted device, and thus the influence of the gate is somewhat reduced. However, once surface accumulation occurs ($V_{GS} \geq V_N$), variations of the channel charge with V_{GS} occur at the surface, and, thus, the slope is determined by the surface mobility and the oxide thickness only. The solid line in Fig. J.3a represents a real device, and no breakpoint is observed in this case. Mobility degradation at high V_{GS} can be seen. The deviation from straight line at very low V_{GS} is due to the presence of diffusion currents, which we have not included in the simple model we have presented. This phenomenon is reminiscent of weak inversion in unimplanted devices.⁸

For the plot of $\sqrt{I_{DS}}$ vs. V_{GS} in saturation, the model predicts again a changing slope effect, although not so pronounced, as shown by the broken line in Fig. J.3b. The slight upward trend of the slope tends to be counteracted by the degradation of mobility with V_{GS} , and the resulting behavior is as shown by the solid line,

[†] A simple approximate model suggested for this³ is $\mu_{s,eff} = \mu_{s,0} / [1 + \theta(V_{GS} - V_N)]$.

[‡] It has, however, been suggested⁷ that, because the electrons move inside a thin slab of undepleted material (Fig. 9.13a), their mobility is limited to values below those for free bulk conduction. This is attributed to effects similar to those responsible for surface scattering, which become more severe as the "slab" becomes thinner. Since the thickness of the slab depends on V_{GS} , a dependence of μ_B on V_{GS} is suggested.

**FIGURE J.3**

Characteristics of an nMOS transistor with an n implant: (a) I_{DS} as a function of V_{GS} with very small V_{DS} ; (b) saturation $\sqrt{I_{DS}}$ vs. V_{GS} . Dashed line: model with constant mobility; solid line: real device.

a significant part of which is practically straight. Thus, a saturation equation like (4.7.24b) can again be used in such cases in the form

$$I_{DS} = k(V_{GS} - V_T)^2 \quad (\text{J.27})$$

only with the constant of proportionality k somewhat reduced from that of an unimplanted device, and V_T negative for low V_{SB} . For some depletion devices, the preceding equation has been shown to be accurate, even more so than for enhancement devices on the same substrate.⁹ In fact, in some circuit analysis computer programs, the complete equation (4.7.24), in both nonsaturation and saturation, is used to model implanted depletion devices, using a negative V_T . Depending on device details, however, significant errors can result from such approximations. Models involving 3/2 powers, analogous to (4.7.2), have also been proposed.^{1,10}

The model we have presented is not valid when the two depletion regions are about to meet. This is both because then our assumptions of "abrupt" depletion region edges are inadequate and because diffusion currents become important. Depending on the value of V_{SB} , the "meeting point" of the two regions can occur at various depths, a fact that affects both the distance from the gate and the effective implant concentration at that point (where most of the current flows). Thus, the dependence of I_{DS} on V_{SB} in this part of the operating range is complicated.

References

1. J. S. T. Huang and G. W. Taylor, "Modeling of an ion-implanted silicon-gate depletion mode IGFET," *IEEE Transactions on Electron Devices*, vol. ED-22, pp. 995–1001, 1975. See also W. Marciniaik and H. Madura, "Comments on the Huang and Taylor model of ion-implanted silicon-gate depletion-mode IGFET," *Solid-State Electronics*, vol. 28, pp. 313–315, 1985.
2. E. Demoulin and F. Van de Wiele, "Ion implanted MOS transistors," pp. 617–676, in *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.

3. G. Merckel, "Ion implanted MOS transistors—Depletion mode devices," pp. 677–688, in *Process and Device Modelling for Integrated Circuit Design*, F. Van de Wiele, W. L. Engl, and P. G. Jespers (editors), Noordhoff, Leyden, The Netherlands, 1977.
4. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed., Oxford University Press, Oxford, 1999.
5. G. Baccarani, F. Landini, and B. Ricco, "Depletion-mode MOSFET model including a field-dependent surface mobility," *IEE Proceedings*, vol. 127, part I, pp. 62–66, 1980.
6. S. Haque-Ahmed and C. A. T. Salama, "Depletion mode MOSFET modelling for CAD," *IEE Proceedings*, vol. 130, part I, pp. 281–286, 1983.
7. C.-Y. Wu and K. C. Hsu, "Mobility models for the I - V characteristics of buried-channel MOSFETs," *Solid-State Electronics*, vol. 28, pp. 917–923, 1985.
8. C. Turchetti and G. Masetti, "Analysis of the depletion-mode MOSFET including diffusion and drift currents," *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 773–782, 1985.
9. M. R. Wordeman and R. H. Dennard, "Threshold voltage characteristics of depletion-mode MOSFETs," *IEEE Transactions on Electron Devices*, vol. ED-28, pp. 1025–1030, 1981.
10. Y. A. El-Mansy, "Analysis and characterization of the depletion-mode IGFET," *IEEE Journal of Solid-State Circuits*, vol. SC-15, pp. 331–340, 1980.

K

MOSFET Model Benchmark Tests

The physics that underlies MOS transistor behavior, including both fundamental operation as well as many nonideal effects, has been presented in this book. Nevertheless, as stressed in Chap. 10, it can be very difficult to merge all of this analysis into a MOSFET model in such a way that it completely avoids unphysical and undesirable behavior, which can cause significant problems for circuit design. In particular, although simulation of digital CMOS circuits can be forgiving of even major errors in certain aspects of MOSFET models (such as output conductance), precision analog circuit design can be compromised by seemingly small or innocuous inaccuracies.^{1–3} Modern MOSFET models have significantly improved over the last couple of decades, and many of the worst problems that were systemic in widely used MOS models are no longer encountered.

However, given the complexity of MOSFET models, it is not difficult for unphysical or undesired behavior to inadvertently creep into a model. A quality assurance procedure to prevent this from happening is to define a set of benchmark tests that are designed to expose and highlight known problems, and then apply these tests to a model (over bias, geometry, temperature, and allowable parameter ranges) to verify that it does not exhibit problematic behavior.³ This appendix describes such benchmark tests, many of which have helped to drive the improvements seen in MOSFET models over the past years. Where relevant, we also outline the characteristics of models that are likely to fail each test.

The benchmarks are a necessary (but not sufficient) set of tests that a model should pass before we can begin to trust it for analog design. A model should pass the tests both qualitatively, showing the expected physical behavior, and quantitatively, by matching measured device characteristics. Although plots of dc current can highlight some problems, inaccuracies are often amplified by differentiation; therefore,

some tests of dc behavior are better done through small-signal conductance tests. Unless otherwise stated, the tests are for wide and long devices at room temperature.

K.1 DC Tests

Continuity and Smooth Behavior Many problems with dc I_D modeling occur at or around the boundaries between operating regions, where I_D and its derivatives can exhibit discontinuities and nonsmooth or nonmonotonic behavior.³ These problems are most prevalent in older “regional” models, such as the SPICE level 2 and 3 models⁴ (which are not in common use today), but not in modern “single-piece” models. Nevertheless, it is always appropriate to test a new model under development for current continuity and smooth behavior (e.g., for models that are part of student projects). Plot I_D vs. V_{DS} for fixed V_{GS} , as well as $\log I_D$ vs. V_{GS} for fixed V_{DS} , and inspect the plots for possible discontinuities or kinks. The general shape of the curves should be as expected from theory, with no such problems. Such issues can be easier to detect in small-signal conductance plots; see Sec. K.2.

Behavior at Zero Bias Using a set of model parameters that turn on all parasitic currents, especially gate current and GIDL, apply a bias of zero to every terminal of a transistor (i.e., short every terminal to ground) and check if all terminal currents are zero. Some models will unphysically predict that currents flow even though there are no voltages driving the device.

Weak-Inversion Behavior A physical behavior that can be difficult to model accurately is the V_{DS} dependence of the drain-source current in weak inversion, which is proportional to $1 - e^{-V_{DS}/\phi_t}$; see (4.8.11). Although this behavior is inherent in the ideal surface potential-based model, it can be disrupted in practice when velocity saturation is incorporated into a model. (Velocity saturation is not important in weak-inversion operation. However, to make a model smooth, the effect of velocity saturation is gradually removed as operation transitions from strong inversion to moderate inversion to weak inversion and is not suddenly switched off; the effective V_{DS} calculation, e.g., as in (10.4.3) or (10.4.4), is still included in a model in weak inversion, and this can cause deviation from the expected physical $1 - e^{-V_{DS}/\phi_t}$ behavior in weak inversion).

The “Gummel slope-ratio test”^{5,6} investigates this aspect of model behavior by plotting the quantity

$$S_R = \frac{I_D / V_{DS}}{\partial I_D / \partial V_{DS}} \quad (\text{K.1})$$

as a function of V_{GS} (from weak to strong inversion), for a small value of V_{DS} and several temperatures. Although this test was originally defined to be calculated based on temperature-independent values of V_{DS} , by selecting $V_{DS} = 0.5\phi_t$ with ϕ_t calculated at each temperature, the slope ratio (K.1) becomes $2(\sqrt{e} - 1) \approx 1.297$ in weak inversion⁷ and should decrease toward 1 in strong inversion.

Figure K.1 shows the Gummel slope ratio as a function of V_{GS} and temperature for $V_{DS} = 0.5\phi_t$. An ideal surface potential-based model gives the correct behavior of Fig. K.1a; a source-referenced threshold voltage-based interpolation model gives the incorrect behavior of Fig. K.1b. The latter model uses both effective V_{GS} and effective V_{DS} computations similar to those described in Sec. 10.4, and those are the cause of the inaccuracies.

Failure to pass this test does not mean a model is not good for design, as the behavior investigated is for small V_{DS} in weak inversion, which is often not a critical biasing region for circuits. However, it does highlight that a model is not physically correct.

Symmetry A fundamental property of the MOS transistor is source-drain symmetry, and this property is exploited in some circuit applications, such as passive mixers and MOSFET-C filters; these are circuits in which the transistor is biased at $V_{DS} = 0$, and the signal makes V_{DS} vary symmetrically around this point, alternating sign. Inaccuracies in modeling high-order derivatives around $V_{DS} = 0$ are manifest as inaccuracies in distortion prediction of such circuits.

The behavior around $V_{DS} = 0$ can be checked using the circuit of Fig. K.2,^{5,6} which implements the “Gummel symmetry test.” A negative body bias ensures that, despite the fact that the potential at these terminals can become negative, the source-body and drain-body junctions do not become forward-biased. A voltage-controlled voltage source ensures that the voltage applied to the source is always the negative of that applied to the drain, thus driving the two terminals in a “balanced” fashion. The channel current should then be an odd function of V_X . However, in the presence of gate and substrate current, I_D (which is what can be measured in practice, as opposed to the channel current) can be nonzero when $V_{DS} = 0$, and so cannot be an odd function of V_X . This problem is solved by monitoring, instead of I_D , the quantity⁸

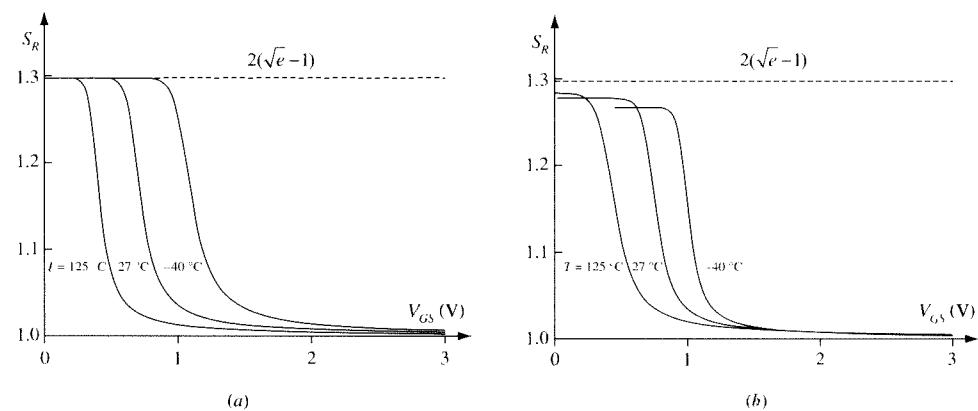


FIGURE K.1

Gummel slope-ratio test. (a) Qualitatively correct curves. (b) Curves from an interpolation model with unphysical V_{DS} dependence in weak inversion. $V_{DS} = 0.5\phi_t$, with ϕ_t calculated at each temperature $V_{SB} = 0$.

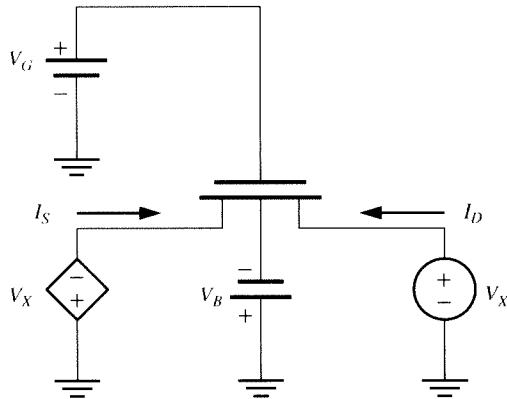


FIGURE K.2
Circuit for the modified Gummel symmetry test, $V_{DS} = 2V_X$.

$$I_X = \frac{I_D - I_S}{2} \quad (\text{K.2})$$

In this quantity, the gate and substrate currents are contained in equal amounts in I_D and I_S at $V_{DS} = 0$ and therefore cancel out in the difference. This makes I_X an odd function of V_X . Plotting I_X , dI_X/dV_X , and d^2I_X/dV_X^2 , as a function of V_X allows the symmetry, or lack thereof, of a model to be visualized; these plots should all be smooth and continuous.

Figure K.3 shows I_X and its first and second derivatives with respect to V_X for four models. The solid lines are for a properly symmetric model with qualitatively correct behavior. The dotted lines are a model with a properly symmetric core I_{DS} model but an asymmetric gate current model; the modified Gummel symmetry test also evaluates the symmetry of gate and substrate current models. The dashed lines are a model with a properly symmetric core formulation but with velocity saturation included in a manner that breaks symmetry. The dot-dashed lines are from a source-referenced threshold voltage-based model.

K.2 Conductance Tests

Differentiation amplifies errors, so it should not be surprising that benchmarks based on small-signal conductance are good at exposing weaknesses in MOS transistor models.

Weak- and Moderate-Inversion Behavior Several similar tests have been proposed to verify behavior from weak through strong inversion.^{3,5,6} A single test, a modification of the “Gummel tree-top test,”^{5,6} that checks both continuity and asymptotic behavior is to plot the transconductance-to-current ratio, g_m/I_D (an important quantity for analog design), vs. $\log I_D$, for several values of V_{SB} and for a finely spaced V_{GS} sweep from weak through strong inversion. Figure K.4a shows the expected physical behavior. The curves should vary smoothly with V_{GS} , and in weak inversion the ratio

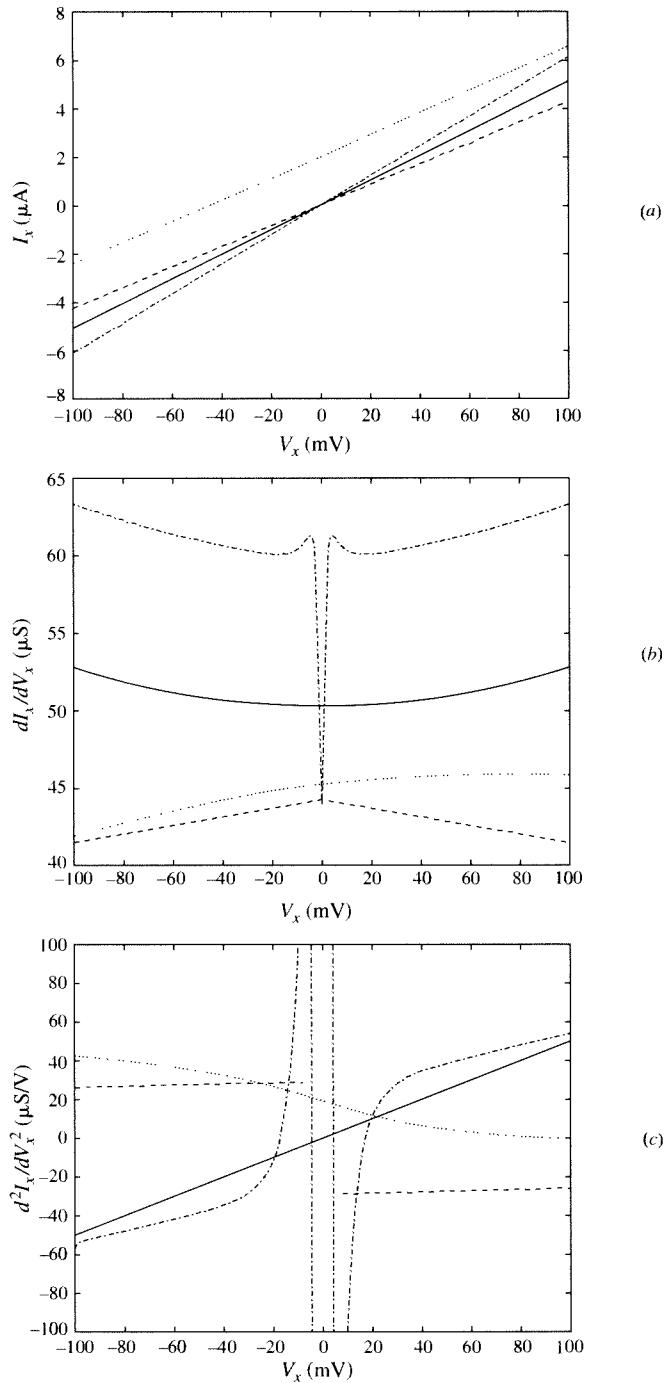


FIGURE K.3
Modified Gummel symmetry test results for four models; see the text for a description of the models.

should asymptotically approach $1/\phi_t$ as V_{SB} increases.^{5,6} More important, because I_D is not *exactly* exponentially dependent on V_{GS} in weak inversion, the curves should peak and then decrease as I_D decreases. Models that approximate weak inversion I_D as being exactly exponentially dependent on V_{GS} will exhibit an incorrect constant value for g_m/I_D in weak inversion.

Figure K.4b shows curves for a model that has an exact exponential $I_D(V_{GS})$ dependence in weak inversion, which causes the curves to be straight lines for small currents. This model also has an artificial transition that links separate weak- and strong-inversion models. The linking function guarantees continuity of both I_D and g_m , yet it is apparent this is insufficient to give proper modeling of g_m/I_D .

Figure K.4c shows curves for an interpolation model (see Sec. 10.4). This model has relative smooth behavior, but is unable to accurately represent the physically expected decrease in g_m/I_D with decreasing I_D .

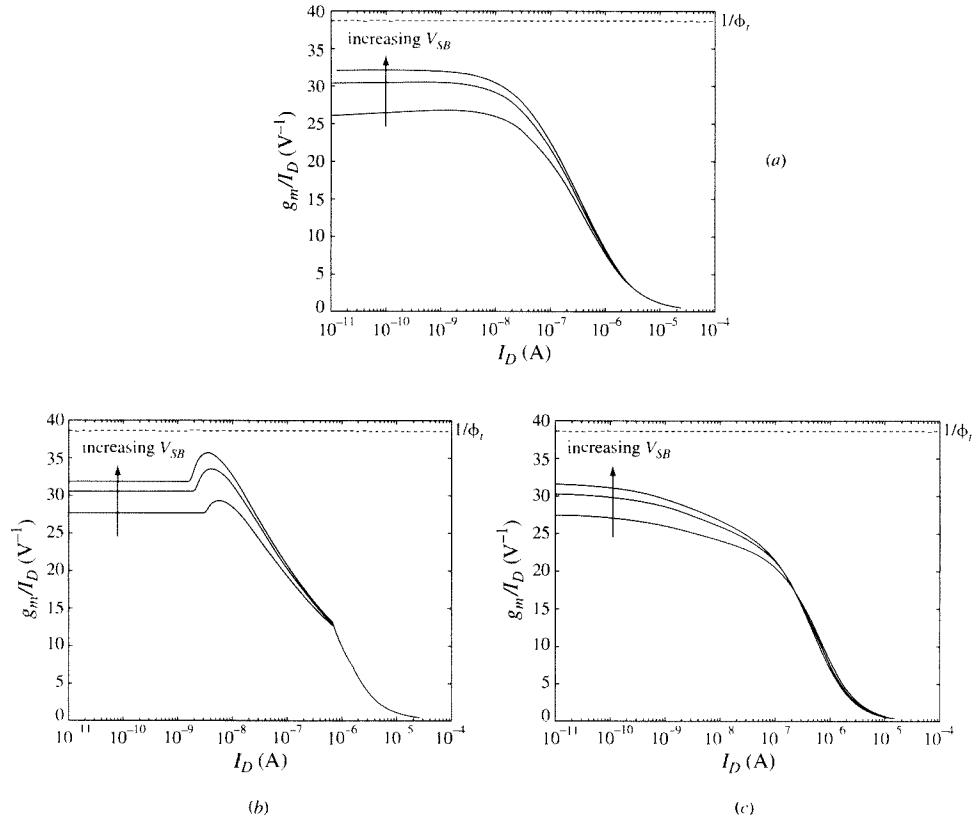


FIGURE K.4

g_m/I_D vs. $\log I_D$. (a) Qualitatively correct curves. (b) Curves from an unphysical regional model. (c) Curves from an unphysical single-piece interpolation model.

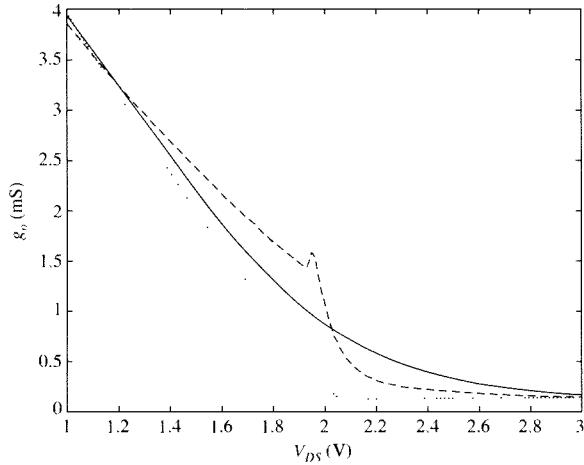


FIGURE K.5
 g_o vs. V_{DS} around the transition from nonsaturation to saturation.

Output Conductance Plot $g_o = \partial I_D / \partial V_{DS}$ vs. V_{DS} (with a fine V_{DS} spacing, concentrating on the region around saturation) for one or more values of V_{GS} in strong inversion. Some models have discontinuities or unnatural behavior in the modeling of g_o through this transition. Such problems are mostly found in older models; recent models generally have appropriately smooth transitions, in part thanks to using interpolation functions to form an effective drain-source voltage $V_{DS,\text{eff}}$ (see Sec. 10.4). Figure K.5 shows such a plot for three models. The solid curve exhibits the qualitatively correct, smooth behavior. The other curves are from models that include “if” conditions at the transition from nonsaturation to saturation; they predict a continuous g_o but produce physically incorrect shapes.

K.3 Capacitance Tests

Often it is not possible to directly access the values of charges for MOSFET models. As we have noted, modeling problems can be highlighted in derivatives; therefore, benchmarking of charge models is best done by inspecting capacitances. Although MOSFET models have improved markedly for dc current and conductance modeling in the past decade, properly incorporating small geometry and other effects into charge models has proven difficult. Many models still have problems with capacitances, such as glitches around the boundaries between regions of operation, unphysical negative capacitance coefficients, and asymmetry at $V_{DS} = 0$ (where we should have $C_{gs} = C_{gd}$ and $C_{bs} = C_{bd}$).

To investigate the behavior of capacitances, the nine capacitances introduced in Chaps. 7 and 8 should be plotted vs. V_{GS} for one or more values of V_{DS} (the expected behavior is shown in Figs. 7.23, 8.9, and 8.10). Also, they should be plotted vs. V_{DS} , in strong inversion (the expected behavior is shown in Figs. 7.20 and

8.7). It can be useful to obtain this plot for a small positive value of V_{SB} , so that the initial V_{DS} value for the sweep can be slightly negative; this highlights asymmetries in capacitances at V_{DS} without causing the drain-bulk junction to become forward-biased. Fine steps should be used for the swept bias, to highlight glitches that can occur over small bias ranges. The capacitances can be determined from small-signal analysis (see note below), and parasitic junction, overlap, and fringing capacitances should be turned off so that the intrinsic MOS model capabilities can be tested (however, the tests can also be run with these parasitics turned on to test for problems in these components).

Inspection of nine capacitances in one plot can be difficult to interpret even for experts in the field, so we provide some limited examples that highlight common problems.

Figure K.6 shows four capacitance coefficients as functions of V_{GS} for $V_{DS} = 0$. This shows four basic problems with the model used to generate these plots (which is a source-referenced threshold voltage-based interpolation model). First, C_{gs} is unphysically nonmonotonic with V_{GS} and has a “glitch” above the threshold voltage (at approximately $V_{GS} = 0.8$ V). Second, C_{gs} is unphysically negative in accumulation. Third, $C_{gs} \neq C_{gd}$ even though $V_{DS} = 0$, i.e., with the device biased symmetrically. Fourth, $C_{gb} \neq C_{bg}$, even though detailed simulations show that these capacitances should be reciprocal at $V_{DS} = 0$.

Figure K.7 shows four capacitance coefficients as functions of V_{DS} for operation in strong inversion; this is a different source-referenced threshold voltage-based interpolation model from that used to generate the capacitances of Fig. K.6. Clearly there are some serious inaccuracies; a large asymmetry around $V_{DS} = 0$ is apparent. More important, C_{bs} should not be negative, and the relative behavior of C_{bs} and C_{bd} should track that of C_{gs} and C_{gd} ; see (7.3.16) and (7.3.17).

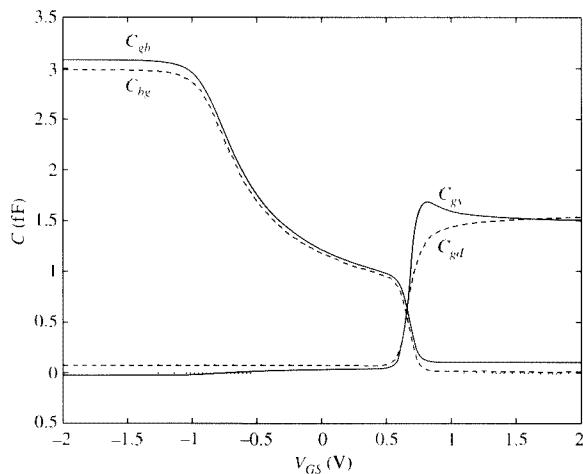
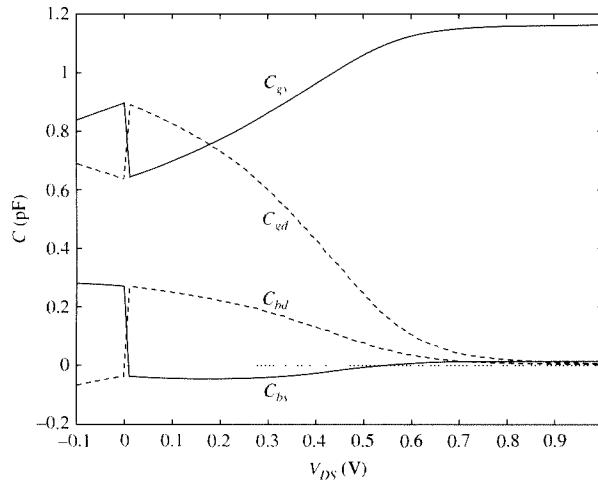


FIGURE K.6
Capacitances vs. V_{GS} , $V_{DS} = V_{SB} = 0$ showing physically incorrect behavior (see text). Dotted line is the zero reference.

**FIGURE K.7**

Capacitances vs. V_{DS} for a V_{GS} value in strong inversion, showing physically incorrect behavior (see text). Dotted line is the zero reference.

Capacitance modeling tends to be more problematic than dc and conductance modeling for most MOSFET models, so detailed benchmarking of capacitances is an important test of the capabilities of any MOSFET model; it generally highlights model deficiencies that are not apparent from only testing I_D behavior. More detailed tests of MOSFET capacitance have also been developed.^{8,9}

A Note on Checking Small-Signal Conductances and Capacitances For small-signal analysis, circuit simulators use first-order linearization of currents and charges with respect to voltages; responses are thus linear with respect to the corresponding excitations. For the general case of voltage excitations applied to all four terminals, the current phasors can be expressed as functions of the voltage phasors using y parameters, as in (8.3.6). The current phasor flowing into terminal k (d , g , s , or b) is of the form

$$I_k = \sum_{l=d,g,s,b} (g_{kl} + j\omega C_{kl}) V_l \quad (\text{K.3})$$

where g_{kl} and ωC_{kl} represent the real and imaginary parts of y_{kl} , respectively, and the sign of C_{kl} can be positive or negative (see Sec. 7.3.2). Consider excitation of only one terminal l , and assume that the voltage phasor excitation (which can be specified in the ac analysis command for a simulator) is $V_l = (1 + j0) \times (1 \text{ V})$. Further, assume that we do the analysis at $\omega = 1 \text{ rad/s}$, i.e., at a frequency of $f = 1/(2\pi) \text{ Hz}$, then

$$I_k = [g_{kl} + j(1 \text{ rad/s})C_{kl}] \times (1 \text{ V}) \quad (\text{K.4})$$

Hence the numerical values of the real and imaginary parts of I_k , which can be printed in most simulators, directly give the numerical values of the conductance and

capacitance (with appropriate sign consideration) between nodes k and l . This is a very simple way to determine conductances and capacitances from circuit simulators and to verify that a model predicts such quantities correctly.

K.4 Noise Tests

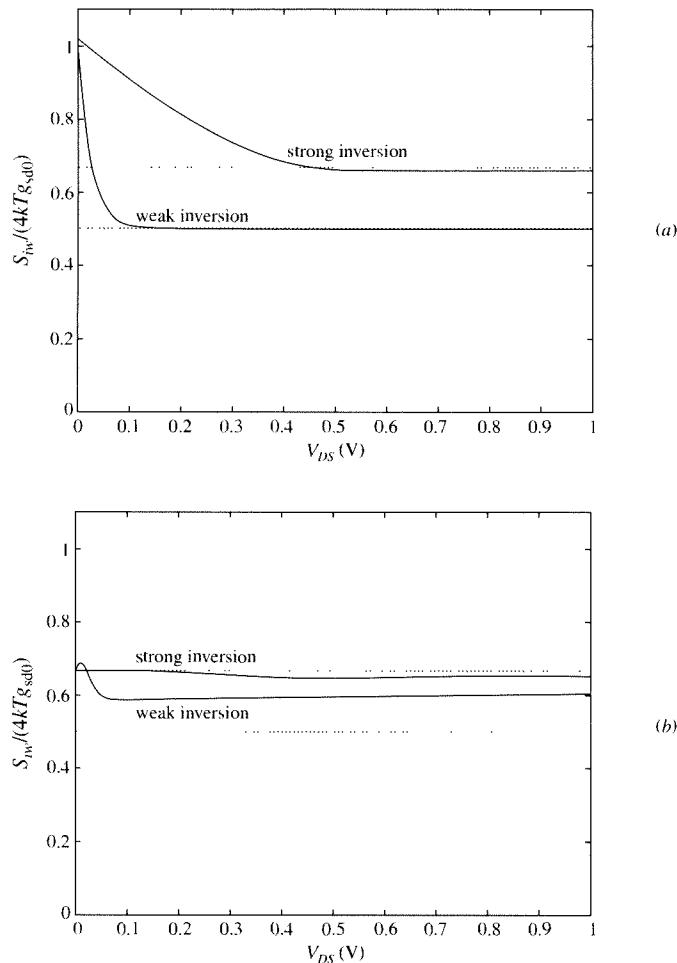
Thermal Noise In a MOSFET biased at $V_{DS} = 0$, the channel is equivalent to a resistor of value $R = 1/g_{sd0}$, where g_{sd} is the drain-source conductance at that bias point; therefore, the channel should show thermal noise consistent with that resistance.³ This is true both in weak and strong inversion (see Sec. 7.5.2). In strong inversion, the noise current power spectral density (PSD), S_{iw} , is described by (7.5.23); thus, this PSD should decrease from $S_{iw0} = 4kTg_{sd0}$ at $V_{DS} = 0$ to $2S_{iw0}/3$ in saturation (see Fig. 7.35). In weak inversion, (7.5.26) shows that the PSD for $V_{DS} > 5\phi_i$ should approach $S_{iw0}/2$. Therefore simulating and plotting $S_{iw}/(4kTg_{sd0})$ as a function of V_{DS} for a value of V_{GS} in weak inversion should give a curve that decreases from 1 to 1/2 over a V_{DS} range of about $5\phi_i$; in strong inversion, it should give a curve that decreases from 1 to 2/3 for $V_{DS} > V'_{BS0}$.¹⁰ If the simulator being used does not allow printing a noise current spectral density, one can use a unity transresistance current-controlled voltage source to drive a noiseless 1- Ω resistor (implemented by a unity transconductance voltage-controlled current source); the noise voltage PSD across the resistor is then numerically equal to S_i .

Figure K.8 shows results from this test with a being results from a reasonable model and b being results from a poor one. Some models can give much worse results than shown in Fig. K.8b, including unphysically predicting zero noise at $V_{DS} = 0$.

Flicker Noise Bias a device in strong-inversion saturation and run a noise simulation to determine S_{iw} at a frequency where the flicker noise should be dominant. Keeping the same W/L ratio, increase the channel area by a factor of 10 and repeat the noise simulation. The PSD of the drain noise current should also decrease by a factor of 10, as is observed in practice (Sec. 7.5.3). This behavior should be observed independent of the value of V_{GS} .

K.5 High-Frequency Test

High-Frequency Transadmittance³ Take the simplest possible model (*remove* all parameters having to do with parasitics, such as junction and overlap capacitances, series resistances, etc.). In the device statement, make sure that the source and drain areas and perimeters are set to zero. We suggest these simplifications to make clear what the following problem is caused by. Bias a 100- μm -long MOSFET in strong-inversion saturation, where the intrinsic gate-drain capacitance is zero. Use an ac source in series with the gate bias as shown in Fig. K.9a, and obtain a frequency response for the ac drain current magnitude up to 10 GHz. Now break the device into two 50- μm -long devices, with their channels in series and with common gate and common substrate, and bias the combination as before. The combination should

**FIGURE K.8**

Noise test results. (a) Correct behavior. (b) A result from a poor model. Dotted lines are at points 1/2 and 2/3.

be equivalent to the single 100- μm device (remember, no junction area is supposed to exist at the intermediate point). Obtain the frequency response again. It should be the same as before. However, several common models give the behavior shown in Fig. K.9b. The behavior is totally different at high frequencies. This is a result of the fact that the models used do not take into account non-quasi-static effects (Chap. 8); the behavior predicted for the 100- μm single device is totally unreasonable,[†] and contradicts both non-quasi-static models and measurements. The two-device combination does a better job at approximating reality because it is a two-element lumped

[†] In fact, the use of transcapacitors in some models can produce even worse errors, predicting that the ac current magnitude goes up with frequency (Fig. 8.25) (this effect will be seen only if it is not masked by extrinsic capacitance effects).

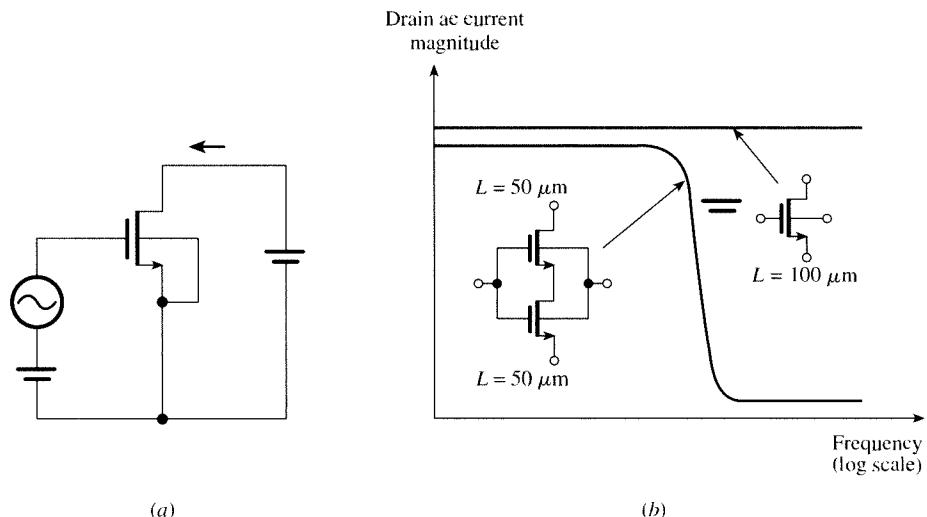


FIGURE K.9
(a) Circuit for obtaining ac response; (b) drain ac current magnitude vs. frequency for a 100- μm -long MOSFET and an equivalent combination of two 50- μm -long devices, using the level 2 SPICE model³ (© 1994 by IEEE).

approximation of what is actually a distributed channel effect. (In fact, such combinations, with two or more elements, can be used in lack of non-quasi-static models, for high-frequency small-signal work; one should be careful, though, not to activate artificial short-channel effects in the subtransistors and not to include extrinsic elements at intermediate channel points.) We note that non-quasi-static behavior has been experimentally demonstrated even in short-channel devices.¹¹

K.6 Small-Signal and Large-Signal Consistency Test

For models that are hand-coded in a low-level programming language such as C, the most common source of problems in implementation is from errors in derivatives. The results of these errors are often slight inaccuracies, but on occasion they can cause significant errors in the prediction of small-signal conductances and capacitances, which can lead to problems in modeling the transition frequency ω_T . In addition, for some models, non-quasi-static effects are formulated differently for different analyses within a simulator. Both of these cases lead to inconsistency between small-signal and large-signal modeling.

In modern simulators with RF analysis capabilities there is a simple procedure than can be used to check consistency of small-signal and large-signal models.¹² Small-signal analysis in simulators is based on linearization of the device relations in a circuit about their dc operating points, and so is fundamentally determined by the derivatives that are coded for a model. A harmonic balance analysis¹³ solves for the nonlinear periodic behavior of a circuit, by determining the waveforms for voltages at each node in a circuit that cause all Fourier components of total current at each

node to sum to zero. This is a form of transient analysis used for nonlinear elements; although the model derivatives are used in it, they are used only for achieving convergence and do not directly affect the final simulation results. Therefore doing a harmonic balance analysis with a small amplitude sinusoidal excitation, and ensuring that harmonic components are small, in essence performs a small-signal analysis without directly using derivatives. The results of the harmonic balance analysis for the fundamental frequency component (i.e., the component at the frequency of the input) can be compared with the results of a small-signal analysis, as described near the end of Sec. K.3, and if there are differences these indicate that there are errors in the derivative calculations of the model. In hand-coded models, these errors, whether accidental or deliberate (e.g., dropping small but computationally intensive terms), are not uncommon.

References

1. Y. Tsividis, "Problems with precision modeling of analog MOS LSI," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 274–277, December 1982.
2. Y. Tsividis and G. Masetti, "Problems in precision modeling of the MOS transistor for analog applications," *IEEE Transactions on Computer-Aided Design*, vol. CAD-3, no. 1, pp. 72–79, 1984.
3. Y. P. Tsividis and K. Suyama, "MOSFET modeling for analog circuit CAD: Problems and prospects," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 210–216, 1994.
4. A. Vladimirescu and S. Liu, "The simulation of MOS integrated circuits using SPICE2," *Memorandum No. UCB/ERL M80/7*, Electronics Research Laboratory, University of California, Berkeley, 1980.
5. H. K. Gummel, "Model implementation and verification facilities for PSIM," *AT&T Bell Laboratories Technical Note*, October 1990.
6. C. McAndrew, H. Gummel, and K. Singhal, "AT&T perspective on compact models," presented at SEMATECH Compact Models Workshop, March 1995.
7. X. Li, W. Wu, A. Jha, G. Gildenblat, R. van Langevelde, G. D. J. Smit, A. J. Scholten, et al., "Benchmark tests for MOSFET compact models with application to the PSP model," *IEEE Transactions on Electron Devices*, vol. 56, pp. 243–251, 2009.
8. C. C. McAndrew, "Validation of MOSFET model source-drain symmetry," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2202–2206, 2006.
9. U. Feldmann, A. Rahm, and M. Miura-Mattausch, "Benchmarking MOS transistor models with respect to capacitances and charges for analog applications," *Proceedings of the IEEE Symposium on Circuits and Systems*, vol. 2, pp. 1352–1355, 1995.
10. A. J. Scholten, R. van Langevelde, L. F. Tiemeijer, and D. B. M. Klaassen, "Compact modeling of noise in CMOS," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 711–716, September 2006.
11. R. Singh, A. Juge, R. Joby, and G. Morin, "An investigation into the nonquasistatic effects in MOS devices with a wafer S-parameter technique," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, Barcelona, March 1993.
12. C. C. McAndrew, Z. Huszka, and G. Coram, "Verilog-A modeling of BJT NQS excess phase," *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 73–76, October 2008.
13. K. Kundert and A. Sangiovanni-Vincentelli, "Simulation of nonlinear circuits in the frequency domain," *IEEE Transactions on Computer-Aided Design*, vol. 5, no. 4, pp. 521–535, 1986.

Index

- Abrupt (step) *pn* junction, 32–43
Acceptors, 7
Accumulation region, 75–76
 charge model, 86–87, 124, 352–353
 small-signal model, 104–109, 431–432
ACM model, 178, 642
Admittance matrix representation, 495. *See also*
 y-parameter models
All-region models
 body-referenced simplified drain-source current
 model, 173–176
 charge formulation drain-source current model,
 177–181
 charge model, 350–351
 complete drain-source current model,
 158–171
 quasi-Fermi potential drain-source current model,
 181–182, 674–677
 simplified charge model, 351–352
 simplified drain-source current model,
 172–182
 source-referenced simplified drain-source current
 model, 176–177
 symmetric linearization drain-source current
 model, 173–175
Alpha-power law model, 604
Avalanche multiplication, 227, 280
Ballistic operation, 285–288
Back gate, 49
Backward propagation of variation
 (BPV), 591
Band gap, 3
Band-to-band tunneling, 302
Barrier lowering, 35, 118–119, 153–154,
 170–171, 271–275
Benchmark tests, 699–711
“Black box” models, 605
Body (substrate, back gate), 44–45, 49
Body (substrate, back gate) terminal,
 45–46
Body charge linearization coefficient, 172,
 173, 176, 178, 196–200
Body charge per unit area
 effect on mobility, 214–219
 general all-region model, 175–176
 linearization, 172–179
 MOS structure, 161
 three-terminal MOS structure, 125, 136
 two-terminal MOS structure, 88, 90–93
Body current, 260
 breakdown (avalanche, impact ionization), 227,
 283–285
 dc current model, 280–283
 small-signal model, 391–393

- Body effect
 coefficient, 84–85, 228
 qualitative description, 48–49, 130–132
- Body referenced vs. source referenced models, 210–212, 609–610
- Body-referenced simplified all-region drain-source current model, 173–176
- Body-referenced simplified strong-inversion drain-source current model, 192
- Body-referenced weak inversion drain-source current model, 204–205
- Boltzmann limit, 408, 412
- Boltzmann transport equation (BTE) models, 601
- Bonds between atoms, 2
- Breakdown
 channel, 226–228, 283–285
 oxide, 227–228, 301
 pn junction (reverse), 40, 46, 226–228
- BSIM models, 574, 603, 638–639
- Built-in potential, 34
- Bulk mobility, 19–20, 159, 695–696
- Bulk punchthrough, 276
- Buried channel devices, 566–570, 690–698
- CAD models. *See* Circuit simulation models
- Capacitances (extrinsic), 371–374, 435–437, 529–538
 contact-related, 373
 fringing, 372–373
 inner fringing, 374
 overlap, 372–373
 pn junction, 40–43, 373
- Capacitances (intrinsic)
 accumulation region, 104–109, 431–432
 all-region models, 430–431, 489–491
 benchmark tests, 705–707
 depletion region, 104–109, 431–432
 five-element medium frequency model, 414–427
 from circuit simulation, 707–708
 gate-body, 372–373
 high frequency models, 474–478, 481–482
 intrinsic, 414–420, 474–478
 medium frequency models, 414–434
 moderate inversion model, 429
 nine-element high-frequency model, 474–478, 481–482
 plots of long-channel capacitance vs. V_{DS} , 424, 485, 486
 plots of long-channel capacitance vs. V_{GS} , 431, 489, 490
- plots of short-channel capacitance vs. V_{DS} , 434, 437
 plots of short-channel capacitance vs. V_{GS} , 434, 437, 491
- polysilicon depletion effect on, 298–291
 quantum mechanical effect on, 293–295
 ratios at low V_{DS} , 423–424, 431–432
 short-channel behavior, 432–434, 490–491
 strong inversion models, 420–427, 484–489
 three-terminal MOS structure, 124
 two-terminal MOS structure, 102–111
 weak inversion model, 489, 489
- Capacitance vs. charge formulation, 613–614
- Capacitor, MOS. *See* Two-terminal MOS Structure
- Carrier velocity saturation, 244–253
 transit time, 357
- Carriers (electrons, holes)
 in extrinsic semiconductors, 4–7
 in intrinsic semiconductors, 2–4
- Channel, 45
 breakdown, 227, 283–285
 gradual channel approximation, 155
 length, 45–46
 width, 45–46
- Channel-length modulation (CLM), 253–259
- Charge balance, 74–75, 123
- Charge density, 13
- Charge neutrality, 13–14
- Charge partitioning, 339, 360–361, 684–686
- Charge per unit area, 19, 75
- Charge pumping, 362
- Charge sharing, 259–260
 effect on threshold voltage, 260–261
 model limitations, 270–271
 narrow-channel effect, 266–270
 short-channel effect, 261–264
- Charge sheet approximation, 89, 158
- Charge vs. capacitance formulation, 613–614
- Charges (intrinsic) for large-signal modeling, 330–334
 accumulation, 352–353
 all-region model, 350–351
 depletion, 352–353
 drain-source charge partitioning, 339, 360–361, 684–686
 moderate inversion, 348
 plots of long-channel charge vs. V_{DS} , 346
 plots of long-channel charge vs. V_{GS} , 355
 quasi-static operation, 330–334
 relation to terminal currents, 334–341, 354

- simplified all-region model, 351–352
- source-referenced simplified strong inversion model, 343–348
- strong inversion, 342–348
- terminal currents from, 334–341
- weak inversion, 98–101, 138–141, 348–350
- Circuit simulation, 616–617
- Circuit simulation models
 - ACM model, 178, 603, 642
 - attributes of good models, 606–608
 - benchmark tests, 699–711
 - BSIM models, 574, 603, 638–639
 - EKV model, 178, 603, 640
 - formulation considerations, 210–212, 608–615
 - HiSIM model, 603, 642
 - implementation, 617–618
 - parameter extraction, 231–233, 618–635
 - PSP model, 175, 176, 603, 640–641
 - source referenced vs. body referenced models, 210–212, 609–610
 - types of models, 602–606
 - UCCM model, 178
 - use in circuit simulation, 615–616
- Circuit simulator implementation, 617–618
- CLM (channel-length modulation), 253–259
- CMOS technology, 54–57
- Compact models, 602–604. *See also* Circuit simulation models
 - Complete all-region drain-source current model
 - current equations, 158–162
 - effective mobility, 218–219
 - surface potential for, 162–165, 169–170
- Complete strong inversion drain-source model, 186–192
- Conductance, 21–22
- Conductance (small-signal)
 - apparent transconductance overshoot, 407
 - body (substrate) transconductance, 387, 397–399, 403, 408, 411–413
 - effect of extrinsic resistance on, 405–407
 - error in modeling source-drain conductance, 399–400
 - gate and body current conductances, 390–394
 - gate transconductance, 387, 396–397, 405, 407–408, 410–414
 - gate transconductance to drain-source current ratio, 409–410, 702–704
 - output conductance, 403–407, 704–405
 - plots of conductance vs. V_{DS} , 403, 404, 406
 - plots of conductance vs. V_{GS} , 410, 413, 415
 - ratios, 397–399, 401–402, 408
 - ratios at low V_{DS} , 423–424, 431–432
 - small-signal models, 386–395
 - source-drain current conductance, 387, 399–403, 408–409, 410–414
 - substrate (body) transconductance, 387, 397–399, 403, 408, 411–413
- Conductance, silicon slab, 21
- Conduction
 - diffusion, 22–25
 - drift, 17–22
 - transit time, 15–16
- Conduction band, 3
- Conductivity, 21
- Constant current threshold voltage, 233, 625
- Constant-field scaling, 308–311
- Constant-voltage scaling, 311–312
- Contact potential, 26–32, 34, 66–69
- Contacts, 57–59, 372–373
- Continuity equation for channel current, 364–365, 503
- Corner frequency for noise, 439–440
- Corner statistical model files, 592
- Coulomb scattering, 213–214
- Crystal lattice, 2
- Cutoff (transition) frequency, 427–429, 534–536
- Degenerate semiconductors, 6, 8, 10, 29, 34
- Depletion approximation, 32, 89
- Depletion mode MOS transistor, 230–231
- Depletion region (MOS transistor operation), 76–77
 - charge model, 86–87, 124, 352–353
 - small-signal model, 104–109, 431–432
- Depletion region (physical)
 - depth for *pn* junction, 36–39
 - MOS, 47, 76, 89–90
 - pn* junction, 32
- Device analysis (TCAD) models, 601–602
- DIBL (drain-induced barrier lowering), 271–275
 - source-drain conductance effect, 401–404
- Dielectric constant, 14
- Diffusion, 22–25
- Diffusion constant, 23
- Direct tunneling, 285, 297
- Donors, 5–6
- Doping, 4
- Drain, 44–45
- Drain terminal, 45–46
- Drain (output) conductance, 403–407

- Drain-induced barrier lowering (DIBL), 271–275
 source-drain conductance effect, 401–404
- Drain-body current. *See* Body current
- Drain-source (intrinsic) current models
 all-region models, 158–182
 body-referenced simplified all-region model, 173–176
 body-referenced simplified strong inversion model, 192
 body-referenced weak inversion model, 204–205
 channel-length modulation effect on, 253–259
 complete all-region model, 158–171
 complete strong inversion model, 186–192
 fluid dynamical analog, 49–52
 interpolation, 209–210, 682–683
 inversion-charge based all region model, 177–181
 moderate inversion models, 209–210
 qualitative description, 47–49
 single-piece models, 209–210, 682–683
 simplified all-region models, 172–182
 strong-inversion models, 186–202
 source-referenced simplified all-region model, 176–177
 source-referenced simplified strong inversion model, 192–199
 source-referenced weak inversion model, 206
 symmetric linearization all-region model, 173–176
 threshold voltage based strong inversion model, 192–199
 velocity saturation effect on, 244–253
 weak inversion models, 204–208
- Drain/source extension implant, 548–551
- Drain (extrinsic, parasitic) resistance, 222–224, 374–375, 435–436, 529–532
- Drift, 17–22
 velocity, 17–18
 velocity saturation, 244–253
- Drift-diffusion device simulation, 601
- Ebers-Moll equations, 205
- Effective channel length, introduced, 46
 channel length modulation, 253–259
- Effective channel width, 46
- Effective interface charge, 69–70
- Effective surface mobility, 212–222
- Effective threshold voltage, 260–261
 halo implant effect, 264–266, 575–577
 narrow-channel charge sharing effect, 266–270
 short-channel charge sharing effect, 261–264
- Effective transverse field, 214–215
- Einstein relationship, 23
- EKV model, 178, 603, 640
- Electric field, 14–15
 drain saturation, 256–258
 pn junction, 37–38
- Electronic charge, 3
- Electrons
 free, 2
 valence, 2–3
- Electrostatic analysis, introduced in one dimension, 653–658
 pn junction, 37–39
 three-terminal MOS structure, 122
 two-terminal MOS structure, 81–82, 89–91
- Empirical compact models, 604
- Energy bands, 3–4
 ballistic transport, 286–287
 bending, 11–12, 35
 drain-induced barrier lowering (DIBL), 271–273
 four-terminal MOS structure, 153–154
 pn junction, 34–36
 three-terminal MOS structure, 118, 120–121
 two-terminal MOS structure, 71–73, 78–79, 81
- Energy barrier, 35, 36, 118–119, 153–154, 170–171, 271–275
- Enhancement mode MOS transistor, 230–231
- Epitaxy, 56
- Equilibrium, 2, 7–12
- Equilibrium concentration, 5
- Equivalent input noise voltage, 452
- Equivalent oxide thickness, 70
- Error criteria, 619–620
- Extrapolated threshold voltage, 96–98, 136–137
- Extension (implant) regions, 57–58, 285, 373–375, 548–550
- Extrinsic parasitics, 371
 capacitances, 371–374, 435–436, 529–538
 gate resistance, 375–377, 435–436, 529–538
 resistance effect on conductances, 405–407
 simplified model, 378–379
 source and drain resistance, 222–224, 374–375, 435–436, 529–538
 substrate resistance, 377, 435–436, 529–538
- Extrinsic semiconductors, 4–7
- Fast surface-state (interface trap) charge, 69, 110–111
- Fermi-Dirac statistics, 7
- Fermi energy level, 8–10
- Fermi potential, 8–10, 34
- FinFET structure, 314–316

- Five-capacitance medium frequency small-signal model, 414–427
- Flatband condition, 71–73, 75
- Flatband voltage, 66–73
- Flicker ($1/f$) noise, 439–440, 450–455
- Forward component of drain-source current, 180–181
- Forward saturation, 189–191
- Fowler-Nordheim tunneling, 296
- Free electrons, 2–4
- Fringing capacitance, 373–374
- Gate, 44–45
- Gate terminal, 45–46
- Gate current
- in circuit applications, 305–307
 - dc current model, 295–301
 - small-signal model, 391–393
- Gate resistance
- minimizing for RF, 533–534
 - modeling, 375–377
- Gate transconductance, 387, 396–397, 405, 407–408, 410–414
- Gate-all-around (GAA) structure, 315–316
- Gate-body extrapolated threshold voltage, 136
- Gate-induced drain leakage (GIDL), 302–306
- Gate-source extrapolated threshold voltage, 193
- Gauss's law, 14, 655
- Generalized scaling, 312
- Global (statistical) variability, 585–586
- Gradual channel approximation, 155
- Gummel slope-ratio test, 700–701
- Gummel symmetry test, 701–703
- Gummel tree-top test, 702–704
- Halo implant, 548–551, 571–573
- complication of parameter extraction, 627–629
 - device characteristics effect, 573–574
 - threshold voltage effect, 264–266, 575–577
- High frequency (intrinsic) small-signal models.
- See also* Small-signal (intrinsic) modeling
 - benchmark test, 708–710
 - complete quasi-static model, 474–491
 - nine-capacitance high-frequency model, 474–478, 481–482
 - noise, 524–529
 - non-quasi-static models, 499–524
 - y-parameter models, 492–499
- High-low doping profile, 555–561
- HiSIM model, 603, 642
- Holes, 2–4
- Hot carrier effects
- body (substrate) current, 280–283, 391–393
 - breakdown, 226–228, 283–285
 - noise, 448
- IGFET (Insulated-gate field-effect transistor), 46
- Impact ionization, 227–280–283
- Impurities, 4–7
- Induced gate noise, 448, 524–528
- Insulated-gate field-effect transistor (IGFET), 46
- Inner fringing capacitance, 374
- Interface trap (fast surface-state) charge, 69, 110–111
- Interpolation drain-source current models, 209–210, 682–683
- Intrinsic carrier concentration, 4
- Intrinsic semiconductors, 2–4
- Intrinsic MOS transistor, 329–330. *See also*
- Capacitances (intrinsic); Drain-source (intrinsic) current models
 - cutoff (transition) frequency, 427–429
 - transition (cutoff) frequency, 427–429
- Inversion, 47–48. *See also* Moderate inversion; Strong inversion; Weak inversion
- described, 76–78
 - moderate inversion, 101–102, 141
 - regions in terms of surface potential, 78–80, 91–92, 110–111, 119–120, 133, 135
 - regions in terms of terminal voltages, 93–98, 110–111, 132–135
 - in three-terminal MOS structure, 124–130, 132–141
 - in two-terminal MOS structure, 76–78, 88–102
 - strong inversion, 94–98, 134–138
 - weak inversion, 98–101, 138–141
- Inversion charge based all-region drain-source current model, 177–181
- Inversion charge per unit area
- all-region model, 161
 - all-region drain-source current model, 177–181
 - drain-source partitioning, 339, 684–686
 - partitioning, 339, 684–686
 - three-terminal MOS structure, 125–127, 138–141, 145–146
 - two-terminal MOS structure, 88–96, 98–101
 - weak inversion, 98–101, 138–141
- Inversion layer, 47
- Ion implantation, 548–551, 577–578
- Johnson (thermal, Nyquist) noise, 438–449, 456, 524–529
- Junction. *See pn junction*

718 INDEX

- Large-signal modeling, 329–330. *See also* Charges
(intrinsic) for large-signal modeling
continuity equation for channel current, 364–365,
503
extrinsic parasitics, 371–379
multi-segment models, 361–363
non-quasi-static operation, 362–371
quasi-static operation, 330–334
terminal currents, 334–341
test for consistency with small-signal modeling,
710–711
- Latchup, 56
- Lateral diffusion, 45, 46
- Layout
layout extraction, 579, 616
folded RF layout, 533–534
- Leakage currents
body (substrate) current, 280–283, 391–393
in circuit applications, 305–307
gate current, 295–301, 391–393
gate-induced drain leakage (GIDL), 302–306
*p**n* junction, 39–40, 46
- Level 1 SPICE model, 200
- Level 2 SPICE model, 187
- Level 3 SPICE model, 195
- Lightly doped drain (LDD) structures, 285
- Limiting (smoothing) functions, 209–210, 610–612,
682–683
- Local oxidation of silicon (LOCOS) CMOS process,
54–55
narrow-channel effect, 268–270
- Local (statistical) variability, 585–592
- Lorentzian noise spectra, 450–451, 454
- Low frequency (intrinsic) small-signal models.
See also Small-signal (intrinsic) modeling
all-region low frequency models, 410–414
complete low frequency model, 393–395
conductance models, 396–403, 407–409,
410–414
- Low-high doping profile, 561–562
- Majority carriers, 6
- Mathiessen rule for mobility, 215
- Maximum frequency of oscillation, 536–538
- Maxwell-Boltzmann statistics, 7–8
- Medium frequency (intrinsic) small-signal models.
See also Small-signal (intrinsic) modeling
all-region medium frequency models, 430–431,
456–458
conductance models, 396–403, 407–409, 410–414
- equivalent circuit topologies, 478–484
five-capacitance medium frequency model,
414–427
transition frequency, 427–429
- Metal oxide semiconductor (MOS) transistor, 43–54
fabrication, 54–60
- Minority carriers, 6
- Mismatch, 585–592
- Mobile ionic charge, 69
- Mobility
bulk, 19–20
effective surface, 212–222
Mathiesen rule for, 215
- Model classification, 203, 602–605
- Model parameter file, 615–616
- Moderate inversion, 91–95
careful definition of limits, 666–668
charge model, 101–102, 141, 348
drain-source current model, 209–210
four-terminal MOS structure, 156–158, 209–210
small-signal modeling, 409–410, 429
three-terminal MOS structure, 135, 141
two-terminal MOS structure, 101–102
- Monte Carlo
circuit simulation model, 592
device simulation model, 601
- MOS capacitor. *See* Two-terminal MOS Structure
- MOS (metal oxide semiconductor) transistor
introduced, 43–54
fabrication, 54–60
- Multi-segment models, 361–363, 708–710
all-region medium frequency models, 430–431,
456–458
conductance models, 396–403, 407–409, 410–414
equivalent circuit topologies, 478–484
five-capacitance medium frequency model,
414–427
transition frequency, 427–429
- Nanowire structure, 315–316
- Narrow-channel effects
charge sharing effect on effective threshold voltage,
266–270
local oxidation of silicon (LOCOS) process,
268–270
shallow-trench isolation (STI) process, 266–268
- n*-Channel MOS transistor, 44–45
- Nine-capacitance high-frequency small-signal model,
474–478, 481–482
- Nonequilibrium, 12–13

- Noise, 436–440
 all-region model, 447–448, 458
 benchmark tests, 708–709
 corner frequency, 439–440
 correlation between drain and gate noise, 526–529
 equivalent input noise voltage, 452
 extrinsic elements, 456
 flicker (1/f) noise, 439–440, 450–455
 high frequency model, 524–529
 induced gate noise, 448, 524–528
 Lorentzian spectra, 450–451, 454
 mobility fluctuation theory, 452
 moderate inversion, 447–448
 Norton equivalent circuit, 438–439
 1/f (flicker) noise, 439–440, 450–455
 random carrier number fluctuation theory, 450–452
 random telegraph noise, 454–455
 short-channel effects, 448–449, 453–455
 shot noise, 439
 strong inversion, 445–446
 thermal (Nyquist, Johnson) noise, 438–449, 456, 524–529
 Thevenin equivalent circuit, 438–439
 unified flicker noise theory, 452–453
 weak inversion, 446–447
 white noise model, 440–449, 456, 524–529
 Non-quasi-static behavior, 362–363
 buried channel devices, 507–519
 described, 362–363
 Non-quasi-static models
 drain current, 500–507
 high frequency models, 499–524
 higher order model, 519–522
 large-signal modeling, 365–371
 model comparison, 522–524
 multi-segment modeling, 362–363
 strong inversion model, 500–519
 y-parameter model, 509–519
 Nonsaturation region, 48, 52–54, 186–188
 Nonuniform transverse doping, 551–555
 buried channel devices, 566–570, 690–698
 drain current, effect on, 564–566
 threshold voltage, effect on, 555–563
 Norton equivalent circuit for noise, 438–439
n-type semiconductor, 5–6
n-well process, 55–56
 Nyquist (thermal, Johnson) noise, 438–449, 456, 524–529
 Ohm's law, 20, 21
 1/f (flicker) noise, 439–440, 450–455
 Output conductance, 403–407, 704–705
 Overlap capacitance, 373
 Oxide, 44
 Oxide capacitance per unit area, 70
 Oxide fixed charge, 69
 Oxide trapped charge, 69
 Pao-Sah model, 169, 676–677
 Parameter binning, 632–633
 Parameter extraction, 231–233, 618–635
 Parasitics. *See* Extrinsic parasitics
 Passivation layer, 58
p-Channel MOS transistor, 45, 228–230
 Permittivity, 14
 Phonon scattering, 213
 Phonons, 213
 Physical compact models, introduced, 602–604
 attributes of good models, 606–608
 benchmark tests, 699–711
 formulation considerations, 210–212, 608–615
 Pinchoff voltage, 145–147, 188–190, 678–683
pMOS transistor, 228–230
pn junction, 32–43
 band-to-band tunneling, 306
 capacitance, 40–43, 373
 dc current, 35, 40, 373
 electrostatic analysis, 37–39
 energy band diagram, 34–36
 forward bias, 34–36
 gate-induced-drain-leakage (GIDL), 302–305
 leakage, 302–307
 reverse bias, 36–40
 Pocket implant. *See* Halo implant
 Poisson's equation, 15, 81, 122, 661, 669
 Polysilicon (poly), 44–45
 Polysilicon depletion, 288–292
 Potential balance, 73–74, 123
 Potentials
 built-in, 34
 Fermi, 8–10, 34
 quasi-Fermi, 181–182, 671, 674–677
 surface, 73
 vs. depth in MOS structure, 74, 90
 vs. position in *pn* junction structure, 38
 Power spectral (noise) density, 437
 Proximity effects, STI stress, 581–584
 well-proximity, 577–581
 PSP model, 175, 176, 603, 640–641

- p*-type semiconductor, 6–7
 Punchthrough, 275–277
- Quantum mechanical (QM) effects, 293–301
 dc gate current, 295–301
 effect oxide thickness increase, 293
 inversion layer centroid shift, 293–294
 tunneling, 295
- Quasi-constant voltage scaling, 311–312
- Quasi-Fermi level, 12–13, 26, 35, 659–660, 671
- Quasi-Fermi potential, 181–182, 671, 674–677
- Quasi-static behavior, 330–334
 complete high frequency model, 474–491
 large-signal terminal currents, 334–341
- Quasi-static model, 330–334
 accumulation charges, 352–353
 all-region charge model, 350–351
 depletion charges, 352–353
 drain-source charge partitioning, 339, 360–361, 684–686
 frequency range of validity, 420
 limitations, 357–362
 moderate inversion charges, 348
 simplified all-region charge model, 351–352
 source-referenced simplified strong inversion charge model, 343–348
 strong inversion charges, 342–348
 terminal currents for, 334–341
 transit time, 355–357
 weak inversion charges, 348–350
- Radio frequency (RF) models. *See also* High frequency models
 gate resistance, 533–534
 model topologies, 529–533
- Random telegraph noise (signal), 454–455
- Recombination, 3
- Regions of operation, 47–49, 75–77, 156–158, 189–192, 194–196
- Regions of inversion, 91–102, 110–111, 132–141, 156–158, 183–185
- Relaxation time non-quasi-static modeling, 614
- Residuals (error criteria), 619–620
- Resistance (sheet), 21
 gate, 376–377
 source and drain, 375
- Resistances (extrinsic, parasitic)
 effect on conductances, 405–407
 gate resistance, 375–377, 435–436, 529–538
- source and drain resistance, 222–224, 374–375, 435–436, 529–532
 substrate resistance, 377, 435–436, 529–538
- Resistivity, 21
- Retrograde doping profile, 561–562
- Reverse (*pn* junction) bias, 36–40
 capacitance, 40–43
- Reverse (*pn* junction) breakdown, 40–41, 226–227
- Reverse component of drain-source current, 180–181
- “Rev 0” models, 602
- Reverse saturation, 191
- Room temperature, 4
- Root mean square (RMS) error, 620, 630
- Salicide (self-aligned silicide) gate, 372, 375–376
- Saturation
 forward, 189–191
 reverse, 191
- Saturation current, 194–198
 channel length modulation effect on, 253–259
 velocity saturation effect on, 250–253
- Saturation region, 48, 52–54, 188–191
- Scaling, 59–60, 307–308
 classical, 308–312
 constant-field scaling, 308–311
 constant-voltage scaling, 311–312
 generalized, 312
 modern, 312–316
 quasi-constant voltage, 311–312
- Scattering, 17
 Coulomb, 213
 phonon, 213
 surface roughness, 213
 transport models, 285–299, 601
- Semiconductors, 1–7
 degenerate, 6, 8, 10, 29, 34
 extrinsic, 4–7
 intrinsic, 2–4
n-type, 5–6
p-type, 6–7
- Shallow-trench isolation (STI) CMOS process, 54–55
- narrow-channel effect, 266–268
 stress effect, 581–584
- Self-aligned silicide (salicide) gate, 372, 375–376
- Sheet resistance, 21–22
- Short-channel effects
 drain-induced barrier lowering (DIBL), 271–275
 channel-length modulation (CLM), 253–259

- charge sharing effect on effective threshold voltage, 261–264
- punchthrough, 275–277
- velocity saturation, 244–253
- Shot noise, 439
- Silicide, 57–58
- Silicon-on-insulator (SOI) CMOS process, 55–56
- Simplified all-region models
 - body-referenced drain-source current model, 173–176
 - charge formulation drain-source current model, 177–181
 - charge model, 351–352
 - complete drain-source current model
 - depletion charge linearization, 172–173
 - source-referenced drain-source current model, 176–177
- Simplified strong inversion models
 - body-referenced drain-source current model, 192
 - source-referenced drain-source current model, 192–199
- Single-piece (interpolation) drain-source current model, 209–210, 682–683
- Single-well process, 55–56
- Simulation. *See* Circuit simulation
- Simulation models. *See* Circuit simulation models
- Skew statistical model files, 592
- Small-dimension effects
 - channel-length modulation (CLM), 253–259
 - drain-induced barrier lowering (DIBL), 271–275
 - gate current, 295–301
 - narrow-channel charge sharing effect on effective threshold voltage, 266–270
 - punchthrough, 275–277
 - quantum mechanical effects, 293–295
 - scaling, 307–316
 - short-channel charge sharing effect on effective threshold voltage, 261–264
 - velocity saturation, 244–253
- Small-signal extrinsic models, 435–437, 529–534
- Small-signal (intrinsic) modeling. *See also* Noise accumulation, 431–432
 - all-region low frequency models, 410–414
 - all-region medium frequency models, 430–431, 456–458
 - body (substrate) transconductance, 387, 397–399, 403, 408, 411–413
 - capacitance models, 420–427, 429–432, 484–491
 - complete low frequency model, 393–395
 - complete quasi-static model, 474–491
- conductance models, 396–403, 407–409, 410–414
- depletion, 431–432
- drain-source current, 386–390
- effect of extrinsic resistance on effective conductance, 405–407
- equivalent circuit topologies, 478–484
- five-capacitance medium frequency model, 414–427
- frequency of validity of models, 491, 522–524
- gate and body currents, 390–393
- gate transconductance, 387, 396–397, 405, 407–408, 410–414
- moderate inversion, 409–410, 429
- nine-capacitance high-frequency model, 474–478, 481–482
- non-quasi-static models, 499–524, 687–689
- non-reciprocity of capacitances, 475–476, 481–482, 485–487, 489
- radio-frequency (RF) models, 529–538
- relationship to terminal charges, 414–419, 474–478
- small-dimension effects, 432–434
- source-drain conductance, 387, 399–403, 408–409, 410–414
- strong inversion low frequency models, 396–405, 420–427
- strong inversion high frequency models, 484–491
- substrate (body) transconductance, 387, 397–399, 403, 408, 411–413
- test for consistency with large-signal modeling, 710–711
- transition frequency, 427–429, 534–536
- two-terminal MOS capacitance, 102–111
- weak inversion, 407–409, 429–430, 489
- y-parameter models, 492–499, 509–518, 687–689
- Smoothing (limiting) functions, 209–210, 610–612, 682–683
- SOI (silicon-on-insulator) CMOS process, 55–56
- Source, 44–45
- Source terminal, 45–46
- Source-drain conductance, 387, 399–403, 408–409, 410–414
- Source referenced vs. body referenced models, 210–212, 609–610
- Source (extrinsic, parasitic) resistance, 222–224, 374–375, 435–436, 529–532
- Source/drain extension implant, 548–551
- Source-referenced simplified all-region drain-source current model, 176–177

- Source-referenced simplified strong inversion
 drain-source current model, 192–199
- Source-referenced weak inversion drain-source current model, 206
- Spline collocation non-quasi-static modeling, 614
- Statistical variability modeling, 584–592
- STI (shallow-trench isolation) CMOS process, 54–55
 narrow-channel effect, 266–268
 stress effect, 581–584
- Stress structure, 58–59
 device characteristics, effect on, 581–584
- Strong inversion, 91–98
 body-referenced simplified drain-source current model, 192
 charge model, 342–248
 complete drain-source current model, 186–192
 drain-source current model including small-dimension effects, 277–279
 four-terminal MOS structure, 156–158, 186–203
 non-quasi-static small-signal model, 500–522
 onset of, 97–98
 pinchoff based drain-source current model, 680–681
 polysilicon depletion effect, 291–292
 potential vs. position, 200–202
 relation to extrapolated threshold voltage, 97
 small-signal model, 396–407, 420–427, 484–489
 source-referenced simplified drain-source current model, 192–199
 three-terminal MOS structure, 124–129
 transit time, 356
 two-terminal MOS structure, 94–98
- Structural effects, complication of parameter extraction, 633–634
- Substrate (body), 44–45
- Substrate (body) terminal, 45–46
- Substrate current. *See* Body current
- Substrate modeling for RF applications, 529–531
- Substrate (body) transconductance, 387, 397–399, 403, 408, 411–413
- Subthreshold slope (swing), 207–208
- Surface effective mobility, 212–222
- Surface potential, introduced, 73
 in deep depletion, 87, 128–129
 and effective mobility, 218
 MOS transistor, 162–165, 169–170
 vs. position, 169–170
 solution for, 162
 two-terminal MOS structure, 84–86, 93
 three-terminal MOS structure, 125–129, 141–143
- Surface punchthrough, 276
- Surface roughness scattering, 213
- Symmetric linearization, 173–175
- Symmetry, 169, 701–703
- Table lookup models, 605–606
- TCAD (device analysis) models, 285–288, 601–602
- Temperature effects, 224–226, 378
- Terminal charges. *See* Charges (intrinsic) for large-signal modeling
- Thermal (Johnson, Nyquist) noise, 438–449, 456, 524–529
- Thermal voltage, 9
- Thevenin equivalent circuit for noise, 438–439
- Thin-oxide effects
 gate current, 295–301
 quantum mechanical effects, 293–295
- Three-terminal MOS structure, 115–116
 accumulation condition, 124
 body effect, 130–132
 circuit applications, 147
 depletion charge per unit area, 125
 depletion condition, 124
 electrostatic analysis, 122
 energy bands, 118, 120–121
 extrapolated threshold voltage, 136–137
 general analysis, 122–124, 669–673
 inversion charge per unit area, 125
 inversion condition, 124–130
 inversion surface carrier concentration, 119–121
 mismatch, 590
 moderate inversion, 135, 141
 pinchoff voltage, 145–147
 potential and charge balance, 133
 semiconductor charge per unit area, 122–123
 small-signal capacitance, 124
 strong inversion, 134–138
 surface electron concentration, 119
 surface potential, 125–129, 141–143
 VCB control point of view, 141–147
 weak inversion, 138–141
- Threshold adjust implant, 549–550
- Threshold voltage, 96–98
 charge sharing, 259–271
 constant current, 233, 625
 drain-induced barrier lowering (DIBL), 271–275
 extrapolated, 96–98, 136–137
 gate-body extrapolated, 136
 gate-source extrapolated, 193
 halo implant, effect of, 264–266, 575–577

- narrow-channel charge sharing effect, 266–270
- nonuniform transverse doping, effect of, 555–563
- p*-channel MOS transistor, of, 229
- short-channel charge sharing effect, 261–264
- small-dimension effects, 259–271
- source-referenced simplified strong inversion drain-source current model, 193–194
- warnings about, 233
- Transconductance
 - apparent transconductance overshoot, 407
 - body current transconductance, 392–394
 - body (substrate) transconductance, 387, 397–399, 403, 408, 411–413
 - effect of extrinsic resistance on, 405–407
 - gate transconductance, 387, 396–397, 405, 407–408, 410–414
 - ratio to drain-source current, 409–410, 702–704
 - substrate (body) transconductance, 387, 397–399, 403, 408, 411–413
- Transconductance-to-current ratio, 409–410, 702–704
- Transit time, 15–16
 - in dc operation, 355–357
 - diffusion, 25
 - drift, 20
 - weak inversion, 356
- Transition (cutoff) frequency, 427–429, 534–536
- Transport current, 330–331
- Tri-gate structure, 314–316
- Triple-well process, 55–56
- Tunneling
 - band-to-band, 302
 - dc gate current, 295–301
 - direct, 285, 297
 - Fowler-Nordheim, 296
 - gate-induced drain leakage (GIDL), 302–307
 - quantum mechanical, 295
 - trap-assisted, 302
- Twin-well (-tub) process, 56
- Two-terminal MOS structure, 65–66
 - accumulation condition, 75–76, 86–87
 - depletion charge per unit area, 88, 90–93
 - depletion condition, 76–80, 86–87
 - electrostatic analysis, 81–82, 89–91
 - energy bands, 71–73, 78–79
 - extrapolated threshold voltage, 96–98
 - flatband condition, 75
 - flat band voltage, 66–73
 - general analysis, 80–86, 661–665
- inversion charge per unit area, 88–96, 98–101
- inversion condition, 76–80, 88–102
- inversion surface carrier concentration, 78–80
- moderate inversion, 101–102
- potential and charge balance, 73–75
- semiconductor charge per unit area, 82–84
- small-signal capacitance, 102–111
- strong inversion, 94–98
- surface electron concentration, 78–80
- surface potential, 84–86, 87
- weak inversion, 98–101
- UCCM model, 178
- Unilateral power gain, 536–538
- Unity power gain frequency, 536–538
- Unconventional operation of the MOS transistor, 171
- Vacancies, 2
- Vacuum energy level, 27
- Valence band, 3
- Valence electrons, 2–3
- Variability modeling, 584–592
- Velocity overshoot, 285
- Velocity saturation, 244–253
 - transit time, 357
- Verilog-A, 617
- Ward-Dutton charge partitioning, 339, 684–686
- Weak avalanche, 280
- Weak inversion, 91–95
 - body-referenced drain-source current model, 204–205
 - charge model, 98–101, 138–141, 348–350
 - four-terminal MOS structure, 156–158, 204–208
 - pinchoff based drain-source current model, 681
 - polysilicon depletion effect, 290–291
 - small-signal model, 407–409, 429–430, 489
 - source-referenced drain-source current model, 206
 - three-terminal MOS structure, 138–141
 - transit time, 356
 - two-terminal MOS structure, 98–101
- Well proximity effect (WPE), 577–581
- White noise, 440–449, 456, 524–529
- Work function potential, 27–32
- y-parameter models, 492–499, 509–518, 687–689

Selected List of Symbols

Certain symbols used only locally within a section, or whose meaning is clear from the context, are not included in this list.

Symbol	Description	Section	Symbol	Description	Section
C'_b	Depletion region capacitance per unit area	2.7	g_{mb}	Body small-signal transconductance	7.2.2
C_{bd}	Body-drain intrinsic capacitance	7.3	g_o	Output small-signal conductance	7.2.5
C_{bs}	Body-source intrinsic capacitance	7.3	g_{sd}	Source-drain small-signal conductance	7.2.2
C_{gb}	Gate-body intrinsic capacitance	7.3	g_{ss}	Source small-signal conductance	7.2.2
C'_{gb}	Gate-body capacitance per unit area	2.7	I_B	Body current	5.8, 6.2
C_{gd}	Gate-drain intrinsic capacitance	7.3	I_D	Drain current	5.8
C_{gs}	Gate-source intrinsic capacitance	7.3	I_{DB}	Drain-to-body current	5.8
C'_i	Inversion layer capacitance per unit area	2.7	I_{DS}	Drain-to-source current (channel current)	4.3-4.8
C_j	pn junction capacitance	1.5	I'_{DS}	Value of I_{DS} at the onset of saturation	4.7
C'_j	pn junction capacitance per unit area	1.5	I_{DSN}	Nonsaturation I_{DS}	4.7
C_m	Difference between C_{dg} and C_{qd}	8.2	I_{DS1}	Component of I_{DS} due to drift	4.3
C_{mb}	Difference between C_{db} and C_{bd}	8.2	I_{DS2}	Component of I_{DS} due to diffusion	4.3
C_{mx}	Difference between C_{bg} and C_{gb}	8.2	I_G	Gate current	5.12
C_{ox}	Total oxide intrinsic capacitance	7.3	I_M	Value of I_{DS} at upper limit of weak inversion	7.3
C'_{ox}	Oxide capacitance per unit area	2.2	i_{DA}	Charging component of drain current	6.3
d_B	Depletion region depth	2.6	i_{SA}	Charging component of source current	6.3
E_c	Conduction band edge	1.2	k	Boltzmann's constant	1.2
E_f	Fermi energy level	1.2	L	Effective channel length	1.6
E_i	Intrinsic energy level	1.2	l_p	Length of pinchoff region	5.3
E_v	Valence band edge	1.2	N_A	Acceptor concentration	1.2
ϵ	Electric field	1.2, 6	N_D	Donor concentration	1.2
ϵ_x	Longitudinal electric field	5.2	n	1. Free electron concentration	1.2
ϵ_y	Transverse electric field	4.11		2. The quantity $(d\phi_w/dV_{FB})^{-1}$	2.6
g_{bd}	Body-drain small-signal conductance	7.2.3	n_i	Intrinsic carrier concentration	1.2
g_{bs}	Body-source small-signal conductance	7.2.3	p	Hole concentration	1.2
g_{gb}	Gate-body small-signal conductance	7.2.3	Q_B	Depletion region charge	6.2
g_{gd}	Gate-drain small-signal conductance	7.2.3	Q'_B	Depletion region charge per unit area	2.6, 3.2
g_{gs}	Gate-source small-signal conductance	7.2.3	\hat{Q}_B	Effective depletion region charge	5.4
g_m	Gate small-signal transconductance	7.2.2	Q_I	Inversion layer charge	6.2
g_{mx}	Difference between g_{gb} and g_{bs}	7.2.3	Q'_I	Inversion layer charge per unit area	2.6, 3.2
			Q_H	Value of Q'_I at drain end of channel	4.4.4
			Q'_{H0}	Value of Q'_I at source end of channel	4.4.4

Symbol	Description	Section	Symbol	Description	Section
Q_o	Gate charge	6.2		weak inversion and depletion, for a given V_{GB}	3.5, 4.6
Q'_o	Gate charge per unit area	2.3	v_d	Drift velocity	1.3, 5.2
Q''_o	Effective interface charge per unit area	2.2	x	Distance along the channel	4.1
q	Magnitude of electronic charge	1.2	y	Distance in direction perpendicular to the surface	2.6
q_D	Drain-associated inversion layer charge	6.3	y_m	Gate transadmittance	8.3
q_S	Source-associated inversion layer charge	6.3	y_{mb}	Body transadmittance	8.3
T	Absolute temperature	1.2	W	Effective channel width	1.6
t_{ox}	Oxide thickness	2.2	α	Coefficient of first-order term in expansion for $-Q'_B/C'_{ox}$	4.4, 4.7.3
V_A	Characteristic voltage in first-order channel length modulation formula	5.3	α_1	Value of α for expansion around the source potential	4.7.3
V_B	Body voltage	6.2	γ	Body effect coefficient	2.4.4, 3.3
V_{CB}	Channel-body voltage	3.2	$\dot{\gamma}$	Effective body effect coefficient	5.4
V_D	Drain voltage	6.2	Δ	Symbol denoting a change in the quantity following it	
V_{DS}	Drain-source voltage	1.6, 4.1	$\Delta\phi$	Difference between the actual strong-inversion surface potential and its classical value of $2\phi_f$	2.6
V'_{DS}	Value of V_{DS} at onset of saturation	4.7.3	μ, μ_{eff}	Permittivity of free space	1.2
V_{FB}	Flatband voltage	2.2	ϵ_0	Permittivity of SiO_2	2.2
V_G	Gate voltage	6.2	ϵ_{ox}	Permittivity of silicon	1.2
V_{GS}	Gate-source voltage	1.6, 4.1	ϵ_s	Degree of nonsaturation	4.7.3
V_H	Value of V_{Gc} , or of V_{GS} , at onset of strong inversion	3.4, 4.6	η	Effective surface mobility	4.3, 4.11
V_{HB}	Value of V_{Gh} at onset of strong inversion	3.4	μ_B	Bulk mobility	1.3
V_{H0}	Value of V_{HB} for two-terminal structure	2.6	τ	Transit time	1.3, 6.5
V_I	Value of V_{Gc} , or of V_{GS} , at onset of weak inversion	3.4	ϕ_{bi}	Built-in potential pn junction	1.5
V_{IB}	Value of V_{Gh} at onset of weak inversion	3.4, 4.6	ϕ_f	Fermi potential	1.2
V_{I0}	Value of V_{IB} for two-terminal structure	2.6	ϕ_{MS}	Work function difference potential	2.2
V_M	Value of V_{Gc} , or of V_{GS} , at onset of moderate inversion	3.4, 4.6	ϕ_{z0}	Thermal voltage, kT/q	1.2
V_{MB}	Value of V_{Gh} at onset of moderate inversion	3.4		Moderate inversion region width in terms of surface potential	3.4
V_{M0}	Value of V_{MB} for two-terminal structure	2.6		Moderate inversion region width in terms of surface potential for two-terminal MOS structure	2.6
V_P	Pinchoff voltage	3.5.2, 4.7.1		Surface potential of two-terminal MOS structure in strong inversion	2.6
V_Q	Value of channel-body voltage at the boundary between strong and moderate inversion, for a given V_{GB}	3.5, 4.6	ψ_{ox}	Oxide potential	2.2
V_S	Source voltage	6.2	ψ_s	Surface potential	2.3
V_{SR}	Source-body voltage	1.6, 4.1	ψ_{sl}	Surface potential in the absence of inversion layer, for a given V_{GB}	2.5, 3.2
V_T	Extrapolated threshold voltage in terms of V_{Gc} or V_{GS}	3.4, 4.7.3	ψ_{sl}	Value of surface potential at drain end of channel	4.3
\hat{V}_T	Effective V_T	5.4	ψ_{so}	Value of surface potential at source end of channel	4.3
V_{TR}	Extrapolated threshold voltage in terms of V_{GB}	3.4	ω_o	Characteristic angular frequency	7.3
V_{T0}	Value of V_{TB} for two-terminal structure	2.6, 3.4	ω_{Ti}	Intrinsic transition angular frequency	7.3
V_w	Value of channel-body voltage at the boundary between				

Values for some useful quantities

Magnitude of electronic charge, q	1.602×10^{-19} C
Thermal voltage, $\phi_t = kT/q$, at 300 K	0.0259 V
Permittivity of silicon, ϵ_s	1.05×10^{-12} F/cm
Permittivity of silicon dioxide, ϵ_{ox}	3.45×10^{-13} F/cm
$\sqrt{2q\epsilon_s}$	5.8×10^{-16} F·V ^{1/2} · cm ^{1/2}